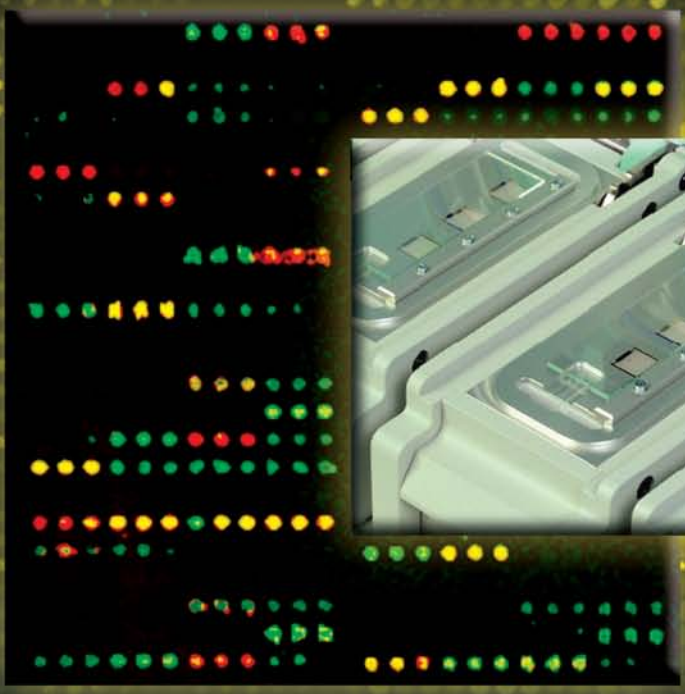
Drug Discovery Series/8

Functional Protein Microarrays in Drug Discovery



CRC Press
Taylor & Francis Group

edited by
Paul F. Predki

Functional Protein Microarrays in Drug Discovery

Drug Discovery Series

Series Editor

Andrew Carmen

Johnson & Johnson PRD, LLC
San Diego, California, U.S.A.

1. Virtual Screening in Drug Discovery, *edited by Juan Alvarez and Brian Shoichet*
2. Industrialization of Drug Discovery: From Target Selection Through Lead Optimization, *edited by Jeffrey S. Handen, Ph.D.*
3. Phage Display in Biotechnology and Drug Discovery, *edited by Sachdev S. Sidhu*
4. G Protein-Coupled Receptors in Drug Discovery, *edited by Kenneth H. Lundstrom and Mark L. Chiu*
5. Handbook of Assay Development in Drug Discovery, *edited by Lisa K. Minor*
6. *In Silico* Technologies in Drug Target Identification and Validation, *edited by Darryl León and Scott Markel*
7. Biochips as Pathways to Drug Discovery, *edited by Andrew Carmen and Gary Hardiman*
8. Functional Protein Microarrays in Drug Discovery, *edited by Paul F. Predki*

Functional Protein Microarrays in Drug Discovery

edited by

Paul F. Predki

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an informa business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 0-8493-9809-6 (Hardcover)
International Standard Book Number-13: 978-0-8493-9809-4 (Hardcover)

Library of Congress Cataloging-in-Publication Data

Functional protein microarrays in drug discovery / [edited by] Paul Predki.
p. ; cm. -- (Drug discovery series ; 8)
Includes bibliographical references and index.
ISBN-13: 978-0-8493-9809-4 (hardcover : alk. paper)
ISBN-10: 0-8493-9809-6 (hardcover : alk. paper)
1. Protein microarrays. 2. Drugs--Design. I. Predki, Paul. II. Series.
[DNLM: 1. Proteins--analysis. 2. Drug Design. 3. Protein Array Analysis. QU 55 F965 2007]

QP551.F96 2007
572'.636--dc22 2007005653

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Table of Contents

Preface

Since their introduction in the 1990s, microarray-based technologies have had a tremendous impact on the biological sciences. One of the most exciting recent developments in this field is functional protein microarrays: microarrays with large numbers of correctly folded and functional proteins. Initially considered an impractical if not impossible goal, high-content functional protein microarrays have now proven their utility in a multitude of applications. While the ''field's" early successes have set the stage for the rapid growth now being witnessed, it is not without its challenges. Indeed, challenges are to be expected in a fast-moving interdisciplinary endeavor such as this, where molecular biology, protein chemistry, bioinformatics, engineering, and physical sciences all intersect.

Currently no book has addressed all aspects of functional protein microarrays in a coherent and integrated fashion. This book is intended to provide the first comprehensive reference for the field, addressing basic principles, methods, and applications. While intended primarily as a reference for industrial, academic, and government scientists, it is also suitable as a graduate-level supplementary text. The book is divided into five main sections, each addressing critical aspects of the field. The first focuses on the generation of functional protein content, which is the first and perhaps most challenging aspect of protein microarrays. The second section describes both "standard" and state-of-the-art fabrication methods, focusing on issues of particular significance to functional protein microarrays. Similarly, the third section reviews current and next-generation approaches to assay detection, which hold one key to the future of the field. The fourth and largest section is dedicated to applications. This section spans the breadth of published applications, from biomolecular interaction discovery and characterization (proteins, antibodies, DNA, small molecules) to humoral response biomarker profiling, enzyme substrate identification, and drug discovery. The final section addresses fundamental computational issues including image and data analysis as well as data visualization.

The intent of this book is to provide the first integrated reference for functional protein microarrays. In doing so, I have aspired to create a volume worthy of the promise of functional protein microarrays, a practical resource capable of conveying the excitement and enabling the development of this field. This book would not have been possible, however, without the hard work of its many authors and Kathie McCoy, to whom I am truly grateful.

Contributors List

Sergei Bavykin
Argonne National Laboratory
Argonne, Illinois

Gloria Biddlecome
Amgen Incorporated
Thousand Oaks, California

John D. Brennan
Department of Chemistry
McMaster University
Hamilton, Ontario

Patrick Burr
The Institute for Genomic Research
The Pathogen Functional Genomics Resource Center
Rockville, Maryland

Darrell P. Chandler
Argonne National Laboratory
Argonne, Illinois
and
Akonni Biosystems, Incorporated
Frederick, Maryland

Thomas G. Chappell
Invitrogen Corporation
Carlsbad, California

Grace Y. J. Chen
Departments of Chemistry and Biological Sciences
Medicinal Chemistry Program of the Office of Life Sciences
National University of Singapore
Republic of Singapore

Arul M. Chinnaiyan
Department of Pathology
and
Department of Urology
and
Bioinformatics Program
and
Comprehensive Cancer Center
University of Michigan Medical School
Ann Arbor, Michigan

Kevin Clancy
Informax
Frederick, Maryland

Scott Clarke
Molecular Probes
Eugene, Oregon

Brian Cunningham
Department of Electrical and Computer Engineering
Micro and Nanotechnology Laboratory
University of Illinois at Urbana-Champaign
Urbana, Illinois

Geeta Devgan
Department of Molecular, Cellular and Developmental Biology
Yale University
New Haven, Connecticut

Ye Fang
Biochemical Technologies, Science and Technology Division
Corning Incorporated
Corning, New York

Igor M. Gavin
Argonne National Laboratory
Argonne, Ilinois
and
University of Illinois at Chicago
Chicago, Illinois

Aparna Girish
Departments of Chemistry and
Biological Sciences
Medicinal Chemistry Program of
the Office of Life Sciences
National University of Singapore
Republic of Singapore

Athena Guo
MicroSurfaces, Inc.
Minneapolis, Minnesota

Steven H. Harwood
Eugene, Oregon

Mingyue He
Technology Research Group
Protein Technologies Laboratory
The Babraham Institute
Cambridge, U.K.

Steve H. Herrmann
Biological Technologies
Wyeth Research
Cambridge, Massachusetts

Yulong Hong
Biochemical Technologies, Science
and Technology Division
Corning Incorporated
Corning, New York

Jing Huang
Department of Molecular and Medical
Pharmacology
David Geffen School of Medicine
University of California
Los Angeles, California

Michael Johnson
Amgen Incorporated
Thousand Oaks, California

Federico Katzen
Invitrogen Corporation
Carlsbad, California

Farid Khan
Technology Research Group
Protein Technologies Laboratory
The Babraham Institute
Cambridge, U.K.

Wieslaw Kudlicki
Invitrogen Corporation
Carlsbad, California

Joydeep Lahiri
Biochemical Technologies, Science
and Technology Division
Corning Incorporated
Corning, New York

Lance G. Laing
SRU Biosystems
Woburn, Massachusetts

Stefan Löfås
Biacore AB, Rapsgatan
Uppsala, Sweden

Brad Love
Corporate Research Laboratory
Invitrogen Corporation
Carlsbad, California

David Lubman
Department of Chemistry
and
Department of Surgery
University of Michigan Medical School
Ann Arbor, Michigan

Alan McWhirter
Biacore AB, Rapsgatan
Uppsala, Sweden

Elizabeth Palmer
Technology Research Group
Protein Technologies Laboratory
The Babraham Institute
Cambridge, U.K.

Scott N. Peterson
The Institute for Genomic Research
The Pathogen Functional Genomics Resource Center
Rockville, Maryland

Brian Rasnow
Amgen Incorporated
Thousand Oaks, California

Keith Robison
Computational Biology
Millennium Pharmaceuticals Incorporated
Cambridge, Massachusetts

Steven Roman
Invitrogen Corporation
Carlsbad, California

Minzi Ruan
VigeneTech Incorporated
Boston, Massachusetts

Nicholas Rupcich
Department of Chemistry
McMaster University
Hamilton, Ontario

Vehary Sakanyan
Unité Biotechnologie, Biocatalyse, Biorégulation
Université de Nantes
and
ProtNeteomix
Nantes, France

John Salon
Amgen Incorporated
Thousand Oaks, California

Jun-ichi Satoh
Department of Immunology
National Institute of Neuroscience
Tokyo, Japan

Daniel S. Schabacker
Argonne National Laboratory
Argonne, Illinois

Barry Schweitzer
Protein Array Center
Invitrogen Corporation
Branford, Connecticut

Dee Shen
Protein Array Center
Invitrogen Corporation
Branford, Connecticut

Marina Snapyan
Centre de Recherché Université Laval Robert-Giffard (CRULRG)
Quebec, Quebec

Michael Snyder
Department of Molecular, Cellular and Developmental Biology
Yale University
New Haven, Connecticut

Arun Sreekumar
Department of Pathology
and
Comprehensive Cancer Center
University of Michigan Medical School
Ann Arbor, Michigan

Michael J. Taussig
Technology Research Group
Protein Technologies Laboratory,
The Babraham Institute
Cambridge, U.K.

Barry S. Taylor
Department of Pathology
and
Bioinformatics Program
University of Michigan Medical School
Ann Arbor, Michigan

Getahun Tsegaye
The Institute for Genomic Research
The Pathogen Functional Genomics Resource Center
Rockville, Maryland

Pratap Venepally
The Institute for Genomic Research
The Pathogen Functional Genomics Resource Center
Rockville, Maryland

Mingwei Wang
National Center for Drug Screening
Shanghai Institute of Materia Medica
Shanghai, China

Xiaoju Wang
Department of Pathology
University of Michigan Medical School
Ann Arbor, Michigan

Brian Webb
Biochemical Technologies, Science and Technology Division
Corning Incorporated
Corning, New York

Shao Q. Yao
Departments of Chemistry and Biological Sciences
Medicinal Chemistry Program of the Office of Life Sciences
National University of Singapore
Republic of Singapore

Harry H. Yim
Invitrogen Corporation
Carlsbad, California

Fang X. Zhou
Protein Array Center
Invitrogen Corporation
Branford, Connecticut

Heng Zhu
Department of Pharmacology and HiT Center
The Johns Hopkins University School of Medicine
Baltimore, Maryland

X.-Y. Zhu
Department of Chemistry
University of Minnesota
Minneapolis, Minnesota

Introduction

As central actors in most biological responses, proteins are the subject of intense study for both basic and drug research. This, in turn, has driven the development of increasingly sophisticated approaches for the study of proteins, which, in recent years has extended to proteomic level methodologies. Despite this need, however, microarray technologies for proteins have lagged behind those for nucleic acids. This has been particularly evident in the case of functional protein microarrays, where formidable technical challenges must be surmounted. However, while technical challenges still remain, the past few years have witnessed movement of the field from basic proof of concept[1,2] to the use of microarrays for important scientific work, landmark discoveries[3] to proteomic characterizations.[4] At the same time, the number of publications in the field has increased exponentially. The purpose of this book is to provide the reader with an up-to-date overview of the field, as well as the background required to actually design and develop arrays or perform and analyze array experiments. The five sections of this book reflect five key considerations in the field: protein content, array fabrication, assay detection, applications and data analysis.

FUNCTIONAL PROTEIN CONTENT

The development of functional protein content is one of the most challenging, and often rate-limiting, aspects of protein microarray experimentation. These challenges can be largely eliminated in cases where protein content or even protein arrays can be acquired commercially. However, in many cases protein content must be generated by the investigator. This content is most typically generated from DNA clones using recombinant expression technology. High-throughput methods for expression clone generation have been developed at The Institute for Genomic Research, and are described in detail in Chapter 1. Important considerations such as information management, automation, quality control and clone validation are addressed. The second chapter addresses expression and purification of proteins in heterologous host systems (*E. coli,* yeast and insect cells), and provides guidance for selecting an appropriate system based on a variety of parameters such as yield, functionality, post-translational modifications, throughput and cost. The final chapter of this section reviews cell-free protein expression systems, and discusses specific considerations for protein microarrays. Together, these chapters provide a thorough overview of the basic considerations for protein content generation.

FABRICATION

The functional and structural heterogeneity of proteins makes arraying and functional surface attachment a considerable challenge. Chapter 4 provides a thorough examination of the challenges of surface chemistry for protein microarrays, which include minimizing nonspecific interactions and maximizing the presentation of conformationally correct proteins. A completely different approach is described in Chapter 5 with the entrapment of proteins in a three-dimensional sol-gel. The various strategies are illustrated in Figure 1.

Critical aspects of array manufacture are addressed in Chapter 6. These include a brief review of commercially available printing technologies, the myriad challenges presented by protein microarrays, and quality control in manufacturing.

The final chapters of this section describe novel strategies for generating protein arrays. Chapter 7 focuses on oriented immobilization strategies based on protein engineering and chemistry, while Chapter 8 addresses the *in situ* generation of proteins. Both chapters describe methods that can "compress" the steps involved in making an array, by combining purification (Chapter 7) or expression and purification (Chapter 8) into the array printing process. These simplified techniques promise to make protein microarray technology more accessible to "average" labs, although at the potential cost of less well controlled array content.

DETECTION

The varied applications of functional protein microarrays all require sensitive assay detection technologies. The most common detection method, fluorescence, is described in chapter 9. This chapter provides a detailed discussion of the basics: fluorescent dyes, fluorescent proteins, time-resolved fluorescence, fluorescent quantum dotes, signal amplification, labeling methods and instrumentation. It concludes

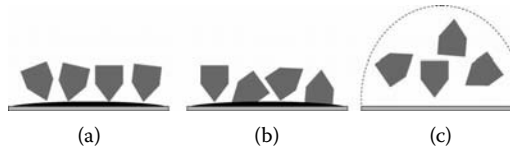


FIGURE 1 Protein immobilization strategies. (a) Proteins are directly attached to the surface based on one (or few) site-specific interactions. This approach has the advantage of a relatively homogeneous presentation of protein to the solution, but regions of the protein may be systematically "hidden." (b) Proteins are attached in a nonspecific orientation. This approach has the advantage of (collectively) displaying a large fraction of the protein surface, but some protein molecules may be functionally blocked. (c) Proteins are not attached but "caged" in an aqueous environment. This approach has the advantage of displaying proteins in a more "native" manner, but can support only a limited range of applications.

by describing a variety of examples of applications enabled by fluorescent detection technology. As useful as fluorescent detection has proven, however, there is a clear need for "label-free" detection methods. This is especially true of small molecule assays, where the addition of a fluorescent group can significantly alter the chemical and biological properties of the compound under investigation. The most common label-free detection technology, surface plasmon resonance (SPR) is described in Chapter 10. This chapter reviews the basic physics behind the SPR phenomenon, and discusses special considerations for the adaptation of SPR to arrays. Chapter 11 describes recent advances in the application of MALDI (matrix assisted laser desorption ionization) mass spectrometry to protein microarrays. In addition to detection, mass spectrometry can be used for molecular identification, potentially enabling highly multiplexed experiments. Chapter 12 describes a recently commercialized alternative to SPR based on photonic crystal biosensors.

APPLICATIONS

Functional protein microarrays have been adapted for a variety of applications in both basic research and drug discovery. Two basic classes of experiments can be performed with functional protein microarrays: interaction assays and activity assays. Interaction assays profile the ability of molecules (or even cells) to bind to proteins on the array surface. Activity assays profile the activity of proteins either in solution or on the arrays themselves (see Figure 2). The breadth of applications generated through these types of experiments is summarized in Table 1, and discussed in more detail in Chapters 13 to 19.

Chapter 13 describes the use of functional protein microarrays for profiling protein-protein interactions, with a focus on 14-3-3 proteins. The use of protein

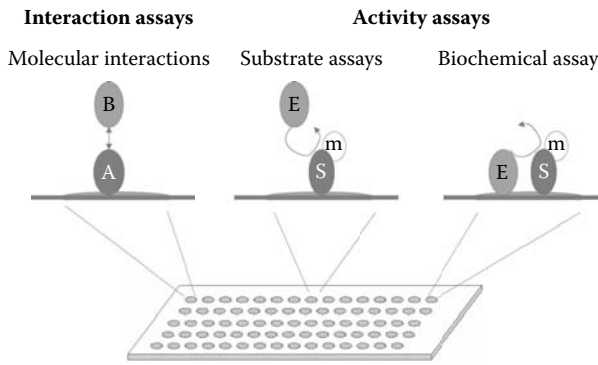


FIGURE 2 Basic types of assays. Interaction assays monitor the ability of a molecule/complex (B) to bind proteins on the array (A). Activity assays monitor the activity of proteins in solution, such as an enzyme (E) modifying (m) a substrate (S) on the array. Alternately, such assays can monitor the activity of proteins on the array. (curved arrow represents a biochemical reaction). (Reprinted with permission, Invitrogen)

TABLE 1
Applications of Functional Protein Microarrays. A Summary of Many of the Basic and Drug Research Applications of Functional Microarray Experiments

Experiment	Basic Research Application	Drug Research Application
Protein–protein interaction profiling Protein–DNA interaction profiling Protein–lipid interaction profiling	Pathway mapping Protein interaction mapping Protein function determination K_d estimation	Target discovery Early target validation
Substrate assays	Pathway mapping Substrate identification	Target discovery Early target validation
Enzyme activity profiling	Pathway mapping Enzyme activity discovery	Target discovery Early target validation
Protein–small molecule interaction profiling	Pathway mapping Metabolomics Chemical genomics	Target/mechanism determination Drug rescue Alternate target identification Specificity profiling IC_{50} estimation Lead optimization Toxicity profiling
Antibody specificity profiling	Antibody characterization	Biotherapeutic development and optimization
Immune response profiling	Biomarker discovery	Diagnostic Biomarkers for efficacy and safety Vaccine design
Enzyme inhibitor profiling	Enzyme characterization	Specificity profiling Lead selection and optimization
Enzyme activity assay	Enzyme kinetics	Specificity profiling IC_{50} determination Lead selection and optimization

Source: Adapted from Predki, P.F., Functional protein microarrays: ripe for discovery, *Curr. Opin. Chem. Biol.*, 8, 8, 2004. With permission.

microarrays for small molecule target identification is described in chapter 14, with an emphasis on the author's pioneering use of protein arrays to study chemical genetics with the compound SMIR4. Chapter 15 discusses the possible uses of functional protein microarrays for biotherapeutic drug development, with a particular

focus on using arrays to identify cross-reactive therapeutic antibodies, as well as monitoring for autoimmune side effects. Chapter 16 describes the use of protein arrays to discover antibody immune response biomarkers, an application which also has implications for vaccine design and testing. The use of protein arrays to study DNA binding is described in Chapter 17, including a discussion of the clinical significance of such investigations. One of the most important and challenging classes of proteins, multi-transmembrane spanning G protein-coupled receptors (GPCRs), is addressed in Chapter 18. This chapter provides a thorough description of this application, from surface chemistry to binding assay protocols and assay validation. Finally, Chapter 19 describes the use of protein arrays for the identification of kinase substrates, focusing on the application of this technique to the yeast proteome.

While exhaustive coverage of all applications is not possible, this section describes all of the major uses of protein microarrays currently under investigation. No doubt, as the field evolves, new applications will be developed. The basics described in this section, though, should provide a good foundation for understanding these future developments.

DATA ANALYSIS

Data analysis is one of the most important, but often underappreciated, aspects of the use of protein microarrays. This section starts with a thorough discussion of image analysis in Chapter 20. Numerous considerations, from spot boundary assignment and contaminant removal to statistical analysis and visualization, are described. Chapter 21 takes off from there, describing approaches to analyzing the numerical data generated directly from the images. Although focusing on biomarker discovery, many of the approaches described in Chapter 21 are directly applicable to the other applications described in this book. Chapter 22 uses computer simulations to help evaluate the potential of protein microarrays for kinase substrate identification. Like the previous chapter, however, the basic approach is applicable to many other applications. The final chapter examines the software requirements for visualizing, sharing and integrating the results of experimentation. It is only with this ability, after all, that the full potential of this technology will be realized.

REFERENCES

1. Zhu, H. et al., Global analysis of protein activities using proteome chips, *Science*, 293, 2101, 2001.
2. MacBeath, G. and Schreiber, S.L., Printing proteins as microarrays for high-throughput function determination, *Science*, 289, 1760, 2000.
3. Hall, D.A. et al., Regulation of gene expression by a metabolic enzyme, *Science*, 306, 482, 2004.
4. Ptacek, J., et al., Global analysis of protein phosphorylation in yeast, *Nature*, 438, 679, 2005.
5. Predki, P.F., Functional protein microarrays: ripe for discovery, *Curr. Opin. Chem. Biol.*, 8, 8, 2004.

Section 1

Functional Protein Content for Microarrays

1 High-Throughput Gene Cloning Using the Gateway® Technology

Scott N. Peterson, Patrick Burr, Getahun Tsegaye, and Pratap Venepally

CONTENTS

INTRODUCTION

The genomic era has produced an ever-increasing number of complete genome sequences from a wide variety of organisms. The large number of annotated gene sequences being produced has driven the advancement of numerous complementary technologies that enable research scientists to exploit the availability of genome

sequence data in new and powerful ways.[1] One such technology is the Gateway® cloning system made available by Invitrogen Inc.[2]

The introduction of this technology was particularly well timed in relation to genomic sequencing, since the Gateway platform provided a vehicle for the cloning and expression of complete open reading frames (ORFs). Prior to the introduction of the Gateway cloning technology, classical cloning strategies using restriction enzymes and DNA ligase were fully entrenched. The primary limitation of traditional cloning procedures was the difficulty in implementing high-throughput approaches. The displacement of the traditional methods would require a clear and substantial improvement in efficiency, ease of use and automation potential. The Gateway technology delivered these requirements. The increased use of this cloning system is in turn driving the development of a novel series of technologies that these expression clones feed directly. Foremost are those technologies associated with *in vivo* and *in vitro* protein expression and purification for functional and structural analysis of proteins. The improved efficiency and ease of generating the raw materials (DNA expression clones, purified recombinant proteins) are supporting a vigorous growth in the use of immobilized proteins on glass surfaces. These advances hold promise for accelerating the discovery of functional roles of genes and provide new strategies for identifying drug targets and therapeutics.

In response to the challenge put forward by the Nation Institute for Allergy and Infectious Disease (NIAID) to generate and distribute cloned ORFs to the scientific community, to enable functional genomics of microbial pathogens, viruses and parasites, the Pathogen Functional Genomics Resource Center (PFGRC) at The Institute for Genomic Research (TIGR) was motivated to identify a cost-effective and efficient cloning technology. It was important to select a strategy that not only provided the necessary efficiency for high-throughput cloning, but also one that would be widely recognized and well-accepted by the diverse scientific end-user. The widespread adoption of the Gateway cloning platform was fortuitous since major cloning efforts performed in other laboratories are now commonly using the Gateway platform and have therefore enabled the collaboration and clone sharing among scientists with diverse scientific interests (see appendix for other users of the technology). While individual applications may vary, the primary purpose of the Gateway platform is the generation of cloned ORFs in one or more expression vectors (destination vector). This is accomplished in two steps that together mimic the recombination reaction that occurs between the genome of *E. coli* and that of phage lambda. The lambda phage genome contains an *att*P site that undergoes recombination, with the aid of lambda phage and *E. coli*–encoded proteins with the 25 bp *att*B site in the *E. coli* genome. Upon recombination, the *att*P site divides in two halves, *att*L and *att*R, that flank the lambda genome. During lytic phase, lambda undergoes a second recombination between *att*L and *att*R thus reconstituting the *att*P site while leaving behind the original *att*B site. In the context of gene cloning, a PCR product (ORF) is generated that is flanked by two nonidentical, primer encoded *att*B sites (*att*B1 and *att*B2). The PCR product undergoes recombination with a vector containing two nonidentical *att*P sites (*att*P1 and *att*P2). The recombination reaction is efficient and directional. The clones derived from this recombination reaction are referred to as entry clones. Entry clones contain inserts that are

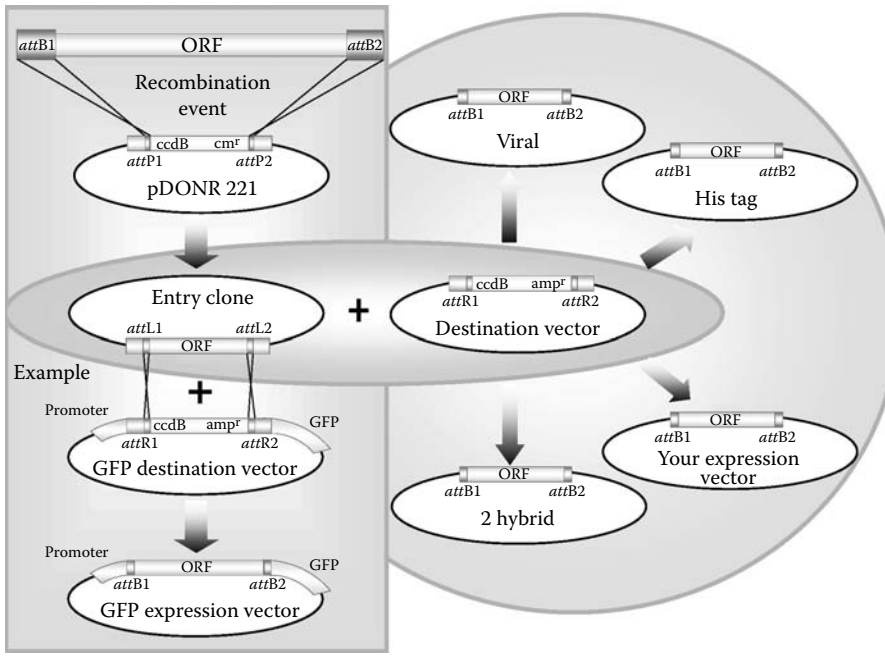


FIGURE 1.1 Gateway Cloning by Recombination.

flanked by two nonidentical *att*L sites. The entry clone has no direct function other than to serve as the substrate for the transfer of the ORF into an expression vector via an LR reaction, named because the *att*L sites in the entry clone recombine directionally with *att*R sites in the destination vector. The entry clone is considered to be a useful resource since the cloned insert can be readily shuttled into any number of commercially available or user-designed expression vectors (destination vectors) without *a priori* knowledge of the intended down-stream application (Figure 1.1).

The generation of entry clones in a high-throughput manner is a multistep process that, taken together, results in a high overall cloning efficiency. This efficiency can be attributed to two features. First, recombination of *att* sites in both BP and LR reactions is nearly stoichiometric and requires the input of a purified PCR product and vector DNA into a proprietary mixture of enzymes that catalyze the recombination of the PCR product into the cloning vector. Second is the system's use of both negative and positive selection in the subsequent transformation of *E. coli*. The successful recombination between vector and PCR product displaces the resident "stuffer fragment" that consists of the markers ccdB and Cmr. The ccdB gene product interferes with gyrA activity and is therefore toxic to *E. coli*. Nonrecombinant vector will retain the ccdB and therefore not be frequently recovered following *E. coli* transformation. The vector backbones used contain standard antibiotic resistance genes for positive selection of transformants. The vast majority (>99%) of colonies that form are recombinant clones. The simplicity of the cloning reaction allows full automation of the steps leading up to and including the cloning reaction itself.

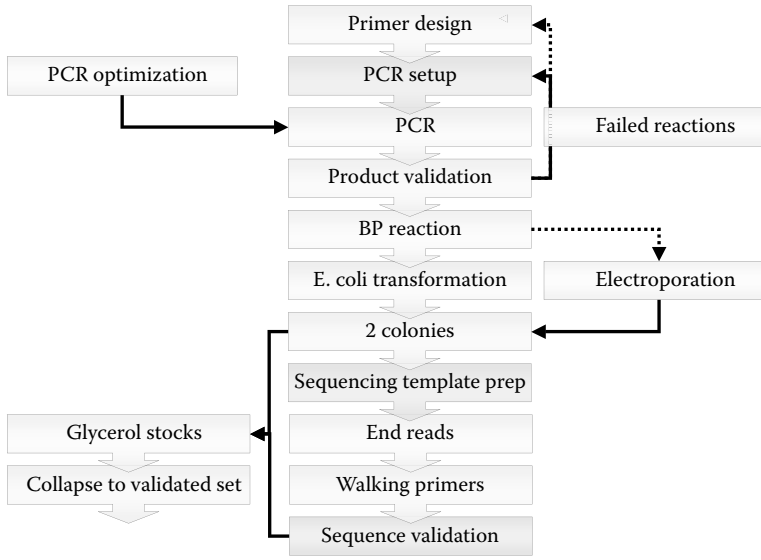


FIGURE 1.2 High-Throughput Clone Production Pipeline.

We have developed a nearly fully automated pipeline for the cloning and sequence validation of ORFs using Gateway. The automation not only provides the potential to generate large numbers of recombinant clones but also the implementation of a process for the tracking of materials through a Laboratory Information Management System (LIMS). The development of a functional LIMS serves to reduce sources of human error and reagent waste. The ability to automate the process is very important to our pipeline since, for several steps, second and even third attempts are made on a small number of failed cases requiring "cherry-picking" and subsequent reintegration with the complete clone set. The creation of a fully functional and automated clone validation sequence analysis process has led to increased throughput and efficiency of clone production. The Gateway clone production pipeline (Figure 1.2) illustrates the integration of this multistep process.

GATEWAY RECOMBINATIONAL CLONING

Primer Design: Each of the unique open reading frames (ORFs) identified and annotated in a genome are potential targets for forward and reverse primer design. Recently duplicated genes displaying a high degree of sequence identity may be difficult, if not impossible, to amplify in pure form. Each forward primer contains a 5′, 25 nt *att*B1 sequence (see Materials and Methods) appended to each gene specific sequence representing the start codon and the neighboring 3′ nucleotides required to achieve a predefined T_m = 60–65°C. The reverse primer has a 5′, *att*B2 sequence appended to gene specific sequence beginning at the nucleotide just upstream of the stop codon. This design feature allows the subsequent flexibility to

create COOH-terminal fusion proteins, wherein a stop codon is conferred by the cloning vector, just downstream of the cloned ORF in each of the three possible reading frames. Some investigators prefer to include a stop codon in the PCR primer. This is accommodated by altering the primer design to include either an endogenous stop codon or a standard (uniform) stop codon. Each primer contains four G residues at the 5′ end. These residues are important for recombination efficiency and serve to internalize the *att*B sequences so that they do appear at the very end of PCR products. We sort our primer pairs with respect to the anticipated PCR product size from smallest to largest. By restricting the T_m of each primer used within a small range, we can define efficient cycling conditions based on the single variable of extension time. For whole genome applications the range of size of ORFs arranged by size in any 384 grouping is relatively small allowing us to define extension times that are nearly optimal for all targets. We have had very good success using oligonucleotides obtained from Illumina Inc. The forward and reverse primers are synthesized in identical well locations in paired 96-well plates, facilitating manual or automated robotic setup of PCR reactions.

PCR Amplification of ORFs

The production pipeline developed in the PFGRC is quite generalized and its overall efficiency is not strongly influenced by the specific ORFs to be cloned. One exception is the prior optimization of PCR conditions for the genomic DNAs of interest. The most pronounced variable to account for is the G+C content of the genome. We have found that species by species optimization of the strategy used for amplification of ORFs using the PCR is critical, especially when large numbers of reactions are to be performed. We have identified four proof-reading polymerases that when applied to particular genomes, perform well (Table 1.1). This list is by no means exhaustive but provides a guideline for robust polymerases for use in a high-throughput cloning process.

Once PCR reaction optimization is complete, high-throughput reaction setup and cycling is ready to begin. We perform PCR in a 35 µl reaction volume in 384-well format. The scale of the reaction provides sufficient yield of product for subsequent cloning reactions. Primer dimers containing both *att*B sites represent clonable products and therefore behave as active competitors with the ORF in BP cloning reactions. An alternative process that we have not investigated thoroughly is the use of a two-step PCR reaction. The primers used differ from those described above and include only the 3′ half of the *att*B sequence. After a limited number of PCR cycles, the products are cleaned-up to remove the initial primers and a second set of universal primers containing a complete *att*B site are used in all reactions. Since the second primer pair is used in all reactions, the cost of primer synthesis can be driven down. After cycling, the PCR products are transferred to 384-well filtration plates (Millipore) using a Beckman Coulter Biomek-FX 96 probe liquid handling robot. For lower throughput applications, a multichannel pipette is a useful alternative. PCR products are purified according to the manufacturer's suggested procedure and products are eluted in 50 µl of H_2O and finally transferred to a clean, 384-well MJ Research hardshell plate using a Beckman Coulter Biomek-FX 96 probe liquid handling robot.

TABLE 1.1
PCR Kits for High-Throughput ORF Amplification

Product	Manufacturer	Catalog	Description	Target Genome
Phusion High-Fidelity DNA Polymerase	Finnzymes New England Biolobs	F-530-L	Proprietary *Pyrococcus*-like enzyme with processivity enhancing domain and proofreading capacity error rate reported: (4.4×10^{-7})	H. pylori, B. anthracis, S. agalactiae, S. typhimurium, S. pneumoniae, V. cholerae, Y. pestis
Platinum PCR SuperMix High Fidelity	Invitrogen	12532–016	Complex of recombinant Taq polymerase and Pyrococcus species GB-D with proofreading capacity error rate reported: 6 × less than native Taq (approx. 1.0×10^{-5})	S. aureus COL
Takara LA PCR kit	Takara	RR013	Proprietary modified Taq polymerase with proofreading capacity error rate reported: 6.5 × less than native Taq (approx. 1.0×10^{-5})	M. tuberculosis
Advantage –HF PCR kit	BD Biosciences	K1909-1	Proprietary mix of a modified Taq polymerase with a Pfu-like proofreading polymerase error rate reported: 20 × less than native Taq (approx. 5.0×10^{-6})	M. tuberculosis

PCR Product Verification and Quantitation

The purified PCR product yield is determined using a Caliper ASM90 SE capillary electrophoresis instrument (Caliper LifeSciences). The Caliper System uses a "sipper" mounted on a robotic arm to remove ~1 µl from each well. Each PCR product is electrophoresed through the single capillary, where its mobility is compared to a set of size standards. The quantity and relative purity (single band) of each PCR fragment is determined in a matter of 30 seconds. The size estimates in our experience are accurate to ±5%. PCR products deviating by more than 10% from an expected size are flagged. In less than 1% of the cases we observe size estimates outside this range. Interestingly, a significant proportion of these are ultimately determined to

TABLE 1.2
Summary of ORFeome Projects

Project	PCR		Transformation		Validation Success Colony 1	Validation Success Colony 2	Overall Success
	1st Round	Follow-up	Heat Shock	Electroporation			
S. aureus COL	93.1%	94.4%	98.0%	99.9%	74.3%	14.3%	88.6%
F. tularensis SHU S4	87.3%	94.6%	96.4%	99.3%	65.8%	18.8%	84.6%

be the expected ORF with no structural rearrangements. The DNA yield, size, and purity are stored directly in the instrument's online computer where it is then classified as passing or failing. Common reasons for scoring a reaction as a failure include low or no yield (<10 ng/μl) or the formation of two or more PCR products (poor primer specificity). Failed reactions (~10%) are identified automatically in a report form that is used to "cherry-pick" appropriate primer pairs for a second pass attempt. A second attempt to amplify failed PCR reactions, using either the identical reaction conditions or those of another kit, generally results in an additional 1 to 8% increase in overall success achieved (~95%). In our experience 0.5 to 1% of PCR failure is attributable to oligonucleotide synthesis. Resynthesis of oligonucleotides for failed reactions imposes additional economic burden on a project with limited returns. A summary of two recent whole genome microbial ORF cloning projects are shown in (Table 1.2). Successful PCR products are then merged together before proceeding to the BP reaction.

BP Clonase Reaction

The output file generated by the Caliper contains the concentration of each PCR product that is then converted to a molar concentration. This information is fed directly into the Biomek-FX SPAN-8 liquid handling robot for automated set up of BP cloning reactions. We have adopted the use of a scaled down version of the BP reaction using 50 fmol of target vector and PCR product insert. A master mix containing all components other than the PCR product insert is prepared and aliquoted into individual wells. Each PCR product is diluted to a concentration of 25 fmol/μl, and subsequently 2 μl of each are added to the master mix. The BP cloning reaction efficiency is inversely proportional to the size of the PCR product to be cloned. This relationship is only strongly limiting for very large genes (~5 Kb). In an earlier version of our pipeline we set up BP reactions at 2 separate scales 25 fmol for PCR products <2.5 Kb and 100 fmol for PCR products >2.5 Kb. More recently we have used 50 fmol scale reactions for all PCR products.

E. coli Transformation

We have not yet identified a reliable 96-well device for electroporation of electrocompetent *E. coli* cells. The efficiency of most cloning efficiency chemically prepared competent cells is more than adequate for recovery of recombinant clones.

E. coli transformations are conducted in 96-well trays and are set up robotically. After heat shock and recovery in nonselective media for 1 hour, cells are robotically plated onto 20 × 20 cm dishes that are sub-divided into 48 grids. Each grid is preseeded with 6 to 8 glass beads (3 mm) and when all 48 transformation reactions are distributed, the plate is shaken gently until all liquid is absorbed into the solid media. The glass beads are removed by inverting the plate into an appropriate receptacle. Any grid yielding one or fewer colonies is considered a failure. A list of failed transformations is compiled and those BP reactions are used a second time to manually transform electro-competent *E. coli* cells. An interesting feature we observe is that virtually all of the PCR reactions scored as failures lead to colony generation, following this two step procedure. Slightly more than half of these cases result in valid full-length recombinant clones.

GATEWAY CLONE RESOURCE VALIDATION PROCEDURE

The validation of Gateway clones is an important aspect of clone production. It is also the most challenging to conduct. Since Gateway clones are most commonly used for expression, it is important to validate the clones for sequence and length integrity. We have not observed substantial DNA rearrangements of cloned inserts; however, nucleotide substitutions introduced during PCR are frequent enough that further consideration is warranted. Thus far, no standards exist for clone validation. In an attempt to initiate such a standard we have adopted a standard set by The Harvard Institute for Proteomics (HIP). The HIP sequence validation standard is stringent but reasonable and involves rejecting any clones containing more than 2 nonsilent substitutions. Clones containing indels (frame shift), nonsense, and/or mutations in the att sites are rejected. These validation criteria have driven our validation strategy to include a two-tier process that begins with the sequence validation of a single colony. If that DNA insert fails the validation criteria, a second colony is then analyzed. Mutations introduced early in PCR cycling will be present in a large fraction of the resulting colonies; however, in practice we find a relatively high degree of utility in going to a second colony in instances where the first colony was deemed unacceptable. Data in Table 1.2 illustrate the utility derived from the analysis of a second colony and our future interest to determine the point of diminished returns in terms of the number of colonies to select for sequence validation. This option must be weighed against the alternative that is to begin the process again from the start.

Initial attempts to sequence validate Gateway entry clones resulted in unacceptably low sequencing success frequencies. This was particularly evident for clones containing small inserts, <600 bp, but also negatively affected end reads obtained for larger inserts. It was suggested that the *att*L sites flanking the cloned inserts have significant potential to form secondary structure that polymerases have difficulty traversing. Specific blocking primers were designed to inhibit the secondary structure formation, thus partially alleviating the barrier to polymerase processivity.[3] We have verified the utility of the blocking primers and observed discrete improvement to our overall entry clone sequencing efficiency. Despite the improvement afforded by the use of blocking primers, our sequencing success was

still below that routinely obtained for other cloning vectors (~90%). We have developed further improvements to the sequence validation of entry clones using phi29 polymerase (Amersham, Inc.) on crude lysates prepared directly from colonies. Templates prepared in this manner are advantageous, although for reasons that are not completely understood. The use of sequencing templates generated through the random priming of plasmid DNAs by rolling circle amplification alleviate the observed sequencing failure for clones containing inserts <600 bp. This strategy has allowed us to generate sequencing results consistent with expected success frequencies and quality. For cloned inserts 600 bp and larger we have had comparable high frequency sequencing success using either templates derived from templiPhi or double-stranded plasmid DNAs. A remaining dilemma in sequence validation of clones involves confirmation of the correctness of the *att*B site itself. Sequence traces from failed validation attempts are often due to abrupt termination in signal strength as the polymerase reaches the *att* site. Yet another possible solution to sequence validation is to direct efforts on the destination clone which is flanked only by *att*B sites.

The generation of DNA templates for sequence validation using templiphi is simple, inexpensive, and easily automated. A sample of 10 µl from an overnight culture is used to prepare lysates by brief heat treatment, 93°C for 3 minutes. A small volume from the cleared lysate is used as template for templiphi reactions. The reaction products are diluted to a final volume of 40 µl with H_2O. We typically obtain yields between 20 and 40 ng/µl, which is sufficient for approximately 20 sequencing reactions.

The number of sequencing reactions performed to validate a cloned ORF is based on its length. For ORFs 500 bp or less, only end-reads are performed. For ORFs larger than 500 bp internal walking primers are designed to generate reads in both directions. The optimal density of walking primers is dependent on average read length and overall sequencing success. We have compared the outcomes of applying walking primers at a regular spacing of 250 bp and 500 bp (Figure 1.3). Given that an average read length from automated sequencing instruments is now in excess of 800 bp, it may be surprising that walking primers at such high density are required for high-throughput sequencing. The failure of some sequencing primers and reactions is a given and if the spacing of sequencing primers is too great, these failures will result in an inability of neighboring primers to fill the gap. The first-pass sequencing attempt (end reads and walking primers) results in a number of outcomes ranging from perfectly validated clones (2X coverage, no mutations) to assemblies with only partial coverage. The classification of sequence validated clones is described below (Table 1.3). We can see that 500 bp spacing among walking primers compares unfavorably to 250 bp spacing in terms of the frequency at which validated clones are identified. From our perspective the choice between 250 bp and 500 bp walking primers spacing is a matter of decision drivers like economics and throughput. The average gene requires seven walking primers for validation and therefore represents a substantial cost. Reducing these costs by nearly 50% is potentially attractive but does carry the consequence of increasing the amount of second-pass sequencing attempts required to fully validate a clone. These additional attempts also carry a cost.

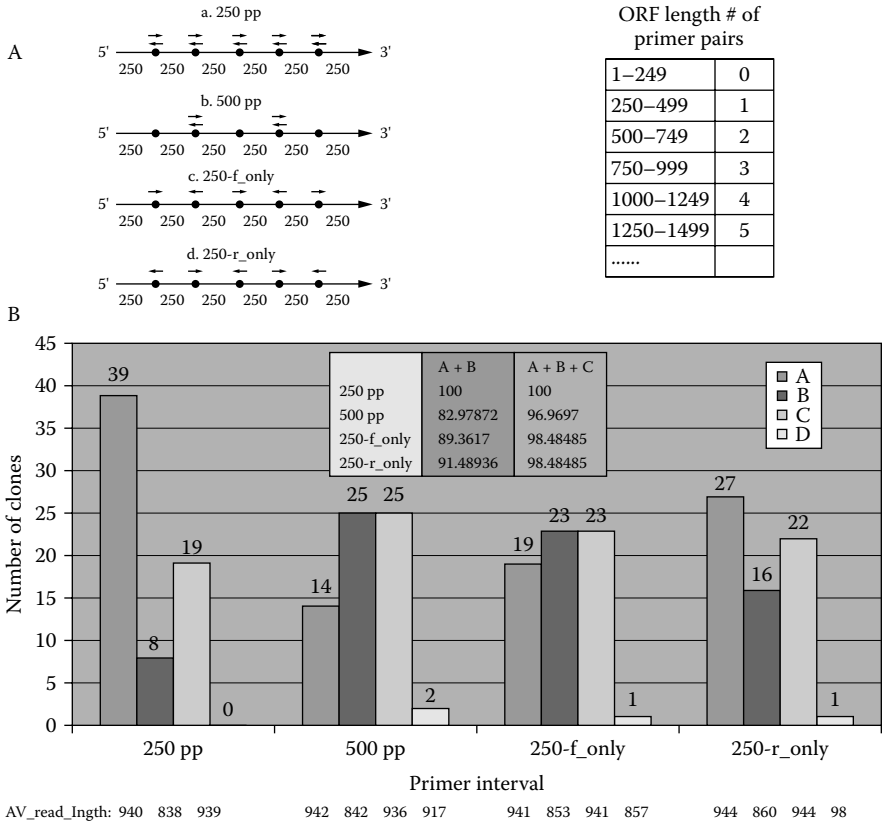


FIGURE 1.3 The effect of walking primer spacing on sequence validation. (A). Primer design schema (left). Walking primer pairs (forward and reverse arrows) at an interval of 250 (a) or 500 bases (b) or single alternating forward and reverse primers at 250 base intervals (c, forward first primer; d, reverse first primer) are used for sequencing. The circles (•) demarcate 250 base intervals. The table on the right lists number of primer pairs designed for various ORF length intervals. (B) Primer intervals vs. sequence coverage. Number of clones in each single-contig sequence validation class A, B, C, and D (see text and Table 1.3) obtained with different walking primer intervals, illustrated in Figure 3A, are plotted. The inset table shows the percentage of full-length A and B wild-type class clones and A, B and C class (full-length clones with mutations) clones seen with different walking primer intervals relative to 250 base interval primer pairs (set to 100). The average length of sequences considered in the validation of clones in each class is shown at the bottom.

Depending on the nature and number of remaining clones, directed efforts are applied to close remaining gaps and confirm sequence ambiguities in the assembled sequence. After applying brute force and manual efforts to elusive clones, a final sequence validation report is generated that directs the representation of acceptable clones for distribution to the scientific community. The list of acceptable ORFs is used to direct the robotic compression of the two freezer copies (colony 1 and 2), into a final set that is replicated into several glycerol stock copies.

TABLE 1.3
Sequence Validation Classes

Valid Classes: The clone has either a full-length or partial-length coverage and shares greater than 90% sequence identity with the wild-type reference	**Invalid Classes: The clone has either a full-length or partial-length coverage and shares less than 90% sequence identity with the wild-type reference**
A — Full-length sequence — 2x or greater sequence coverage at each base, 100% sequence identity with the reference	M — Full or partial-length sequence — less than 90% sequence identity with the reference ORF.
B — Full-length sequence — 1x or greater sequence coverage at each base, 100% sequence identity with the reference	N — Full or partial-length sequence — less than 100% sequence identity with the NON-reference ORF
C — Full-length sequence — sequence variation (< 100% but > 90% sequence identity with the reference)	T — No good quality sequences available for sequence assembly and validation
D — Partial-length sequence — single contig with missing end-sequence (> 90% sequence identity with the reference)	U — Sequencing status unknown
E — Partial-length sequence — multiple contigs with gaps in assembly (> 90% sequence identity with the reference)	W — Full-length sequence — 1 × or greater sequence coverage, 100% sequence identity with the NON-reference ORF
	Z — Failed assemblies due to process errors in the assembly pipeline

Note: Clones are grouped into various valid and invalid classes based on the identity and the coverage of the cloned sequence vis-à-vis the reference.

Sequence Assembly

The sequences obtained from Gateway entry clones are validated using a novel high-throughput assembly pipeline, called CLASP (CLone validation ASsembly Pipeline) developed by the PFGRC bioinformatics group at TIGR. This software can be accessed upon request (www.pfgrc.tigr.org). Initially, the algorithm performs the assembly of individual sequences, generated from the cloned insert, to form a consensus sequence. Subsequently, the consensus is compared to the reference ORF and a validation report detailing the quality of the cloned sequence is generated.

The clone assembly validation pipeline exploits the fact that the sequence of the insert in the vector is already known. As shown in the Figure 1.4, it uses two separate assembler programs to optimize the assembly of the sequence reads and achieve maximum accuracy in the consensus sequence. The first and the most important of these is AMOScmp (Figure 1.4A) which places a premium on the agreement between the reads and the reference — rather than on the phred quality scores generated from the trace files. This guides the selection of which sequences to use for the final contig and thus the consensus. The AMOScmp assembly allows joining of even short overlapping sequences, resulting in a high recovery of single,

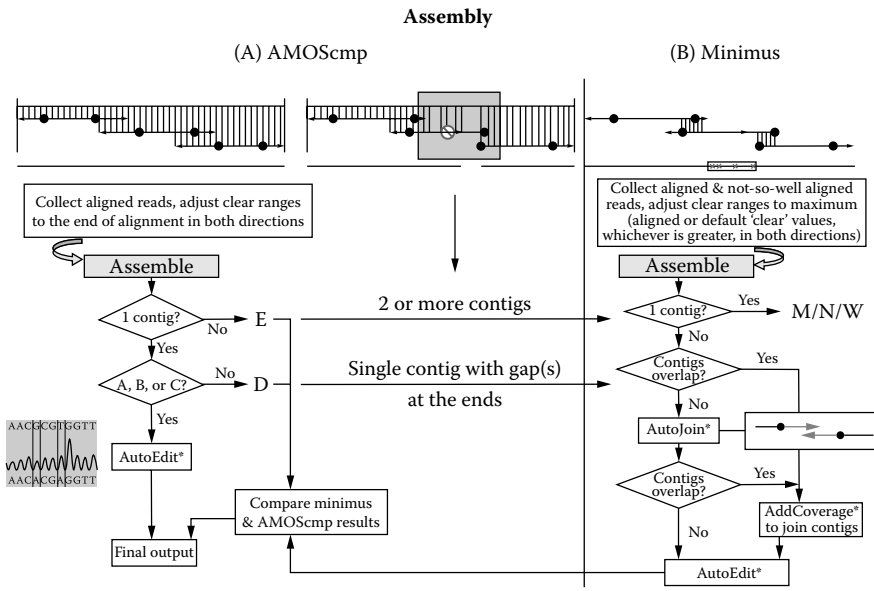


FIGURE 1.4 A schematic diagram of the CLASP assembly pipeline. Two assemblers — one comparative (AMOScmp) and another noncomparative (Minimus) — are used serially to assemble sequencing reads from the inserts cloned into the Gateway vector. The blue circles indicate the ends of individual sequences defined as 'clear' (good quality). AMOScmp does not rely on these but utilizes the alignment with the reference sequence to determine the extent of reads for the assembly of contigs (left panel, A). Minimus (right panel, B) is used to assemble reads that do not align well with the reference sequence. See text for the definitions of various validation classes.

full-length contigs (classes A, B and C, Table 1.3) which otherwise might remain as a single (class D) or multiple (class E) partial-length contigs. For sequences that align very poorly with the reference sequence and result in either partial-length contigs or no contigs after applying AMOScmp, a second assembler, called Minimus (Figure 1.4B), is used. In Minimus the amount of sequence used from any read in the assembly are determined by the phred quality scores instead of how well they align with the reference. For that reason, the consensus sequence(s) obtained from the Minimus assembly have less identity with respect to the reference sequence than those generated by the AMOScmp assembler (classes M, and N; Table 1.3). However, the Minimus results are indispensable for (a) 'catching' clones which are good but mislabeled in any of the steps along the cloning process (class W) and also in the (b) identification of clones that do not meet the acceptance criteria for valid clones and thus require repeat efforts at cloning and validation. After the assembly with AMOScmp or Minimus, the base calls in the consensus sequence(s) obtained for each clone is verified for accuracy against chromatograms using autoEditor*. In the case of multiple contigs with gaps, autoJoiner* is used to join the neighboring reads

* Sequencing closure software developed by TIGR.

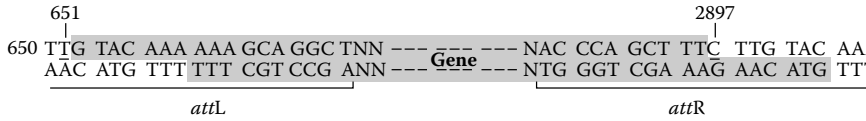
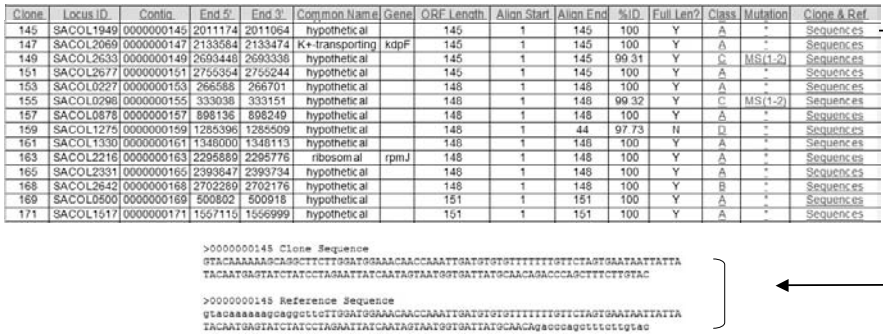


FIGURE 1.5 The recombination sites in the Gateway entry vector. The sequences of *att*L (*att*B1) and *att*R (*att*B2) sites, verified for their integrity along with the enclosed insert, are indicated by the underlines.

by relaxing (extending) the clear ranges if they align well with each other above a set threshold value (Figure 1.4B). The final contig(s) for the clones, which are processed by both assemblers, is chosen based on the best coverage in length and the identity shown vis-à-vis the reference.

Validation and Reports

The consensus sequence generated in the assembly process is analyzed not only for the integrity of the insert but also of the flanking *att*L sites up to BsrGI sites (5' TGTACA 3' sites at 651 on the forward strand and at 2903 on the reverse strand — Figure 1.5.) Following sequence validation by BLASTN and BLASTX analysis against reference nucleotide and protein sequences, respectively, the clones are classified into valid and invalid categories as defined in Table 1.3. The details of the final validation data are presented in two reports. One of them, clone_distribution_report.html (Figure 1.6, partly shown), shows the details of validation for each clone including the sequences for the cloned insert and the reference ORF — via hyperlinks shown in the last column. The Class and Mutations fields in each case are hyperlinked



Clone	Locus ID	Contig	End 5'	End 3'	Common Name	Gene	ORF Length	Align Start	Align End	%ID	Full Len?	Class	Mutation	Clone & Ref
145	SACOL1949	0000000145	2011174	2011064	hypothetical		145	1	145	100	Y	A	-	Sequences
147	SACOL2069	0000000147	2133584	2133474	K+-transporting	kdpF	145	1	145	100	Y	A	-	Sequences
149	SACOL2633	0000000149	2693448	2693338	hypothetical		145	1	145	99.31	Y	C	MS(1-2)	Sequences
151	SACOL2677	0000000151	2755354	2755244	hypothetical		145	1	145	100	Y	A	-	Sequences
153	SACOL0227	0000000153	266588	266701	hypothetical		148	1	148	100	Y	A	-	Sequences
155	SACOL0296	0000000155	333038	333151	hypothetical		148	1	148	99.32	Y	C	MS(1-2)	Sequences
157	SACOL0878	0000000157	898136	898249	hypothetical		148	1	148	100	Y	A	-	Sequences
159	SACOL1275	0000000159	1285396	1285509	hypothetical		148	1	44	97.73	N	D	-	Sequences
161	SACOL1330	0000000161	1348000	1348113	hypothetical		148	1	148	100	Y	A	-	Sequences
163	SACOL2216	0000000163	2295889	2295776	ribosomal	rpmJ	148	1	148	100	Y	A	-	Sequences
165	SACOL2331	0000000165	2393847	2393734	hypothetical		148	1	148	100	Y	A	-	Sequences
168	SACOL2642	0000000168	2702289	2702176	hypothetical		148	1	148	100	Y	B	-	Sequences
169	SACOL0500	0000000169	500802	500918	hypothetical		151	1	151	100	Y	A	-	Sequences
171	SACOL1517	0000000171	1557115	1556999	hypothetical		151	1	151	100	Y	A	-	Sequences

FIGURE 1.6 A partial screenshot of clone_distribution_report.html. Various aspects of the validation data for the clones are displayed in the file. Sequences for the cloned insert and the reference sequences (shown at the bottom) are accessed via hyperlinks present in the last column of the file. See text for the details.

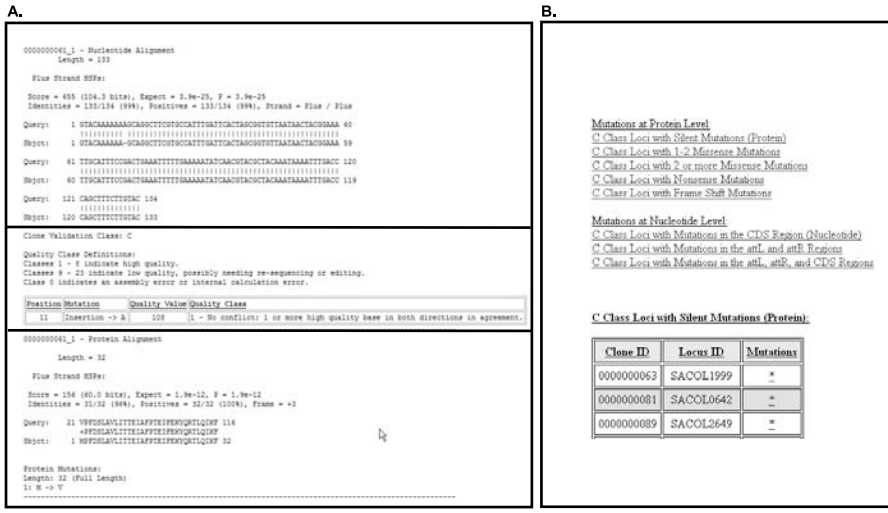


FIGURE 1.7 Clone alignment and summary of mutations. (A) Nucleotide and protein alignments (top and bottom panels) and a summary of mutation(s) (middle panel) are shown. See text for the details. (B) A partial screenshot of clone_distribution_report_C_class_ mutations.html. Various categories of C class clones, grouped based on the number and types of mutations at the protein level, are shown (links at the top and the details of a link at the bottom are shown as an example). In addition, clones categorized on the basis of whether the mutations occur in the CDS or the flanking "*att*" sites, or both at the nucleotide level are shown in the middle (only links shown).

to the files showing nucleotide and protein alignments with the reference sequence (Figure 1.7A, top and bottom panels, respectively) and the summary of mutations, if any (Figure 1.7A, middle panel). Positions of mutations and the associated consensus base call quality values, calculated using the procedure described by Churchill and Waterman,[4] are displayed below each nucleotide alignment. In addition, a quality class is assigned to each mutation suggesting the level of confidence in the base call (Figure 1.7B, middle panel). A second report (clone_distribution_report_C_ class_ mutatations.html) shows the further sub-division of C class clones based on whether the mutations occur within the coding DNA sequence (CDS) or the flanking '*att*' sequences and whether they represent silent or missense or nonsense mutations at the protein level.

The first pass attempt results in a number of outcomes ranging from perfectly validated clones (classes A and B, see above) to assemblies with only partial coverage. In cases where the first pass validation attempts indicate that a clone has more than 2 nucleotide substitutions relative to the reference sequence, or no cloned insert, the second colony, held as a glycerol, stock is used for template production and sequence validation. Upon generating comparable sequence data from the second colony and the validation process, additional clones which pass the acceptable

criteria are identified and consolidated with those selected from the first colony. At this juncture, depending on the nature and number of remaining invalid clones, further sequencing on both first and the second colony templates are performed in an effort to close remaining gaps and resolve any sequence ambiguities in the assembled sequence.

Following the final consolidation of the validated clone set, reports with various details on the validation are generated. The list of acceptable ORFs from these reports is used to direct the robotic compression of the two freezer copies (colony 1 and 2), into a final set that is replicated into several glycerol stock copies (5 to 10). Since accurate storage and retrieval of samples is essential to a facility managing the distribution of thousands of clones, PFGRC has acquired and installed a Biophile Storage and Individual Vial Retrieval System (Biophile, TekCel) for this purpose. The system consists of five –80°C storage units (BSU) and a –40°C individual vial retrieval unit (IVR). The system is integrated into a relational database, utilizing bar code information to identify each clone and its location in storage. The IVR system automatically records, stores, and retrieves each requested vial based on a prepared worksheet. The "rearraying" of individual vials into new sets allows accurate retrieval of clones or clone sets.

DESTINATION VECTOR CLONING

The ability to automate a large portion of the Gateway Clone Resource production pipeline accounts for the high-throughput capabilities that the PFGRC now offers. The ease of use of the Gateway technology makes possible the construction and validation of 10,000 or more expression clones annually. The scale up of the procedure is largely dependent on the acquisition of additional robots. The ability to transfer cloned and validated entry clone inserts into Gateway expression vectors is straightforward. Purified entry clones and destination vectors are mixed and recombination occurs faithfully via an LR clonase reaction. The screening of recombinant clones and validation of their quality can be performed in a streamlined manner as well, since there is no need to repeat the sequence validation, the only important check being to establish that the complete ORF is present in the expression vector. In general, a single colony can be selected for insert validation. Direct PCR from selected colonies using forward and reverse Gateway® primers allow a rapid and cost-effective method for qualifying the expression clones.

TECHNOLOGY AND ROBOTICS

The PFGRC utilizes two versions of the Beckman, BioMek FX platform to automate most steps in the process, including PCR reaction setup, PCR product purification, cloning reaction setup, transformation of chemically competent cells, plating of transformed cells, plasmid isolation, setup of sequencing plate format, and replication of clone stock copies for distribution. The FX-96 platform transfers equal volumes of

96 samples in parallel, and is used for processes that have pre-equalized concentrations of all reagents, such as PCR reaction setup, and clone stock replication. The FX-Span8 platform transfers samples individually with one of eight pipetting tips with independent volume and sample well location control. Additionally the range of movement and software flexibility allows the deposition of transformation reactions onto oversized 48-well, sectored agar culture plates for colony isolation. Together these two instruments have provided the necessary flexibility and accuracy to make high-throughput Gateway cloning efficient and reliable. One adaptation to reaction setup necessitated by the use of robotics is to ensure that each individual pipetting step delivers 2 μl or more, since overall pipetting accuracy below these volumes is lower.

Accurate qualitative and quantitative assessment of PCR amplification products is an essential part of the high-throughput application of Gateway technology. PCR products must be screened for the presence of multiple bands, incorrect size products, and failed reactions. The precise size (bp) and concentration of successfully amplified products must be known to ensure that optimal recombinational cloning efficiency is achieved. The PFGRC utilizes the Caliper ASM 90 SE Capillary Electrophoresis platform for this purpose. Other similar technologies offered by Agilent Inc and others, perform in a similar way. The Caliper instrument has the ability to accomplish these tasks in parallel with a high degree of accuracy and walk away automation. The system performs electrophoresis in a single gel filled micro capillary channel with high resolution and processing speed (100 to 5000 bp, 30 seconds per sample), allowing the characterization of a 384-well microtiter plate in approximately three hours. The PCR products are detected using a fluorescent dye that provides high sensitivity detection of secondary products and smears that could go unnoticed using traditional agarose or polyacrylamide slab gels. Sizing of the detected PCR products is automated and accurate within ±5%. The concentration of any bands detected is also calculated automatically based on a standardization sample. The area under the peak is calculated for product bands and compared to the standard. This method provides results that are more accurate than traditional absorbance readings taken at 260 nm as it relies on an intercalating fluorescent dye rather than absorbance that can be skewed by multiple factors. The output from the system is then transferred to liquid handling robots for subsequent automated reaction set up.

To achieve a high level of success in a high-throughput endeavor such as the Gateway clone validation pipeline, tracking various laboratory and data processing steps in a systematic way is very critical. A software system like LIMS (Laboratory Information Management System) or a similar resource will be very helpful in that effort and can aid in the creation of high-quality Gateway clones in the following ways: (a) by capturing measurement data, a LIMS can ensure that the correct values are used for PCR evaluation and calculations, (b) by automatically generating robot rearray scripts, a LIMS can prevent plate-to-plate transfer errors, as well as speed up lab processing, and (c) by providing analysis and reporting tools, it can provide valuable metrics that allow lab personnel to evaluate the quality of their techniques over time to improve them.

METHODS AND MATERIALS

PCR Amplification of ORFs

Forward and reverse primers are designed to amplify each ORF from a reference genome sequence. Each oligonucleotide sequence is then appended 3' of the *att*B1 (forward) and *att*B2 (reverse) sequence.

Forward Primer:

5′ GGGG ACAAGTTTGTACAAAAAAGCAGGCTTC (N18-25) Gene Specific Seq 3'

Reverse Primer:

5′ GGGG ACCACTTTGTACAAGAAAGCTGGGTC (N18-25) Gene Specific Seq 3'

Forward and reverse primer pairs (Illumina/Invitrogen, Carlsbad, CA) at a concentration of 25 μ*M* are combined into a master mix containing 0.15 μM of each dNTP, reaction buffer and 40 ng of genomic DNA with total reaction volume of 35 μl. Typical cycling conditions after a 1 minute initial denaturation at 98°C are as follows: 98°C for 10 seconds, 55°C for 30 seconds and 72°C for 1 minute per kb intended product size. Reactions are cycled through these temperatures 25 times followed by a 72°C final extension of 10 minutes. After cycling, the PCR products are transferred to 384-well Millipore filter plates (Millipore, Billerica, MA) using a Beckman Coulter (Beckman Coulter, Fullerton, CA) Biomek FX 96 probe liquid handling robot. Filter plates are then subjected to a vacuum of 10 inches of Hg for approximately 10 minutes. Then, 50 μl of Milli-Q water is added to each filter plate, PCR products are eluted by aspiration and then transferred to a clean, 384-well MJ Research hardshell plate (Bio-Rad Waltham, MA) using a Beckman Coulter Biomek FX 96 probe liquid handling robot.

PCR Product Verification and Quantitation

Each PCR product is analyzed by capillary electrophoresis on the Caliper LifeSciences (Hopkinton, MA) AMS 90 SE Instrument using LabChip HT 2.4.1 software. A 384-well or 96-well plate containing PCR sample volumes no less than 25 μl is placed in the instrument. Two trays containing ladder and buffer respectively are equipped alongside a "Caliper Chip" which must be cleaned, primed, and prepared with gel dye and marker. Before beginning a run, an input file, containing only ORF IDs and ORF lengths are loaded into the computer along with user input of the plate type and allowed percent deviation from actual ORF length. This information will be used to later calculate the concentration of the fragments found. A 96-well plate will take about an hour to resolve; consequently a 384-well plate will resolve within 4 hours.

After finishing the run, two files are generated and saved. One output file contains the actual electronic gel images and the second output file (of entirely text format) contains the summary data obtained from every well including concentration, size, and if the original fragment was found. This second output file is converted by virtue

of a simple script into a .csv file and is directly imported into the Biomek FX Span-8 liquid handling robot for automated set up of BP cloning reactions using equimolar quantities (50 fmol) of target vector and PCR product insert.

BP Clonase Reactions

BP cloning reactions are performed in 96-well, MJ Research (Bio-Rad Waltham, MA) plates and are conducted in a 15 µl total volume containing: 50 fmol entry clone vector, pDONR221, 50 fmol PCR product, 2 µl of proprietary BP clonase enzyme, (Invitrogen, Carlsbad, CA) 3 µl of 5x BP clonase buffer (Invitrogen, Carlsbad, CA) and brought to volume with 1 × TE. Reaction plates are then incubated for 16 hours at 25°C in a thermal cycler, followed by 4°C hold until recovered. BP clonase reactions are terminated through the addition of 2 µg of proteinase K (Invitrogen, Carlsbad, CA) for 20 minutes at 37°C.

DH10B-T1 *E. coli* Transformation

Chemically competent, DH10B-T1 *E. coli* cells in 96-well format (Invitrogen, Carlsbad, CA) are thawed on ice and 2 µl of the BP cloning reaction are added using the Biomek FX 96 probe instrument. The plates are sealed with sterile covers and incubated on ice for 30 min. The plates are transferred to a thermal cycler, prewarmed to 45°C. The plates are held for 30 seconds and immediately transferred to ice for 2 min. Cells are allowed to recover by adding 40 µl of SOC media and incubating at 37°C without shaking for 1 hour. Qtray bioassay trays (Genetix Limited, U.K.) with 48 divided areas containing LB media supplemented with 50 µg/ml kanamycin and 2% agar are warmed to room temperature. Several 3 mm glass beads are added to each well and 30 µl of cells are then pipetted onto the agar surface. The Qtrays are shaken gently until all visible liquid has been absorbed into the plates. The beads are discarded and the plates are incubated at 37°C for 16 to 18 hours. Transformation efficiencies are scored by colony count estimations (1–10, 10–50, >50) for each transformation. The Qtrays are held at 4°C.

Clone Sequence Validation

Colonies are picked with sterile toothpicks into 1250 µl of 2x YT media, supplemented with kanamycin 50 µg/ml, in 96 deep well blocks. The blocks are sealed with an airpore tape pad strip and incubated at 37°C for 17 hours (11 hours static and 6 hours shaking at 800 rpm). These inoculations are performed in duplicate, one being specified for stock generation. Freezer copies of clone sets are generated in 96-well Matrix Track Mate 2-D bar-coded vials (Matrix Technologies Hudson, NH) by combining 50 µl of overnight culture to an equal volume of 75% glycerol using the Biomek FX 96 probe liquid handling robot.

Plasmid Extraction

Plasmid DNA is purified using what is essentially the Qiagen (Qiagen, Valencia, CA) R.E.A.L preparation method using Qiagen's QIAfilterTM and appropriate buffers.

E. coli overnight cultures are collected by centrifugation of deep-well blocks at 3200 rpm for 15 minutes at 4°C. The growth medium is decanted and the pellets are resuspended in 300 µl of R1 buffer (50 m*M* Tris pH 8) containing a final RNase concentration of 100 ng/ml. The cells are lysed by the addition of 300 µl of R2 buffer (1% [w/v] SDS, 200 m*M* NaOH) with gentle mixing. The lysates are incubated at room temperature for 5 min. The lysates are neutralized by the addition of 300 µl of R3 buffer (3 *M* KOAc, pH 5.5) followed by mixing and incubation on ice for 10 min. The Biomek FX 96 probe instrument is used to transfer lysates into QIAfilter TM filter plates. The lysates are cleared by vacuum filtration. The plasmid DNA is precipitated through the addition of 625 µl of isopropanol. After mixing, the plates are spun at 3200 rpm for 30 min. at 4°C. The supernatants are decanted and the pellets are washed with 300 µl 70% (v/v) ethanol (–20°C). The plates are spun at 3200 rpm for 15 minutes. The supernatants are decanted and allowed to air dry to completion. The plasmids are resuspended in 50 µl of Blue Tris dye (1 m*M* Tris pH 8.0, bromophenol blue 1.25 mg/ml) by shaking for 30 minutes on a platform shaker.

Sequencing Template Production by TempliPhi

Sequencing templates generated by TempliPhi (Amersham Biosciences, U.K.) uses the Phi29 DNA polymerase and rolling cycle amplification to generate linear concatenated copies of plasmid templates. The Biomek 96 probe instrument is used to transfer 10 µl of overnight culture into 384-well plates. The cells are collected by centrifugation at 3200 rpm for 5 minutes. The media is decanted by inverting the plates on absorbent material followed by low-speed centrifugation at 500 rpm for 2 to 3 sec. The Biomek Span-8 is used to add 2 µl of lysis buffer to cell pellets/well. The plates are then placed in a thermal cycler and incubated at 93°C for 3 min. The plates are returned to the Biomek and 34 µl of Milli-Q H_2O is added to each well. The plates are then sealed and spun at 3200 rpm for 5 min. Two microliters of the supernatant are transferred to a clean 384-well MJ Research plate (Bio-Rad Waltham, MA). The enzyme (4 µl) is added to the supernatants and the plates are sealed and incubated at 30°C for 16 hours. The reactions are stopped by heat treatment at 96°C for 5 min. The Biomek FX 96 probe liquid handling robot then adds 34 µl of Milli-Q H_2O Blue Tris dye to each well.

REFERENCES

1. Marsischky, G. and LaBaer, J., DNA many paths to many clones: A comparative look at high-throughput cloning methods, *Genome Res.*, 14, 2020, 2004.
2. Hartley, J.L., Temple, G.F., and Brasch, M.A., DNA cloning using *in vitro* site-specific recombination, *Genome Res.*, 10, 1788, 2000.
3. Esposito, D., Gillette, W.K., and Hartley, J.L., Blocking oligonucleotides improve sequencing through inverted repeats, *Biotechniques*, 35, 914, 2003.
4. Churchill, G.A. and Waterman, M.S., The accuracy of DNA sequences: Estimating sequence quality, *Genomics*, 14, 89, 1992.

2 Protein Expression for MicroArrays

Harry H. Yim, Thomas G. Chappell, and Steven H. Harwood

CONTENTS

INTRODUCTION

With the completed sequence of the human genome, as well as the sequencing of the genomes of hundreds of other species, the structure and function of literally hundreds of thousands of proteins are of potential interest in diverse fields of biology. Proteomic studies increasingly require the expression of large numbers of proteins in parallel. No one expression system has proven be ideal for all types of downstream applications; each host having advantages and disadvantages when evaluated for protein yield, functionality, posttranslational modifications, high-throughput (HTP) capacity and cost. Trade-offs are required to optimize high-throughput

output based on downstream requirements, which are frequently mutually exclusive. Posttranslational modifications, for instance, might be critically important in drug screening applications but incompatible with crystallization studies. As a result, open reading frames are often expressed in a variety of heterologous host systems; an approach that has been facilitated by the development of improved cloning methods such as the Gateway® system, which utilizes *in vitro* recombination,[1] allowing the rapid and flexible cloning of recombinant DNA into a variety of expression vectors.

E. COLI EXPRESSION

Introduction

E. coli remains the most widely used system for rapidly expressing large numbers of proteins and is in many respects a model system for high-throughput protein production. Protein expression in *E. coli* is relatively reliable, robust, simple, amenable to HTP expression, and cost-effective. Continual improvements have resulted in increased throughput and decreased growth volumes. There are also well developed protocols for cloning, expression and purification, many which have been highly optimized and automated for small scale.[2,3] Despite the advantages that the *E. coli* protein expression system provides, there are some drawbacks to using *E. coli* as an expression host. These include lack of posttranslational modifications and contamination of protein product with endotoxin. However, the most significant problem for proteomics applications is that proteins expressed in *E. coli* often accumulate as insoluble and inactive aggregates. One possibility for the high fraction of insoluble proteins may be related to the use of the popular and well established T7 expression systems. In this approach, one employs the bacteriophage T7 late promoter on medium copy number plasmids. The highly active T7 RNA polymerase is provided by the host cell and regulated by the IPTG-inducible *lac*UV5 promoter. While this system provides very high concentrations of recombinant protein, it may be a victim of its own success in that it may produce more protein than the cell is capable of properly folding.

Optimizing Soluble Protein Expression

One common approach to improving solubility for T7 and other systems is to alter the expression conditions to promote folding. These are generally applicable for HTP format and are geared towards reducing the rate of protein synthesis to provide more time for folding. Methods to decrease the expression levels include using lower concentrations of IPTG or coexpression of phage T7 lysozyme (which degrades T7 RNA polymerase) from compatible pLysS and pLysE plasmids. Another approach is to use a promoter with the native *E. coli* RNA polymerase, rather than T7 RNA polymerase, to transcribe the mRNA. This allows more efficient coupling of transcription and translation, potentially leading to more soluble product. Another easy and effective method is to reduce the growth temperature and allow a longer period for protein synthesis. For example, rather than performing the expressions at 37°C for 3 hours, temperatures are reduced to between 18°C and 30°C and proteins

expressed for longer periods of time. One interesting method involves growing the preinduction cultures at 42°C in order to induce the expression of heat shock proteins and supply the cell with chaperones to improve folding.[4] Upon induction, the cultures are grown at lower temperatures to enhance folding.

Other strategies to promote the expression of properly folded recombinant protein, include coexpression of molecular chaperones[5] and foldases,[6] two classes of proteins play an important role in *in vivo* protein folding. Molecular chaperones (GroES-GroEL, DnaK-DnaJ-GrpE, ClpB) promote the proper isomerization and cellular targeting by transiently interacting with folding intermediates. Foldases, such as peptidyl prolyl cis/trans isomeases (PPI) or disulfide oxidoreductase (DsbA) and disulfide isomerase (DsbC), accelerate rate-limiting steps along the folding pathway.

The choice of growth medium can also have effects on protein expression. Standard LB media is an inexpensive and easy to prepare media, however it is poorly buffered and is not supplemented with a carbon source. Newer medias have become commercialized (AthenaES™, Baltimore, MD) that increase biomass of the culture and expression of recombinant proteins. Additionally, adding osmolytes to increase in osmotic pressure causes the cell to accumulate osmoprotectants in the cell, which may stabilize the native protein structure. Other reagents such as ethanol, low molecular weight thiols and disulfides, and NaCl also may improve folding.[7]

The latest improvement to *E. coli* growth media has been developed by William Studier at the Brookhaven Labs.[8] The media has been optimized to not only promote high density growth, but also contains a combination of sugars to first repress expression from lac promoters (including the T7 system) and then automatically induce them in late log-phase growth due to the depletion of carbon sources other than lactose. This media is ideal for HTP applications in that it eliminates the need to monitor cell density for adding IPTG and thus eliminates a very laborious part of the process.

Another method to increase the likelihood of obtaining a soluble protein is to fuse a highly soluble fusion partner (usually derived from an *E. coli* gene) to the N-terminus of the protein of interest (reviewed by Waugh[9]). A side benefit of this approach is that the solubility fusion partners often increases the yield of protein. A variety of solubility fusion partners have been used to significantly increase the solubility of target proteins. Fusion partners include thioredoxin,[10] NusA,[11] glutathion-S-transferase (GST),[12] and the maltose binding protein (MBP).[13,14] These all appear to work for certain proteins, however, the best characterized and most effective appear to be NusA and MBP.[15–18]

Recently, newer solubility tags have been described including the ubiquitin-like molecule SUMO. The protein was studied in Dr. Christopher Lima's laboratory at Weill Medical College of Cornell University, where the complex between the *S. cerevisiae* SUMO (Smt3p) and its cognate protease Ulp1p was characterized.[19] During the course of this investigation, it was discovered that fusing recombinant proteins and peptides to Smt3p improved recombinant protein expression and solubility to the fusion partner. SUMO is an ideal fusion partner in that it is relatively small, highly soluble, and monomeric. Additionally, when cloned appropriately, the

SUMO moiety can be efficiently and specifically cleaved by the Ulp1 protease, resulting in a native recombinant protein. Independently, another group used the human SUMO to demonstrate increased solubility and expression levels (LifeSensors Inc., Malvern, PA).[20,21] One of the criticisms of using solubility tags is that many of the proteins convert into an insoluble aggregate as soon as they are cleaved from the fusion partner. No solubility tag is universal and not all tags work equally well and the exact mechanism for enhancing solubility is not known. However, it may be worthwhile to try and optimize for solubility by trying several different tags and determining which works best for that application.

Other approaches to reduce the likelihood of insoluble protein product include procedures that refold the protein. In general, insoluble proteins are denatured under reducing conditions and then resolubilized by removing the denaturing reagent through exchange against an assortment of refolding buffers. Although this method can be highly successful on an individual basis, conditions and refolding buffers are protein-dependent and are therefore not universal; making it relatively unattractive in an HTP workflow. Additional concerns include the loss of yield (obtaining less refolded product than the starting material) and not being sure that the final protein product represents a legitimate, native, and active structure even if it is "soluble."

HTP Methods for Detecting Protein Solubility

Although the methods mentioned above may increase the probability of obtaining a soluble protein product, many other proteins will remain insoluble. To quickly determine the solubility of an expressed protein, several methods have been developed for distinguishing the solubility of a sample.[2,18,22] Most of the HTP methods grow small liquid cultures and separate the proteins on the basis of standard separation techniques (e.g., centrifugation, affinity purification) and immunological detection and/or SDS-PAGE analysis. An alternative method has been developed where clones are screened directly from colonies on plates.[23] A filter is placed on top of a plate containing colonies containing expression constructs and is induced for expression by placing on a second plate containing LB + inducer. The cells are lysed and soluble proteins pass through the filter where they are bound to a nitrocellulose filter. After blocking and immunological detection (all the clones contain a common epitope tag), clones expressing soluble protein are detected and can be picked from the master plate for further expression. This method has a greater than 80% positive correlation with results obtained from traditional lysis and centrifugation methods and improves throughput, eliminates the need for centrifugation and SDS-PAGE.

Another recent approach has been designated "Pooled ORF Expression Technology" (POET),[24] and involves cloning and pooling hundreds of clones into a His-tagged vector and expressing them in a single tube. The mixed-clone cultures are expressed and soluble protein is identified and isolated by IMAC purification under native conditions. The purified proteins are then separated by 2-D electrophoresis which provides an estimate of the relative expression level. This is followed by picking individual spots for clone identification by mass spectrometry. After deconvoluting the spots, the proper clone can be identified and used for individual clone expression.

Although many of the steps are complex, the procedure for subcloning, expression, and purification are performed on pools of hundreds of ORFs and likely saves significant time vs. individual expression and conventional analysis. Additionally, many of the analysis steps for expression can be automated such as spot identification and picking, preparation of samples for MALDI-TOF/TOF, and peptide identification.

Finally, there is a method that does not require the use of any tag and exploits a unique set of genes that respond to translational misfolding. Promoters for these genes have been fused to reporter genes (*lacZ*) and are upregulated in response to misfolded heterologous protein. The optimal promoter was for the small heat shock protein *ibpA* which was fused to *lacZ* and used to monitor misfolding.[25] Using this approach, the reporter can differentiate between soluble, partially soluble, and insoluble recombinant proteins. However, it could be further developed to fuse the *ibpA* promoter to a lethal gene so as to select for only those clones that are capable of expressing soluble product.

Protein Purification

In addition to improved solubility, easy purification is a prerequisite for HTP expression and analysis. This is due to the many logistical challenges of HTP protein purification such as cell lysis, binding to affinity resins, washing, and elution, all of which may require optimization. To simplify purification, a vector encoded purification tag is usually fused to the protein. Commonly used tags include 6xHis, glutathione S-transferase (GST), STREP tag, Protein A, maltose binding protein, and the FLAG peptide (reviewed Waugh[9] and Lichty et al.[26]). These tags all exhibit high affinity and allow one-step purification by passing cell extracts or supernatants over their cognate matrices. Using many of these purification tags, several hundred human proteins expressed in *E. coli* were efficiently purified in high-throughput format.[2] They can also serve as epitope tags for immuno-detection and can be easily combined with solubility tags.[15]

An interesting strategy is to use the purification tag for direct attachment onto a microarray slide. This method involved fusing full-length *p53* clones to the *E. coli* biotin carboxyl carrier protein (BCCP),[27] which is biotinylated *in vivo* during expression. After the cultures are lysed and cleared, they are used to directly spot onto streptavidin-coated membranes or neutravidin-derivitised, dextran-coated slides. Although this method is efficient, the solubility of each protein is unclear.

YEAST EXPRESSION

Introduction

Ultimately, many eukaryotic proteins cannot be expressed in fully functional form in *E. coli*. This is especially true of secretory and transmembrane proteins that can require the oxidative environment of the eukaryotic secretory pathway for proper folding and disulfide bond formation. For this and other reasons, many high-throughput efforts rely on eukaryotic expression systems either entirely or to supplement *E. coli* work.

A number of yeast species provide an easy transition from bacterial to eukaryotic expression. Much of the equipment used for *E. coli* transformation, growth, and protein induction can be used interchangeably for yeast HTP work. The two most highly developed yeast species for HTP protein expression are *Saccharomyces cerevisiae* and *Pichia pastoris*. Both species can be easily transformed with circular or linear DNA molecules and possess effective *in vivo* homologous recombination pathways that allow stable, directed integration into their genomes.

S. cerevisiae has spent more generations under "domestication" by humans than any other organism, as a workhorse of what has become the food and beverage industry. Because of its economic importance, it was adopted early as a model genetic system and benefited from early uptake of recombinant DNA techniques. *S. cerevisiae* has two anomalous features that accelerated its manipulation with recombinant DNA molecules — small centromeric sequences that allowed development of episomal plasmids that are partitioned with high fidelity during cell division, and *in vivo* homologous recombination that is effective and efficient with short (20 to 50 bp) regions of DNA homology. The development of PCR DNA amplification in the 1980s allowed the *S. cerevisiae* research community to rapidly move to HTP and "whole genome" approaches to molecular and cellular biology. The early completion of the *S. cerevisiae* genome sequence allowed development of DNA microarray-based tools to globally analyze mRNA expression patterns.[28] Recombination with short homology arms allowed the generation of *S. cerevisiae* strain collections containing systematic gene deletions, and tagging of each ORF with GFP[29] and TAP tags.[30,31] In addition, researchers have used *in vivo* recombination to generate systematic collections of *S. cerevisiae* ORF expression constructs, enabling overexpression of nearly all *S. cerevisiae* proteins in the host system itself. One of these collections, encoding GST fusion proteins, provided the basis for the development of yeast protein microarrays.[32]

Early Heterologous Protein Expression in Yeast

With the development of recombinant DNA technology, *S. cerevisiae* almost immediately became a host system for heterologous protein expression.[33,34] Members of the *GAL* gene family were isolated in the late 1970s,[35] and the organization of the GAL1-10 cluster was elucidated shortly afterward.[36,37] This gene family provided a set of regulated promoters that could be up- or downregulated by modulating the carbon source of the yeast culture. This regulation was exploited to optimize the early expression of human insulin,[38] which failed to express from the unregulated *ADH1* promoter. Although a number of other regulated promoters have been used for heterologous protein expression in *S. cerevisiae,* Gal4p regulated promoters continue to be used and improved.[39]

Very early on, it became apparent that *S. cerevisiae* homologous recombination[40–43] could be used to bypass what, at the time, were tedious *in vitro* manipulation steps. Transformation of two DNA molecules into yeast would, under proper selection, result in homologous recombination to generate either episomal plasmids or genomic integrants.[33,44] DNA libraries could be cotransformed with yeast plasmids to create expression constructs without using restriction enzymes to perform "cut-and-paste"

operations *in vitro*. This approach was utilized to isolate the first human Cdc genes, by recombining a human cDNA library with a *Schizosaccharomyces pombe* vector backbone to complement *cdc2* mutations.[45]

Pichia pastoris Expression Systems

Pichia pastoris is one of a number of yeast species that is capable of growth using methanol as its sole carbon source. This ability was first utilized by Phillips Petroleum for biomass production, converting natural gas first to methanol by catalytic oxidation and subsequently to protein using *Pichia*. Because *P. pastoris* prefers respiratory growth, biomass could be produced at very high cell density in fermentation. Changing world markets resulted in *P. pastoris* never being an economical approach to biomass production for animal feed, but in the early 1980s Phillips saw an opportunity to exploit developments in *S. cerevisiae* technology to develop *P. pastoris* as an alternative expression system.

When switching from glucose or glycerol to methanol as its carbon source, *Pichia* requires the expression of a series of enzymes that oxidize the methanol first to formaldehyde and then to formic acid. The first enzyme in this pathway, alcohol oxidase (*AOX1*), is one of the most tightly regulated and strongly induced loci in any organism. Jim Cregg and colleagues created expression vectors containing the *AOX1* promoter for heterologous protein expression[46,47] that could be introduced into *P. pastoris* using slight modifications of existing *S. cerevisiae* techniques. In order to create *P. pastoris* strains that could easily be scaled up to fermentation for large scale protein production, their approach relied on genomic integration rather than episomal plasmids. *Pichia* has proved to be especially useful for producing proteins normally processed in higher eukaryotic secretory pathways. Like *S. cerevisiae*, proteins involved in membrane trafficking are highly homologous to their mammalian counterparts,[48] and native mammalian signal sequences and transmembrane domains are often correctly inserted into *Pichia* membranes, with proper disulfide bond formation. Jim Cregg at the Keck Graduate Institute maintains an updated list of proteins successfully expressed in *P. pastoris* (http://faculty.kgi.edu/cregg/index.htm). A recent high-profile success using *P. pastoris* expression was the structure elucidation of a mammalian *Shaker* gated ion channel by MacKinnon's group at Rockefeller University.[49]

Posttranslational Modifications

Both *S. cerevisiae* and *P. pastoris* perform a wide variety of protein posttranslational modifications (PTMs) that are similar or identical to those found in mammalian cells. In numerous cases, mammalian proteins can complement yeast mutations even in situations where proper functionality requires either static or dynamic PTMs. In contrast to modifications such as phosphorylation, ubiquitization, and isoprenylation, glycosylation in yeasts results in very different structures than in mammalian cells. While yeasts recognize the same protein sequence motif for N-linked glycosylation as mammalian cells and transfer an identical oligosaccharide core structure onto secretory proteins, mechanisms for the modification of this core structure evolved

very differently in yeast and mammals. In general, mammalian cells degrade about half of the mannose core and rebuild structures with diverse sugars, including GlcNAc, galactose, and sialic acids. Yeasts, on the other hand, tend to build on the mannose core by the addition of more mannose units, generating structures that can contain dozens of branched mannose rings.

Engineering the glycosylation pathway in yeast to more closely resemble that of mammalian cells has been a challenge for heterologous protein expression in both *S. cerevisiae* and *P. pastoris*. Oligosaccharide processing occurs in eukaryotic cells in a series of reactions performed by transmembrane, sugar transferase enzymes spatially distributed along the secretory pathway in the endoplasmic reticulum (ER) and the Golgi apparatus. In addition to the enzymes directly involved in oligosaccharide construction, nucleotide sugars need to be synthesized in the cytoplasm and transported through channels into the ER and Golgi. Subtle differences in localization mechanisms between yeast and mammals have usually resulted in poor activity of mammalian sugar transferases in yeast. Recently, however, Choi and colleagues have created combinatorial libraries of the enzymatic domains of mammalian sugar transferases with localization domains from yeast and expressed these libraries in *P. pastoris*.[50] This approach has been successful in generating strains of *Pichia* that produce N-linked oligosaccharides that closely match mammalian structures.[51] Using one *in vitro* "polishing" reaction, they were recently able to produce a human monoclonal antibody in *P. pastoris* with an identical oligosaccharide structure to the commercial product produced using mammalian cell bioreactors.[52]

HTP Yeast Expression

As discussed in the Introduction, HTP protein expression in yeast has been performed for whole genome yeast ORF collections.[32] Extending these protocols to heterologous expression of mammalian ORF collections is straightforward. *In vivo* homologous recombination has been used to clone human ORFs into a copper regulated *S. cerevisiae* expression system.[53] The same group has taken a similar approach with *P. pastoris*, although with a conventional, restriction enzyme-based protocol for the generation of expression vectors.[54] The recent development of Gateway™ vectors for *Pichia* simplifies the initial steps in this process.[55] *Pichia*, however, remains more difficult than *S. cerevisiae* to work with in a HTP format, since expression constructs have to be properly integrated into the genome and methanol induction is more finicky than galactose or copper induction. Recent advances in glycoengineering discussed above may, in many circumstances, make the added effort worthwhile.

One aspect of HTP protein expression and purification in yeast that needed to be solved was cell breakage. Yeast cells are much more difficult to break open than *E. coli*. The typical low throughput approach has been mechanical "crushing" using vigorous agitation in the presence of 0.5 mm glass beads. Commercial products are now available to perform glass bead breakage in 96-well plates (BioSpec Products, Inc.). Low throughout glass bead breakage has also been performed using agitation with a paint mixer typically found at local hardware store[56] and this protocol has been successfully extended to 96-deepwell plates.[32]

INSECT CELL EXPRESSION

Introduction

Since the first publication reporting the use of insect cells to express a heterologous gene,[57] use of baculovirus has become a routine method for protein expression. Eukaryotic proteins expressed using baculovirus are frequently soluble, correctly folded, and active, bypassing many of the problem points often encountered in bacterial expression. For example, baculovirus has been particularly useful for production of G-protein coupled receptors (GPCRs). G-protein coupled receptors are the largest single family of cell surface receptors involved in signal transduction, and thus are important therapeutic targets. However, GPCRs are generally expressed endogenously at very low levels. Typical GPCRs are large-membrane proteins containing seven transmembrane domains. Agonist binding to the receptor triggers phosphorylation of associated trimeric G-proteins. A variety of posttranslational modifications, such as palmitoylation, myristyloylation, prenylation, and carboxymethylation have been reported to be required for G-protein structure and activity and generally occur in baculovirus-infected insect cells as readily as they do in mammalian cells.[58] In general, yields of mammalian proteins are often high (100 to 500 mg/l culture), and because baculovirus does not replicate and is nonpathogenic in mammalian cells, the baculovirus expression system requires no extra safety precautions beyond general sterile tissue culture procedures.

Cloning for Insect Cell Expression

Baculoviruses are large enveloped DNA viruses that infect insects, primarily in the order *Lepidoptera*. The most studied and commonly used baculovirus for biotechnological applications is *Autographa californica* multi-nucleocapsid polyhedrovirus (AcMNPV).[59] The baculovirus life cycle involves two distinct morphological forms of the virus. The polyhedron derived virus (PDV) is responsible for transmission from insect to insect, whereas the budded form of the virus (BV) is responsible for viral transmission within individual insects. See Federici[60] and Williams and Faulkner[61] for details. Polyhedra are easily seen by light microscopy in cells as crystalline inclusion bodies. BV is the form of the virus that replicates in cell culture and is primarily used for heterologous protein expression. The polyhedrin and p10 genes are very highly expressed but are not required for BV transmission. Thus, most (but not all) biotechnology applications of baculovirus use polyhedrin or p10 promoters for expression of heterologous genes. Early use of baculovirus for heterologous gene expression required restriction enzyme cloning of the desired gene into a transfer vector downstream of the polyhedrin promoter, flanked by the viral sequences surrounding the polyhedrin locus of the virus. Following cotransfection of the transfer vector and wild-type virus DNA, homologous recombination across the flanking sequences created recombinant viruses that had to be identified by plaque assay screening for polyhedrin negative plaques.[57] Since recombination occurred at a frequency of less than 1%, creation of a useful recombinant virus stock required months of tedious plaque purification. Significant advances in baculovirus cloning technology came in the early 1990s

with the advent of linearized baculovirus DNA. A baculovirus genome was engineered to have a single *Bsu36* I site in the polyhedrin locus. Following recombination with the transfer vector, the recircularized DNA resulted in a 10-fold increase in the frequency of recombinants.[62] A few years later, vectors were created possessing multiple *Bsu36* I sites positioned such that linearization removed an essential gene that was rescued upon homologous recombination. A *lacZ* fragment added color selection, boosting the frequency of obtaining recombinant plaques to over 90% and made plaque identification easier.[63] These vectors were commercialized and the wide availability of these baculovirus vectors (BacMagic™, Merck KGaA, Darnstadt, Germany; BacPAK™, Takara Bio, Inc., Otsu, Japan; Sapphire™, Orbigen, San Diego, CA) no doubt explains the exponential increase in the number of published reports using baculovirus in the first half of the 1990s. In 1993, Luckow et al.[64] published a method for baculovirus recombination in bacteria that greatly shortened the time required to generate recombinant baculoviruses. A baculovirus genome was engineered to replicate in bacteria via a mini-F replicon (called a bacmid). The bacterial replicon contains a *lacZ* α reading frame containing attTn7 sites, allowing for site-specific transposition from a transfer vector that contains the gene of interest under polyhedrin (or other baculovirus promoter) control, flanked by Tn7 sites. The transposition activity is provided by a helper plasmid encoding the requisite transposase. Recombinant bacmids are selected by antibiotic selection and a color screen. The bacmid DNA is isolated and transfected into insect cells. The incidence of obtaining parental bacmid is low, and plaque purification is generally not required. Expression can be extended to protein complexes using a vector that contains nested cloning sites making possible the simultaneous cloning and expression of eight or more different genes.[65] The flexibility of the baculovirus expression system allows for expression of multicomponent protein assemblies from benchtop to bioreactor scale.

Recently, a baculovirus expression system, called BaculoDirect™ (Invitrogen), incorporating Gateway® cloning was developed. A recombinant baculovirus was created that contains a counter-selection cassette in the polyhedrin locus. The counter-selection cassette contains the *lacZ* α fragment under control of the late p10 promoter, and the thymidine kinase gene (TK) under control of the immediate early ie-0 promoter. The entry clone is recombined with linearized BaculoDirect DNA in a short room temperature reaction that removes the counterselection cassette. The reaction is transfected directly into insect cells, eliminating the *E. coli* manipulation steps necessary in other systems. The transfected insect cells are grown in the presence of ganciclovir, a nucleoside analog rendered toxic by the TK gene product.[66] Thus, replication of residual parental virus from the LR reaction is inhibited because it is both linear and expresses the TK gene. The *lacZ* gene is also recombined out providing a visual confirmation that parent virus was eliminated.

Posttranslational Modifications

One of the key attributes of the baculovirus expression system is that recombinant proteins are posttranslationally modified, often yielding protein that is correctly

folded and biologically active. Insect cells in culture perform most of the post-translational modifications typical of eukaryotes, including glycosylation, phosphorylation, sulfation, acylation, acetylation,[67] and possibly α-amidation.[68] With regards to glycosylation, baculovirus infected insect cells support both N- and O-linked glycosylation. Given the importance of proper glycosylation for the therapeutic utility of recombinant proteins, the N-linked glycosylation capabilities of insect cells have been studied extensively (reviewed by Jarvis[69,70]). Without terminal sialic acid residues, introduced glycoproteins are rapidly cleared from circulation by asialoglycoprotein receptors in the mammalian liver. Insect cells generally have a truncated N-linked processing pathway, resulting in paucimannosidic or high-mannose glycans lacking terminal sialic acid residues.[70] The truncated pathway is a result of diminishingly low levels of key Golgi enzymes in insect cells. Recently, insect cell line derivatives were produced that constitutively express several key mammalian glycosylation enzymes. A cell line expressing bovine β-1,4 galactosyltransferase produced glycans with terminal galactose, unlike the parent *Sf9* cell line.[71] A subsequent cell line expressing five glycosyl transferases produced complex mono-and bi-antennary complex glycans with terminal sialic acid residues[72] and has been commercialized (Mimic™ cells, Invitrogen). Sialylation occurs only when the Mimic cells were grown in media containing serum supplementation of serum-free media with fetuin likewise enabled sialylation, suggesting that the cells are able to scavenge sialic acid from proteins in serum.[73,74] Further metabolic engineering produced a cell line that had enhanced sialic acid processing capabilities resulting in higher levels of glycoprotein sialylation in serum free media.[73,74]

HTP Baculovirus Expression

Jumping from microbial expression systems such as *E. coli* or yeast to higher eukaryotic expression for an HTP pipeline introduces a number of complexities that need to be addressed. For baculovirus in particular, there are three specific areas where microbial techniques and equipment do not necessarily transfer easily. First, baculovirus recombinant DNA molecules are typically 10- to 20-fold larger than *E. coli* or yeast expression vectors. These are produced by recombination techniques, either *in vivo* (*E. coli* or insect cells) or *in vitro* (Gateway). Second, there is an intermediate step between cloning and expression that requires the production of baculoviral stocks that are difficult to titer in HTP format. Since baculovirus expression is a transient technique, lot-to-lot variation in viral titers and protein expression can be significant. Third, insect cell culture in HTP is more difficult than microbial cell growth and there are two cell culture steps (viral production and protein expression) that need to optimized for different endpoints. If one is looking to maximize functional mammalian protein production in a single expression system, using baculovirus has distinct advantages that make it worthwhile to address the difficult intermediate steps.

Albala and coworkers at the Lawrence Livermore National Laboratory have set up a baculovirus protein production system for expression of the I.M.A.G.E. clone collection (http://www.llnl.gov/tid/lof/documents/pdf/304834.pdf).[75] Their approach

uses Baculogold™ (Takara Bio Inc., Otsu, Japan) along with conventional rare restriction enzyme cloning (*Asc* I/*Fse* I) to generate expression clones. Recombination into the baculoviral genome is done in *Sf21* cells grown in normal 96-well plates, followed by protein expression in 96-well deep-well plates. A variety of incubators can be used for insect cell growth, although 96-well plates need an optimized radius in an orbital shaker for adequate oxygen transfer in deep wells. Albala's group has used a magnetic levitation stirrer (V&P Scientific, San Diego, CA), which moves a ball bearing vertically up and down through the culture to generate proper oxygenation.

CHOOSING EXPRESSION TECHNOLOGIES FOR A HTP PIPELINE

As described above, expression systems based on bacterial, fungal or insect each have advantages and disadvantages. Specific attributes of each system are summarized and compared in Table 2.1. No single system is ideal for all types of proteins, and care must be taken in balancing the strengths and weaknesses of each system when developing a HTP pipeline.

TABLE 2.1
Comparison of High-Throughput Expression Systems

Attribute	Expression System Ranking (best to worst)
Speed of expression	*E. coli* > yeast > baculovirus
Cloning complexity	*E. coli* > yeast > baculovirus
Protein yield	*E. coli* > yeast > baculovirus
Expense	*E. coli* < yeast < baculovirus
Ease of cell lysis	baculovirus > *E. coli* > yeast
Purification tags	All about equal
Native protein solubility	baculovirus = yeast > *E. coli*
Solubility tags	*E. coli* > yeast = baculovirus
Protein secretion (mammalian signal sequences)	yeast = baculovirus >> *E. coli*
Membrane proteins	baculovirus > yeast >> *E. coli*
Disulfide bond formation	baculovirus = yeast > *E. coli*
Protein complex formation	baculovirus > yeast > *E. coli*
N-linked glycosylation	*P. pastoris* = baculovirus >> *E. coli*
O-linked glycosylation	baculovirus > yeast > *E. coli*
Other PTMs	baculovirus > yeast > *E. coli*

Note: The three expression systems discussed in the chapter (*E. coli*, yeast, and baculovirus) are compared for a variety of attributes important for HTP cloning and expression.

REFERENCES

1. Hartley, J.L., Temple, G.F., and Brasch, M.A., DNA cloning using *in vitro* site-specific recombination, *Genome Res.*, 10, 1788, 2000.
2. Braun, P. et al., Proteome-scale purification of human proteins from bacteria, *Proc. Natl. Acad. Sci. USA*, 99, 2654, 2002.
3. Vincentelli, R. et al., Automated expression and solubility screening of His-tagged proteins in 96-well format, *Anal. Biochem.*, 346, 77, 2005.
4. Chen, J. et al., Enhancement of the solubility of proteins overexpressed in *Escherichia coli* by heat shock, *J. Mol. Microbiol. Biotechnol.*, 4, 519, 2002.
5. Ikura, K. et al., Co-overexpression of folding modulators improves the solubility of the recombinant guinea pig liver transglutaminase expressed in *Escherichia coli*, *Prep. Biochem. Biotechnol.*, 32, 189, 2002.
6. Wulfing, C. and Pluckthun, A., Correctly folded T-cell receptor fragments in the periplasm of *Escherichia coli*. Influence of folding catalysts, *J. Mol. Biol.*, 242, 655, 1994.
7. Georgiou, G. and Valax, P., Expression of correctly folded proteins in *Escherichia coli*, *Curr. Opin. Biotechnol.*, 7, 190, 1996.
8. Studier, F.W., Protein production by auto-induction in high density shaking cultures, *Protein Expr. Purif.*, 41, 207, 2005.
9. Waugh, D.S., Making the most of affinity tags, *Trends Biotechnol.*, 23, 316, 2005.
10. Lavallie, E.R. et al., A thioredoxin gene fusion expression system that circumvents inclusion body formation in the *E. coli* cytoplasm, *Biotechnology (NY)*, 11, 187, 1993.
11. Davis, G.D. et al., New fusion protein systems designed to give soluble expression in *Escherichia coli*, *Biotechnol. Bioeng.*, 65, 382, 1999.
12. Smith, D.B. and Johnson, K.S., Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase, *Gene*, 67, 31, 1988.
13. Bedouelle, H. and Duplay, P., Production in *Escherichia coli* and one-step purification of bifunctional hybrid proteins which bind maltose. Export of the Klenow polymerase into the periplasmic space, *Eur. J. Biochem.*, 171, 541, 1988.
14. Di Guan, C. et al., Vectors that facilitate the expression and purification of foreign peptides in *Escherichia coli* by fusion to maltose-binding protein, *Gene*, 67, 21, 1988.
15. Busso, D., Delagoutte-Busso, B., and Moras, D., Construction of a set Gateway-based destination vectors for high-throughput cloning and expression screening in *Escherichia coli*, *Anal. Biochem.*, 343, 313, 2005.
16. Hammarstrom, M. et al., Rapid screening for improved solubility of small human proteins produced as fusion proteins in *Escherichia coli*, *Protein Sci.*, 11, 313, 2002.
17. Kapust, R.B. and Waugh, D.S., *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused, *Protein Sci.*, 8, 1668, 1999.
18. Shih, Y.P. et al., High-throughput screening of soluble recombinant proteins, *Protein Sci.*, 11, 1714, 2002.
19. Mossessova, E. and Lima, C.D., Ulp1-sumo crystal structure and genetic analysis reveal conserved interactions and a regulatory element essential for cell growth in yeast, *Mol. Cell*, 5, 865, 2000.
20. Malakhov, M.P. et al., SUMO fusions and SUMO-specific protease for efficient expression and purification of proteins, *J. Struct. Funct. Genomics*, 5, 75, 2004.
21. Marblestone, J.G. et al., Comparison of SUMO fusion technology with traditional gene fusion systems: Enhanced expression and solubility with SUMO, *Protein Sci.*, 15, 182, 2006.

22. Waldo, G.S., Genetic screens and directed evolution for protein solubility, *Curr. Opin. Chem. Biol.*, 7, 33, 2003.
23. Cornvik, T. et al., Colony filtration blot: A new screening method for soluble protein expression in *Escherichia coli*, *Nat. Methods*, 2, 507, 2005.
24. Gillette, W.K. et al., Pooled ORF expression technology (POET): Using proteomics to screen pools of open reading frames for protein expression, *Mol. Cell Proteomics*, 4, 1647, 2005.
25. Lesley, S.A. et al., Gene expression response to misfolded protein as a screen for soluble recombinant protein, *Protein Eng.*, 15, 153, 2002.
26. Lichty, J.J. et al., Comparison of affinity tags for protein purification, *Protein Expr. Purif.*, 41, 98, 2005.
27. Boutell, J.M. et al., Functional protein microarrays for parallel characterisation of p53 mutants, *Proteomics*, 4, 1950, 2004.
28. Lashkari, D.A. et al., Yeast microarrays for genome wide parallel genetic and gene expression analysis, *Proc. Natl. Acad. Sci. USA*, 94, 13057, 1997.
29. Huh, W.K. et al., Global analysis of protein localization in budding yeast, *Nature*, 425, 686, 2003.
30. Gavin, A.C. et al., Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature*, 415, 141, 2002.
31. Krogan, N.J. et al., Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*, *Nature*, 440, 637, 2006.
32. Zhu, H. et al., Global analysis of protein activities using proteome chips, *Science*, 293, 2101, 2001.
33. Beggs, J.D., Transformation of yeast by a replicating hybrid plasmid, *Nature*, 275, 104, 1978.
34. Beggs, J.D. et al., Abnormal expression of chromosomal rabbit beta-globin gene in *Saccharomyces cerevisiae*, *Nature*, 283, 835, 1980.
35. St. John, T.P. and Davis, R.W., Isolation of galactose-inducible DNA sequences from *Saccharomyces cerevisiae* by differential plaque filter hybridization, *Cell*, 16, 443, 1979.
36. St. John, T.P. and Davis, R.W., The organization and transcription of the galactose gene cluster of *Saccharomyces*, *J. Mol. Biol.*, 152, 285, 1981.
37. St. John, T.P. et al., Deletion analysis of the *Saccharomyces GAL* gene cluster. Transcription from three promoters, *J. Mol. Biol.*, 152, 317, 1981.
38. Stepien, P.P. et al., Synthesis of a human insulin gene. VI. Expression of the synthetic proinsulin gene in yeast, *Gene*, 24, 289, 1983.
39. Sil, A.K., Xin, P., and Hopper, J.E., Vectors allowing amplified expression of the *Saccharomyces cerevisiae* Gal3p-Gal80p-Gal4p transcription switch: Applications to galactose-regulated high-level production of proteins, *Protein Expr. Purif.*, 18, 202, 2000.
40. Orr-Weaver, T.L. and Szostak, J.W., Yeast recombination: The association between double-strand gap repair and crossing-over, *Proc. Natl. Acad. Sci. USA*, 80, 4417, 1983.
41. Orr-Weaver, T.L. and Szostak, J.W., Fungal recombination, *Microbiol. Rev.*, 49, 33, 1985.
42. Orr-Weaver, T.L., Szostak, J.W., and Rothstein, R.J., Yeast transformation: A model system for the study of recombination, *Proc. Natl. Acad. Sci. USA*, 78, 6354, 1981.
43. Szostak, J.W. et al., The double-strand-break repair model for recombination, *Cell*, 33, 25, 1983.

44. Struhl, K. et al., High-frequency transformation of yeast: Autonomous replication of hybrid DNA molecules, *Proc. Natl. Acad. Sci. USA*, 76, 1035, 1979.
45. Lee, M.G. and Nurse, P., Complementation used to clone a human homologue of the fission yeast cell cycle control gene cdc2, *Nature*, 327, 31, 1987.
46. Cregg, J.M. et al., *Pichia pastoris* as a host system for transformations, *Mol. Cell Biol.*, 5, 3376, 1985.
47. Tschopp, J.F. et al., Expression of the *lacZ* gene from two methanol-regulated promoters in *Pichia pastoris*, *Nucleic Acids Res.*, 15, 3859, 1987.
48. Payne, W.E. et al., Isolation of *Pichia pastoris* genes involved in ER-to-Golgi transport, *Yeast*, 16, 979, 2000.
49. Long, S.B., Campbell, E.B., and Mackinnon, R., Crystal structure of a mammalian voltage-dependent *Shaker* family K+ channel, *Science*, 309, 897, 2005.
50. Choi, B.K. et al., Use of combinatorial genetic libraries to humanize N-linked glycosylation in the yeast *Pichia pastoris*, *Proc. Natl. Acad. Sci. USA*, 100, 5022, 2003.
51. Hamilton, S.R. et al., Production of complex human glycoproteins in yeast, *Science*, 301, 1244, 2003.
52. Li, H. et al., Optimization of humanized IgGs in glycoengineered *Pichia pastoris*, *Nat. Biotechnol.*, 24, 210, 2006.
53. Holz, C. and Lang, C., High-throughput expression in microplate format in *Saccharomyces cerevisiae*, *Methods Mol. Biol.*, 267, 267, 2004.
54. Bottner, M. and Lang, C., High-throughput expression in microplate format in *Pichia pastoris*, *Methods Mol. Biol.*, 267, 277, 2004.
55. Esposito, D. et al., Gateway cloning is compatible with protein secretion from *Pichia pastoris*, *Protein Expr. Purif.*, 40, 424, 2005.
56. Crosby, B. et al., Purification and characterization of a uracil-DNA glycosylase from the yeast, *Saccharomyces cerevisiae*, *Nucleic Acids Res.*, 9, 5797, 1981.
57. Smith, G.E., Summers, M.D., and Fraser, M.J., Production of human beta interferon in insect cells infected with a baculovirus expression vector, *Mol. Cell Biol.*, 3, 2156, 1983.
58. Massotte, D., G protein-coupled receptor overexpression with the baculovirus-insect cell system: A tool for structural and functional studies, *Biochim. Biophys. Acta*, 1610, 77, 2003.
59. Ayres, M.D. et al., The complete DNA sequence of *Autographa californica* nuclear polyhedrosis virus, *Virology*, 202, 586, 1994.
60. Federici, B.A., Baculovirus pathogenesis, in *The Baculoviruses*, Miller, L.K., Ed., Plenum Press, New York, 1997, p. 33.
61. Williams, G.V. and Faulkner, P., Cytological changes and viral morphogenesis during baculovirus infection, in *The Baculoviruses*, Miller, L.K., Ed., Plenum Press, New York, 1997, p. 61.
62. Kitts, P.A., Ayres, M.D., and Possee, R.D., Linearization of baculovirus DNA enhances the recovery of recombinant virus expression vectors, *Nucleic Acids Res.*, 18, 5667, 1990.
63. Kitts, P.A. and Possee, R.D., A method for producing recombinant baculovirus expression vectors at high frequency, *BioTechniques*, 14, 810, 1993.
64. Luckow, V.A. et al., Efficient generation of infectious recombinant baculoviruses by site-specific transposon-mediated insertion of foreign genes into a baculovirus genome propagated in *Escherichia coli*, *J. Virology*, 67, 4566, 1993.
65. Berger, I., Fitzgerald, D.J., and Richmond, T.J., Baculovirus expression system for heterologous multiprotein complexes, *Nat. Biotechnol.*, 22, 1583, 2004.

66. Godeau, F., Saucier, C., and Kourilsky, P., Replication inhibition by nucleoside analogues of a recombinant *Autographa californica* multicapsid nuclear polyhedrosis virus harboring the herpes thymidine kinase gene driven by the ie-1(0) promoter: A new way to select recombinant baculoviruses, *Nucleic Acids Res.*, 20, 6239, 1992.
67. Jarvis, D.L., Baculovirus expression vectors, in *The Baculoviruses*, Miller, L.K., Ed., Plenum Press, New York, 1997, p. 389.
68. Suzuki, K. et al., Elucidation of amidating reaction mechanism by frog amidating enzyme, peptidylglycine alpha-hydroxylating monooxygenase, expressed in insect cell culture, *EMBO J.*, 9, 4259, 1990.
69. Jarvis, D.L., Modifying insect cell glycosylation pathways with baculovirus expression vectors, WO 98/06835, 1998.
70. Jarvis, D.L., Developing baculovirus-insect cell expression systems for humanized recombinant glycoprotein production, *Virology*, 310, 1, 2003.
71. Hollister, J.R., Shaper, J.H., and Jarvis, D.L., Stable expression of mammalian β-1,4-galactosyltransferase extends the N-glycosylation pathway in insect cells, *Glycobiology*, 8, 473, 1998.
72. Hollister, J. et al., Engineering the protein N-glycosylation pathway in insect cells for production of biantennary, complex N-glycans, *Biochemistry*, 41, 15093, 2002.
73. Hollister, J., Conradt, H., and Jarvis, D.L., Evidence for a sialic acid salvaging pathway in *Lepidopteran* insect cells, *Glycobiology*, 13, 487, 2003.
74. Aumiller, J.J., Hollister, J.R., and Jarvis, D.L., A transgenic insect cell line engineered to produce CMP-sialic acid and sialylated glycoproteins, *Glycobiology*, 13, 497, 2003.
75. Gilbert, M., Edwards, T.C., and Albala, J.S., Protein expression arrays for proteomics, *Methods Mol. Biol.*, 264, 15, 2004.

3 Emerging Trend: Cell-Free Protein Expression

Federico Katzen and Wieslaw Kudlicki

CONTENTS

INTRODUCTION

With the advent of the proteomics era, the cell-free expression field has experienced a technical renaissance expanding into a myriad of applications covering both functional and structural proteomics. Cell-free systems offer several advantages over traditional cell-based expression methods, including the easy modification of reaction conditions to favor protein folding, decreased sensitivity to product toxicity, and suitability for high-throughput strategies owing to the ability to reduce reaction volumes and process time. Moreover, improvements in translation efficiency have resulted in yields that exceed a milligram of protein per milliliter of reaction. Finally, the ability to easily manipulate the reaction components and conditions makes *in vitro* protein synthesis especially amenable to automation and miniaturization, enabling application to the fields of protein arrays, *in vitro* evolution, and multiplexed real-time labeling among others. We review the advances on this expanding technology and highlight the growing list of associated applications for protein microarrays. For further details we suggest the reader to refer to previously published literature.[1–5]

CONFIGURATIONS AND HISTORY

In vitro translation systems are based on the early demonstration that cell integrity is not required for protein synthesis to occur. In its simplest form, this can be accomplished by the use of a crude lysate from any given organism (which provides the translational machinery, accessory enzymes, tRNA, and factors) in combination with exogenously added RNA template, amino acids, and an energy supply. This classical *in vitro* translation scheme is called "uncoupled" in opposition to the "coupled" or "combined" transcription/translation configuration in which the mRNA is transcribed *in situ* from a DNA template added to the reaction (see[3] for more details). Usually coupled systems exhibit higher protein yields and are easier and faster to operate than systems that are not coupled, although they require supplementing the reaction with additional NTPs and a highly processive RNA polymerase such as those encoded by T7, T3, or SP6 bacteriophages. The use of plasmid or PCR templates rather than purified mRNAs has made possible the emergence of a variety of new applications.

Sources of Lysates

Almost any organism could potentially be used as a source for the preparation of a cell-free protein expression system. However, the most popular are those based on *Escherichia coli*, wheat germ, and rabbit reticulocytes (for a review, see Jermutus et al.[3]). *E. coli*–based systems provide yields that range from a few micrograms up to several milligrams per milliliter of reaction depending on the protein and the reaction format.[7] On the other end, eukaryotic-based systems provide a better platform for functional studies, particularly for post-translationally modified proteins. However, yields for these types of systems are in the microgram per milliliter of reaction range. The wheat germ-based translation system is of special interest due to its eukaryotic nature and robustness. Yields can go up to a few hundred of micrograms per milliliter of reaction. Other systems include the use of cell-free extracts derived from insect cells,[8] HeLa cells,[9] and yeast.[10] An advantage that the wheat germ and rabbit reticulocytes systems have over other eukaryotic cell-free systems is that they efficiently translate mRNAs in which the 5' cap has been replaced by an internal ribosome entry site (for a recent example see Shaloiko et al.[11]).

Despite being one of the most complicated basic cellular processes, the whole translational mechanism from *E. coli* can be reconstituted *in vitro* starting from > 100 individually purified components.[12] The system (called the PURE system) exhibits high translational efficiency, with the added advantage of simpler manipulation of the reaction conditions and easy purification of untagged protein products. Also, a eukaryotic translation elongation system could be reconstituted *in vitro* by the assembly of ribosomes onto dicistrovirus genomes that do not require the presence of aminoacylated initiator tRNAs.[13]

Yield and Throughput

The principal limitation of the first generation batch-formatted reactions is their short lifetimes (less than an hour) and consequent low yield. This is primarily owing to the rapid depletion of the high-energy phosphate pool, which occurs even in the absence

of protein synthesis.[14] In turn, this leads to the accumulation of free phosphate, which can complex with magnesium to further inhibit protein synthesis. This problem was first overcome by Spirin and coworkers with the introduction of the continuous-flow cell-free (CFCF) translation system, which relies on the continuous supply of energy and substrates and the continuous removal of the reaction byproducts.[15] The reaction time can then be extended for 20 hours with increases in product yield of up to two orders of magnitude. Despite the improvement in yield, the operational complexities make this system extremely impractical. The technology was later simplified by the development of a semicontinuous or continuous exchange cell-free (CECF) method, in which a passive rather than active exchange of substrates and byproducts extended the reaction lifetime.[16,17] However, semicontinuous systems are not easily applicable to high-throughput processes, which require miniaturization and automation.

To this end, several laboratories have focused either on developing high-throughput friendly systems or maximizing the energetics of batch reactions. Endo and coworkers have devised a highly efficient bilayer diffusion system devoid of membranes that is compatible with high-throughput formats.[6,18] On the other hand, the group of Swartz has consistently sought more efficient and economical alternatives to the traditional ATP/GTP regeneration systems. They have recently developed an economical method for cell-free protein synthesis using glucose and nucleoside monophosphates, reducing substantially the cost while supplying high protein yields.[19] In addition, they have demonstrated that with the use of a thin film reactor configuration it is possible to produce close to a milligram of protein per milliliter of reaction, overcoming issues related to scaling-up batch reactions.[20] In our laboratory, we have developed an alternative technology based on discrete "feeds" that replenish the reaction with the necessary substrates while diluting toxic byproducts. Milligram amounts of protein products can be obtained in small tubes without the use of any special equipment (Figure 3.1). The various approaches to cell-free expression are summarized in Figure 3.2.

FOLDING AND POSTTRANSLATIONAL MODIFICATIONS

A key goal for cell-free translation systems is to synthesize biologically active proteins. Currently, the primary issues are protein folding and posttranslational modifications. A clear advantage that these systems have over *in vivo* protein synthesis is that the environmental conditions can be easily adjusted. Strategies to improve protein folding and posttranslational processing include the addition of a variety of reagents and folding catalysts to the reaction.

Chaperones

Reports on the exogenous supply of chaperones to cell-free protein synthesis reactions suggest that the effect that these catalysts have is protein-dependant. For example, addition of purified DnaK, DnaJ, GroEL, and GroES has been reported to be beneficial for the synthesis of single chain and Fab antibodies[21,22] but appears to have no effect on the folding or activity of luciferase.[23] It has been demonstrated

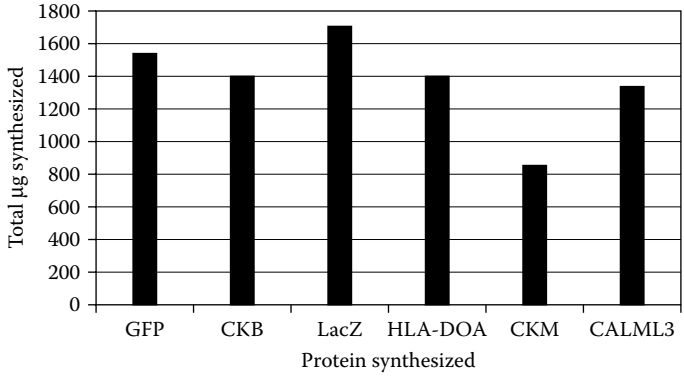


FIGURE 3.1 Efficient batch mode for the *in vitro* synthesis of milligram amount of proteins. Standard 1-ml cell-free reactions were performed for six hours at 37°C, using Expressway Milligram (Invitrogen, Carlsbad, CA) with two subsequent additions of 0.5 ml of feeding buffer at 30 minutes and 2 hours. Abbreviations: GFP, green fluorescent protein; CKB, human creatine kinase from brain; LacZ, *E. coli* ß-galactosidase; HLA-DOA of the human major histocompatibility complex (-chain); CKM, human creatine kinase from muscle; CALML3, human calmodulin-like 3 protein; IL24, human interleukin 24.

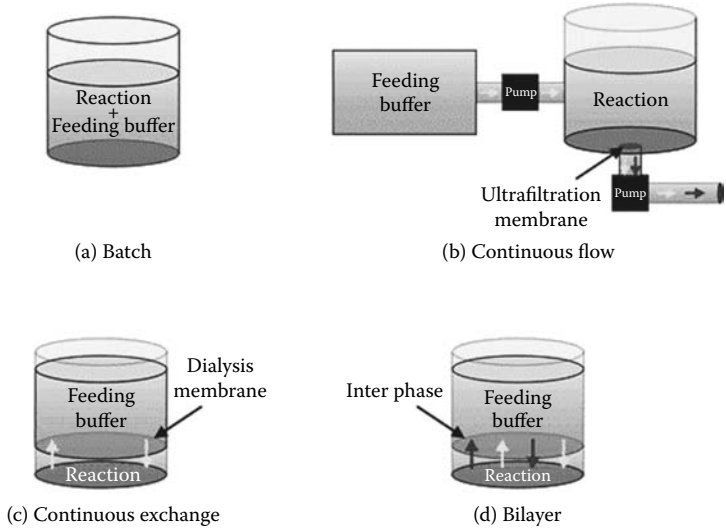


FIGURE 3.2 Current formats of cell-free protein expression systems. Formats are classified according to how the reaction is fed: (a) batch (b) continuous-flow cell-free (c) continuous exchange cell-free and (d) bilayer. Reaction components include ribosome, translation factors, tRNA, aminoacyl tRNA synthetases, template (RNA or DNA), and RNA polymerase (when necessary). The feeding buffer includes amino acids, energy components, NTPs (when necessary), cofactors, and other accessory reagents. Yellow arrows indicate the flow of buffer components and red arrows represent the flow of protein product.

that the use of S30 extract prepared from strains transformed with genes encoding the major chaperones and disulfide bond isomerases produces a cell-free system capable of expressing active eukaryotic proteins, thus eliminating the need for addition of purified folding catalysts.[24]

Membrane Proteins

Over-expression of membrane proteins *in vivo* frequently results in cell toxicity (owing to hydrophobicity or inherent properties of the target), protein aggregation, misfolding, and low yield. Nearly all of these obstacles can be overcome by cell-free expression.

In vitro translation offers a unique opportunity to use the highly efficient bacterial transcription and translational machinery while introducing natural mammalian or other synthetic lipids and detergents. For example, milligram amounts of active transmembrane multidrug transporters has recently been synthesized *in vitro*.[25,26] The presence of mild detergents or lipid mixtures during the reaction notably eased aggregation and insolubility issues and apparently did not interfere with the translation activity. Also, the oligomeric ion channel MscL could be synthesized *in vitro*[27] in a form that is undistinguishable from the one produced *in vivo*.

Ueda and coworkers have adapted the PURE system (see above) for the expression of membrane proteins. Basically the addition of the translocon machinery present in inverted vesicles provided the means for the production of correctly folded integral membrane proteins.[28]

Finally, it has recently been shown that a cell-free expression system can be encapsulated in phospholipid vesicles to build cell-like bioreactors. This configuration prolongs membrane and nonmembrane protein expression for up to 5 hours opening up new avenues of research and generating novel downstream biotechnological applications.[29]

Disulfide Bond Formation

Disulfide-bonded proteins are ordinarily formed in extracytoplasmic compartments, such as the periplasm of prokaryotes and the lumen of the endoplasmic reticulum (ER) of eukaryotes, in which conditions are more oxidizing. Although cell-free protein systems bear two intrinsic features that may prevent the formation of disulfide bonds (reducing agents that stabilize the protein synthesis machinery, and the lack of compartments with oxidizing redox potential), these hurdles can be easily overcome. One method eliminates dithiotreitol from the cell-free extract prior to the translation reaction, which has been shown to result in high yield production of single-chain antibodies with dual disulfide bonds.[30] Also, the combination of alkylation of the extract with iodoacetamide, a suitable glutathione redox buffer and a disulfide bond isomerase added to the *in vitro* reaction can have a profound positive effect on the production of active proteins with multiple disulfide bonds.[21,31]

Glycosylation and Other Posttranslational Modifications

Glycosylation is the most widespread and complex form of posttranslational modification in eukaryotes (for a review see Lowe and Marth[32]). A major problem for

the production of glycoproteins is that they are ordinarily produced as a mixture of glycoforms. Only the glycosidic core remains relatively uniform while the protein is in transit within the ER. In cell free systems, core glycosylation can be achieved by supplementing extracts with microsomal fractions.[33] Proteins are translocated to the lumen of the vesicles in which their leader peptide is cleaved and they acquire the oligosaccharide chain. Given that intracellular transport is disrupted, further processing of the oligosaccharides is prevented. However some variation on the glycosylation pattern can still be observed due to inhomogeneous folding that apparently restricts the access of the glycosylating enzymes.[34] Recently, the generation of a Spodoptera frugiperda 21 cell-based lysate has been reported.[8] The system provides core protein glycosylation enzymes without the need for supplementing the reaction with membrane vesicles. The newest approach for the *in vivo* synthesis of homogeneous samples of glycoproteins exploits the use of a nonnatural amino acid linked to a monosaccharide moiety.[35] This strategy could be easily adapted to cover *in vitro* protein synthesis. When incorporated into a protein, the monoglycosylated amino acid can be further modified by glycosyltransferases added to the *in vitro* reaction resulting in more complex glycoforms.

Other post-translational modifications, such as phosphorylation, myristylation, farnsylation, isoprenylation, and adenylation have been observed in lysates from higher eukaryotes. With all these modifications, the dynamic complexity of post-translational modifications makes it difficult to produce homogeneous protein samples. Methods for creating artificial posttranslational modification mimics (artificial modifications that imitate the structure of the natural ones) have been proposed as a solution for this problem (for a review see Davis[37]). Cell-free systems appear to be the most favorable platform for this novel strategy.

Finally, a technique that makes use of cell-free translation for dissecting components involved in the ubiquitin–proteasome pathway has been recently reported. Basically, substrates of this pathway can be isolated in ER membranes which, when incubated in RRL lacking exogenous hemin, are degraded in an ATP-dependent manner.[38]

Solubility Tags

According to data from several proteomics centers, more than half of all recombinant proteins are insoluble when they are overproduced in *E. coli*. For some of these cases, it has been shown that certain affinity tags have the ability to promote the solubility of their fusion partners (for a recent review see Waugh[39]). This strategy has the added bonus of facilitating the purification of the passenger protein and in other cases (as represented by the use of the Lumio tag), expediting the detection (in-gel or real-time) of the fusion protein.[1,40] Examples of solubility tags include glutathione S-transferase, maltose binding protein, thioredoxin, SUMO, and NusA, among others. Placing the solubility tag at the N-terminus of the protein has the advantage of providing an optimum context for translation initiation increasing the yield of recombinant proteins.

It is worth mentioning that not every protein can be made soluble simply by the incorporation of a solubility-enhancing tag. Also, some proteins will become insoluble

once the tag is removed. However it is clear that this strategy in general leads to the recovery of more soluble and properly folded proteins when compared to the expression of native proteins.

As no affinity tag is ideal for every single situation, combinatorial tagging sometimes has been recommended (for a recent example see Dyson et al.[41]). This approach has also proven successful in combination with the Gateway recombinational cloning technology.[42]

Unnatural Amino Acids and *in Vitro* Protein Labeling

The incorporation of nonnatural amino acids, especially those with chemically or physically reactive side chains, has the potential to be a useful tool for functional and structural proteomics. A variety of labels including fluorescent dyes for functional studies, biotinylated moieties to facilitate purification, and many others including those for structural studies and for posttranslational modifications can be sequence-specifically incorporated into proteins. An efficient way to incorporate artificial amino acids into polypeptides is to supplement the cell free extracts with chemically aminoacylated suppressor tRNAs that recognize a particular stop codon[43] or by reconstructing the genetic code *de novo* using the PURE approach.[44] This fascinating technology has been already applied to the synthesis of nonribosomal peptides by reassigning 35 of the 61 sense codons to 12 unnatural amino acid analogues.[45] It has been recently reported that suppression of the amber codon in cell-free translation systems can be enhanced by *in situ* deactivation the release factor 1 with specific antibodies.[46] A similar technology was applied to incorporate a single label at the N-terminal position, highly desirable for the preparation of protein micro arrays. This has been accomplished by using an amber initiator suppressor tRNA and a DNA template with an amber codon instead of the normal initiation codon.[47]

In vitro cotranslational labeling is not limited to the use of unnatural amino acids. For example, puromycin derivatives can be used in cell-free expression systems to specifically label proteins at the C-terminus.[48] Recently, a novel tetracysteine motif was shown to specifically bind biarsenical ligands that become fluorescent only after binding.[49] Using a fluorometer, these compounds have been directly added to cell-free transcription-translation systems to monitor real-time protein synthesis in high-throughput expression format. This approach is particularly useful for high-throughput screening of pharmacological agents with translation-inhibiting activity. Although some of these labeling techniques can be applied to cell-based systems, problems such as cytotoxicity of the compounds, reduced protein yields, low label incorporation, or transport across membranes are issues largely reduced or eliminated when using a cell-free expression system.

CONSIDERATIONS FOR PROTEIN ARRAYS

A clear application of cell-free protein expression reactions is in the area of miniaturization and protein arrays. For example, protein "macro" arrays can be generated by small 25 μl reactions to synthesize tagged products that are *in situ* immobilized

in separate wells coated with tag-binding beads.[50] Reactions can be downscaled to levels (nanoliter scale) that are unimaginable for cell-based approaches[51] and yet still synthesize enough products to perform individual enzymatic assays in 96-well glass microplates. Also, coupled transcription and translation using a solid phase DNA template on 96-well plates has been reported recently.[52]

Finally, cell free protein synthesis offers tremendous advantages to the construction of protein micro arrays. One of the first reports of the use of cell-free protein expression for protein array assembly describes the use of parallel cell-free reactions following immobilization on a surface.[53] More recently Ramachandran and coworkers developed a self-assembling protein chip starting with DNA gene micro arrays, which are transcribed and translated by a cell-free system. The resulting proteins, fused to glutathione S-transferase (GST), are immediately captured *in situ* by virtue of an antibody anti GST printed simultaneously with the expression plasmid.[54] This technique saves considerable labor, time, and costs by eliminating the need to express, purify and print proteins separately.

FUTURE PERSPECTIVES

Most of the advantages that cell-free expression systems have to offer can only be attained by high productive batch-fed configurations. Although protein concentrations up to a milligram per milliliter of reaction can be now achieved, this is still not enough for certain applications. But there is plenty of room for improvement. For example the incorporation of membrane vesicles loaded with the oxidative phosphorylation enzymes might have a positive effect by recycling ADP and lowering the free phosphate contents.[55] Finding a high efficient energy regeneration system is also a key-issue for lowering the costs of this still pricey technology.

Another area that cell-free can make a significant impact is protein folding. A relatively high fraction of proteins obtained by *in vitro* and *in vivo* systems is usually insoluble or misfolded. The addition of detergents or chaperones to the reaction sometimes has a productive effect but there might be complementary approaches as well. For instance, hybrid systems composed of lysates from different sources, including those from archaea, might provide a more robust folding context.

Cell-free expression is a powerful, flexible, and ever-expanding technology. The ability to manipulate the reaction conditions and to generate novel applications will probably be limited only by our creativity.

REFERENCES

1. Katzen, F., Chang, G., and Kudlicki, W., The past, present and future of cell-free protein synthesis, *Trends Biotechnol.*, 23, 150, 2005.
2. Spirin, A.S., High-throughput cell-free systems for synthesis of functionally active proteins, *Trends Biotechnol.*, 22, 538, 2004.
3. Jermutus, L., Ryabova, L.A., and Pluckthun, A., Recent advances in producing and selecting functional proteins by using cell-free translation, *Curr. Opin. Biotechnol.*, 9, 534, 1998.
4. Swartz, J.R., *Cell-Free Protein Expression,* Springer, Berlin, 2003.

5. Jackson, A.M. et al., Cell-free protein synthesis for proteomics, *Brief Funct. Genomic Proteomic*, 2, 308, 2004.
6. Endo, Y. and Sawasaki, T., Advances in genome-wide protein expression using the wheat germ cell-free system, *Methods Mol. Biol.*, 310, 145, 2005.
7. Kigawa, T. et al., Cell-free production and stable-isotope labeling of milligram quantities of proteins, *FEBS Lett.*, 442, 15, 1999.
8. Katzen, F. and Kudlicki, W. Efficient generation of insect-based cell-free translation extracts active in glycosylation and signal sequence processing, *J. Biotechnol.* 125, 194, 2006.
9. Mikami, S. et al., An efficient mammalian cell-free translation system supplemented with translation factors, *Protein Expr. Purif.*, 2005.
10. Wang, Z., Controlled expression of recombinant genes and preparation of cell-free extracts in yeast, *Methods Mol. Biol.*, 313, 317, 2006.
11. Shaloiko, L.A., Granovsky, I.E., Ivashina, T.V., Ksenzenko, V.N., et al., Effective non-viral leader for cap-independent translation in a eukaryotic cell-free system, *Biotechnol. Bioeng.*, 88, 730, 2004.
12. Shimizu, Y., Kanamori, T., and Ueda, T., Protein synthesis by pure translation systems, *Methods*, 36, 299, 2005.
13. Pestova, T.V. and Hellen, C.U., Reconstitution of eukaryotic translation elongation *in vitro* following initiation by internal ribosomal entry, *Methods*, 36, 261, 2005.
14. Kim, D.M. and Swartz, J.R., Prolonging cell-free protein synthesis with a novel ATP regeneration system, *Biotechnol. Bioeng.*, 66, 180, 1999.
15. Spirin, A.S. et al., A continuous cell-free translation system capable of producing polypeptides in high yield, *Science*, 242, 1162, 1988.
16. Alakhov, Y.B. et al., Method of Preparing Polypeptides in a Cell-Free Translation System, U.S. Patent 5478730, 1995.
17. Kim, D.M. and Choi, C.Y., A semicontinuous prokaryotic coupled transcription/translation system using a dialysis membrane, *Biotechnol. Prog.*, 12, 645, 1996.
18. Sawasaki, T. et al., A bilayer cell-free protein synthesis system for high-throughput screening of gene products, *FEBS Lett.*, 514, 102, 2002.
19. Calhoun, K.A. and Swartz, J.R., An economical method for cell-free protein synthesis using glucose and nucleoside monophosphates, *Biotechnol. Prog.*, 21, 1146, 2005.
20. Voloshin, A.M. and Swartz, J.R., Efficient and scalable method for scaling up cell free protein synthesis in batch mode, *Biotechnol. Bioeng.*, 91, 516, 2005.
21. Ryabova, L.A. et al., Functional antibody production using cell-free translation: effects of protein disulfide isomerase and chaperones, *Nat. Biotechnol.*, 15, 79, 1997.
22. Jiang, X. et al., Expression of Fab fragment of catalytic antibody 6D9 in an *Escherichia coli in vitro* coupled transcription/translation system, *FEBS Lett.*, 514, 290, 2002.
23. Kolb, V.A., Kommer, A., and Spirin, A.S., *Cell-Free Translation Systems*, Springer, New York, 2002, p. 131.
24. Kang, S.H. et al., Cell-free production of aggregation-prone proteins in soluble and active forms, *Biotechnol. Prog.*, 21, 1412, 2005.
25. Elbaz, Y. et al., *In vitro* synthesis of fully functional EmrE, a multidrug transporter, and study of its oligomeric state, *Proc. Natl. Acad. Sci. USA*, 101, 1519, 2004.
26. Klammt, C. et al., High level cell-free expression and specific labeling of integral membrane proteins, *Eur. J. Biochem.*, 271, 568, 2004.
27. Berrier, C. et al., Cell-free synthesis of a functional ion channel in the absence of a membrane and in the presence of detergent, *Biochemistry*, 43, 12585, 2004.
28. Kuruma, Y. et al., Development of a minimal cell-free translation system for the synthesis of presecretory and integral membrane proteins, *Biotechnol. Prog.*, 21, 1243, 2005.

29. Noireaux, V. and Libchaber, A., A vesicle bioreactor as a step toward an artificial cell assembly, *Proc. Natl. Acad. Sci. USA*, 101, 17669, 2004.
30. Kawasaki, T. et al., Efficient synthesis of a disulfide-containing protein through a batch cell-free system from wheat germ, *Eur. J. Biochem.*, 270, 4780, 2003.
31. Yin, G. and Swartz, J.R., Enhancing multiple disulfide bonded protein folding in a cell-free system, *Biotechnol. Bioeng.*, 86, 188, 2004.
32. Lowe, J.B. and Marth, J.D., A genetic approach to Mammalian glycan function, *Annu. Rev. Biochem.*, 72, 643, 2003.
33. Walter, P. and Blobel, G., Preparation of microsomal membranes for cotranslational protein translocation, *Methods Enzymol.*, 96, 84, 1983.
34. Bulleid, N.J. et al., Cell-free synthesis of enzymically active tissue-type plasminogen activator. Protein folding determines the extent of N-linked glycosylation, *Biochem. J.*, 286 (Pt 1), 275, 1992.
35. Zhang, Z. et al., A new strategy for the synthesis of glycoproteins, *Science*, 303, 371, 2004.
36. Jagus, R. and Beckler, G.S., *Current Protocols in Cell Biology*, John Wiley, New York, 1998, 11.1.1.
37. Davis, B.G., Biochemistry. Mimicking posttranslational modifications of proteins, *Science*, 303, 480, 2004.
38. Carlson, E. et al., Reticulocyte lysate as a model system to study endoplasmic reticulum membrane protein degradation, *Methods Mol. Biol.*, 301, 185, 2005.
39. Waugh, D.S., Making the most of affinity tags, *Trends Biotechnol.*, 23, 316, 2005.
40. Feldman, G. et al., Detection of tetracysteine-tagged proteins using a biarsenical fluorescein derivative through dry microplate array gel electrophoresis, *Electrophoresis*, 25, 2447, 2004.
41. Dyson, M.R. et al., Production of soluble mammalian proteins in *Escherichia coli*: Identification of protein features that correlate with successful expression, *BMC Biotechnol.*, 4, 32, 2004.
42. Tsunoda, Y. et al., Improving expression and solubility of rice proteins produced as fusion proteins in *Escherichia coli*, *Protein Expr. Purif.*, 42, 268, 2005.
43. Noren, C.J. et al., A general method for site-specific incorporation of unnatural amino acids into proteins, *Science*, 244, 182, 1989.
44. Tan, Z. et al., De novo genetic codes and pure translation display, *Methods*, 36, 279, 2005.
45. Josephson, K., Hartman, M.C., and Szostak, J.W., Ribosomal synthesis of unnatural peptides, *J. Am. Chem. Soc.*, 127, 11727, 2005.
46. Agafonov, D.E. et al., Efficient suppression of the amber codon in *E. coli in vitro* translation system, *FEBS Lett.*, 579, 2156, 2005.
47. Olejnik, J. et al., N-terminal labeling of proteins using initiator tRNA, *Methods*, 36, 252, 2005.
48. Tabuchi, I., Next-generation protein-handling method: Puromycin analogue technology, *Biochem. Biophys. Res. Commun.*, 305, 1, 2003.
49. Adams, S.R. et al., New biarsenical ligands and tetracysteine motifs for protein labeling *in vitro* and *in vivo*: Synthesis and biological applications, *J. Am. Chem. Soc.*, 124, 6063, 2002.
50. He, M. and Taussig, M.J., DiscernArray technology: A cell-free method for the generation of protein arrays from PCR DNA, *J. Immunol. Methods*, 274, 265, 2003.

51. Angenendt, P. et al., Cell-free protein expression and functional assay in nanowell chip format, *Anal. Chem.*, 76, 1844, 2004.
52. Ditursi, M.K. et al., Simultaneous *in vitro* protein synthesis using solid-phase DNA template, *Biotechnol. Prog.*, 20, 1705, 2004.
53. He, M. and Taussig, M.J., Functional Protein Arrays, International Patent WO 02/14860 A1, 2002.
54. Ramachandran, N. et al., Self-assembling protein microarrays, *Science*, 305, 86, 2004.
55. Kim, D.M. and Swartz, J.R., *Cell-Free Translation Systems*, Springer, New York, 2002, p. 41.

Section 2

Fabrication of Functional Protein Microarrays

4 The Critical Role of Surface Chemistry in Protein Microarrays

Athena Guo and X.-Y. Zhu

CONTENTS

INTRODUCTION

Protein microarrays are an important tool in proteomics. However, duplicating the success of the DNA chip for protein microarrays has been difficult. This account discusses a key issue in protein microarray development: surface chemistry. Ideally, the surface chemistry for protein microarray fabrication should satisfy the following criteria: the surface resists nonspecific adsorption; functional groups for the facile immobilization of protein molecules of interest are readily available; bonding between a protein molecule and a solid surface is balanced to provide sufficient stability but minimal disturbance on the delicate three-dimensional structure of the protein; linking chemistry allows the control of protein orientation; the local chemical environment favors the immobilized protein molecules to retain their native conformation; and finally, the specificity of linking chemistry is so high that no prepurification of proteins is required. We discuss strategies to achieve such an ideal situation and demonstrate the optimal activity of the immobilized protein molecules via surface molecular engineering. We elucidate how the commonly seen ring structures

in spot morphology on protein microarrays is related to partitioning of protein molecules between the bulk solution and the air-liquid interface due to the large surface-to-volume ratio of a nanoliter droplet. We also show how to eliminate this problem for quantitative applications.

THE PROMISE OF PROTEIN MICROARRAYS

Protein microarrays have been subject of considerable excitement in the last a few years, as evidenced by an exceptionally large number of review articles and commentaries published within a short period of time.[1,2] However, successful applications of the protein microarray technology are few and far between. What is the reason for such a unique situation? Answers to this question lie in the exceptional potential of the protein microarray technology, as well as the exceptional difficulty in developing such a technology.

With the great success in genomics, there is a pressing need for large-scale profiling and functional analysis of the protein molecules encoded by genes. One of the most exciting tools in this endeavor is protein microarray technology in which a large number of proteins or peptides are immobilized on a solid substrate for the high-throughput, parallel analysis of population profiles, biochemical properties and biological activities. A wide range of applications have been envisioned and/or demonstrated for protein microarrays, including expression profiling, interaction profiling, and functional identification. These applications are detailed elsewhere in this book. The first application is most obvious. The concentration profile of proteins in an organism depends on age, physical/chemical environment, and more importantly, disease state. The need to go beyond mRNA profiling arises because of the general presence of translational and post translational modifications as well as protein degradation by proteolysis. Thus, knowing protein levels is the most direct way to phenotype cells and to diagnose disease state, stage, and response to treatment. This task is possible and has already been explored with antibody arrays. The second application is critical to drug discovery. Given the large number of proteins and the fact that their activities are often intimately related to mutual interactions, it is a daunting task to identify and understand the vast possibilities of protein–protein interactions. One may envision the preparation of protein microarrays with the whole or a subset of human proteome and their use for large-scale categorization of protein–protein interactions, including the identification of specific domain–domain interactions. These microarrays can also be used in drug discovery since many drugs function by disrupting protein–protein interactions. The last application is most difficult but is important for fundamental understanding. The functions of only a small population of proteins are known at the present time and the main goal of proteomics is to associate each protein with particular functions. A protein microarray may be used to screen for corresponding targets. The reciprocal process is to use a small molecule array to screen for binding with proteins. In many ways, functional profiling overlaps with interaction profiling.

Despite all the potential and expectations, it is naïve to assume that the success story of DNA microarrays can be duplicated for protein microarrays. The availability of oligonucleotide synthesis and PCR has made the production of DNA molecules a routine task. However, techniques equivalent to PCR do not exist for proteins.

Proteins are produced in small quantities either recombinantly in cells or in cell-free translation systems, neither of which is as simple as PCR. Producing and purifying antibodies from biological samples of animal models based on the natural immune systems is also a difficult and labor intensive process. In terms of handling, proteins are much more difficult than DNA molecules. DNA molecules are relatively simple polyanions which can be chemically modified and easily immobilized on solid surfaces based on electrostatic interactions or covalent bonding through functional groups on either terminus. Protein molecules are much more complex. They possess delicate three-dimensional (3-D) structures, varying chemical and physical properties (e.g., hydrophobic, hydrophilic, and ionic domains). Because the activity or function of a protein molecule is critically dependent on its 3-D structure which is very sensitive to local physical and chemical environment, keeping an immobilized protein molecule in a native state with its 3-D structure intact and with its active domains accessible, is a major challenge.

THE DEMAND FOR SURFACE CHEMISTRY

The challenge in protein microarray development is manifested in the stringent demand on surface chemistry, which is the focus of this chapter. There are two inherent difficulties associated with protein surface chemistry. The first problem is *background*. Proteins tend to adsorb nonspecifically to most solid surfaces. This is because a protein molecule has various hydrophobic domains, charged sites, and hydrogen bond donor/acceptor groups. These groups can bind strongly with hydrophobic surfaces, oppositely charged sites, and hydrogen bond acceptor/donor groups. The hydrophobic interaction (van der Waals) is particularly prevalent and is the dominant reason for the fouling of surfaces. The excessive interaction between a protein molecule and a solid surface often results in the disruption of its 3-D structure and eventually denaturation, i.e., the complete loss of activity. The second problem is *conformation/orientation.* A protein molecule interacts with other molecules through specific functional domains. However, chemical forces responsible for adsorption on a solid surface are oblivious of the presence of any functional domain. If we let nature take its course, chances are we will not have protein molecules with the desired orientation on a solid surface. We must engineer specific chemical functionality to differentiate the domain responsible for immobilization from those of chemical/biological activity. Ideally, we would like the surface chemistry for protein microarrays to meet the following criteria:

- The surface is inherently inert and resists nonspecific adsorption;
- The surface contains functional groups for the facile immobilization of protein molecules of interest;
- Bonding to a solid surface is strong enough to retain the protein on the surface, but sufficiently non intrusive to minimize disturbance to the delicate 3-D structure;
- The linking chemistry allows the control of protein orientation and makes active sites easily accessible to target molecules in the solution phase;
- The immobilization chemistry is highly specific and does not require prepurification of protein samples.

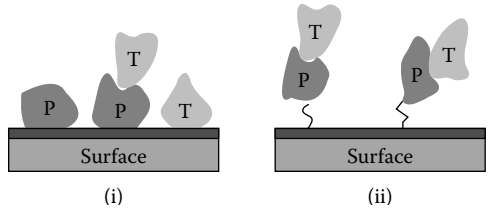


FIGURE 4.1 A comparison of the adsorption of a protein molecule (P) and its interaction with a target (T) on: i) a "sticky" surface; and ii) a repulsive surface.

The particular importance of an inert starting surface can be easily comprehended from the illustration in Figure 4.1. Consider a surface not repulsive towards proteins but containing specific functional groups, such as aldehyde or epoxy, for covalent bonding to $–NH_2$ groups on a protein molecule. Alternatively, the surface functional group may assist noncovalent protein adsorption, e.g., $–NH_2$ functionalized coatings for protein adsorption through electrostatic and hydrogen bonding interactions. The backbones of these surface coatings are inherently "sticky" and permit nonspecific adsorption. Such a surface may lead to excessive protein-surface interaction, resulting in the loss of activity. In addition, a substantial percentage of protein molecules can adsorb on the surface with their active sites inaccessible to target molecules. Finally, a target molecule can also adsorb nonspecifically on the "sticky" surface, thus contributing to background signal. Before the surface can be used for protein-target interaction, there is often a need for the so-called "blocking" step, which consists of adsorption of other protein molecules, e.g., bovine serum albumin (BSA). The blocking step is problematic: small probe molecules can be buried by large blocking molecules and the adsorbed blocking molecules are not completely "nonfouling" and may also interact nonspecifically with targets. Examples of this kind of surfaces include widely used and commercially available aldehyde, epoxy, and amine functionalized silane coatings.

The second type of surface starts with an inert coating. The surface is activated for covalent linking to a protein molecule. Besides the covalent linker, the repulsive nature of the surface ensures that the immobilized protein has little interaction with the surface. In other words, the covalently attached protein molecule prefers to stay away from the solid surface. This leads to optimal activity and accessibility of the immobilized protein to interact with the target. After the immobilization step, remaining active groups on the repulsive coating can be easily removed/titrated by chemical means, thus eliminating the need for blocking with other protein molecules. Examples of these inert surfaces include oligoethyleneglycol (OEG) terminated alkanethiol self-assembled monolayers on Au as introduced by Whitesides' group[3] and applied extensively by Mrksich and coworkers,[4] and other polyether based coatings by our groups,[5,6] and by Moeller and coworkers.[7] High-density polyether brush coatings are now commercially available from MicroSurfaces, Inc.[8]

Recently, Whitesides and coworkers[9] surveyed a large number of surface functional groups and concluded that the most extensively studied oligo or poly-ethyleneglycol (PEG)[10] remains the most "inert" chemical group toward protein adsorption,

FIGURE 4.2 Fluorescence microscopic images taken after the adsorption (spotting) of fibrinogen (1 mg/ml) on polyether brush coated glass (left) or clean glass (right, spot diameter ~ 100 μm). For detection, the surface is first incubated with primary antibody and then with Cy3-labeled secondary antibody.

often referred to as "inert" or "nonfouling." The inertness or nonfouling property of PEG coatings towards protein adsorption is attributed to its hydrophilic nature.[11] The PEG backbone is extensively hydrogen-bonded to water molecules, resulting in the formation of partially structured water extending into the aqueous phase. Adsorption of a protein molecule requires the disruption of this structured water layer and is enthalpically inhibited. In addition, protein adsorption leads to the compression of the PEG layer towards the solid surface and is entropically unfavorable. Note that a partially structured water layer may form on any highly hydrophilic solid surface, such as clean silica.[12] In principle, such a surface is also resistant to protein adsorption because of the enthalpic barrier. However, the nonfouling property of a clean silica surface is quickly lost due to the adsorption of impurity from the background. The presence of the PEG coating prohibits the adsorption of impurities and thus maintains its nonfouling property. As an example, Figure 4.2 compares the adsorption of fibrinogen on a high-density PEG brush coated glass slide (left, MicroSurfaces) and a clean glass slide (right). Here the amount of fibrinogen adsorption is detected via a sandwich assay (immunostaining). While the PEG brush is completely inert, the clean glass surface adsorbs not only fibrinogen (spot) but also antibodies (background).

SURFACE CHEMISTRY FOR THE BINDING OF PROTEINS WITH RANDOM ORIENTATION

Two general approaches that do not meet the above criteria but can be easily implemented for protein immobilization involve either the passive adsorption of protein molecules into a polymer matrix or covalent bonding via $-NH_2$ groups on the surface of protein molecules.

The first approach is mainly derived from conventional methods (such as western blotting) available in biochemical laboratories. It uses filter membranes (e.g., nitrocellulose, nylon, polyvinylidene difluoride) or glass slides coated with a polymer film (e.g., poly-L-lysine and polyacrylamine) for the immobilization of proteins.[13–16] The advantage of using a polymer matrix is the ease of which protein is immobilized and the relatively high load of protein samples in each spot. There are also problems,

including the inability to control protein orientation and local environment, the unknown diffusion kinetics or the inaccessibility of large target molecules into the polymer matrix, and the possibility of wash-off or exchange reactions with solution phase proteins during analysis. Excess washing necessary for these polymer matrixes can potentially denature the adsorbed protein molecules.

The second approach generally relies on covalent bond formation between amine groups on protein molecules and other functional groups on a solid support. For example, MacBeath and Schreiber demonstrated high density protein arrays on glass slides through Schiff's base linkage formed from amine groups on protein molecules and aldehyde groups on silanized glass surface.[17] Zhu et al. fabricated protein arrays in microwells on a silicone elastomer sheet based on covalent bond formation between amine groups and epoxide groups on the silanized surface.[18] Because a protein molecule usually displays many lysines on its surface in addition to the terminal amine group, it can be covalently bonded to a substrate via a variety of orientations. One should recognize that, despite common beliefs, there is little actual experimental evidence for the presence of covalent bonding between protein molecules and the solid support in the above examples. Most of these surfaces are inherently susceptible to nonspecific adsorption. For example, aldehyde or epoxy surfaces obtained from silanization reactions are partially hydrophobic. It is not known what percentage of the immobilized protein molecules are actually results of nonspecific, hydrophobic-hydrophobic interaction.

A major concern with both approaches is that the protein molecules are randomly oriented on the surfaces. As a result, the active sites of a substantial population of immobilized protein molecules are not accessible to targets in the solution phase. The nonspecific nature of these approaches inevitably requires the use of purified protein samples. The wide variation in orientation may give rise to a distribution of binding constants and kinetic constants, thus limiting the fidelity, sensitivity, and resolution of the array. In addition, there is possibility of denaturing when the interaction between randomly immobilized protein and the surface is too strong. An excellent review by Kusnezow and Hoheisel deals with the surface problems in protein microarray technology.[19] They point out many of the practical limitations of current approaches, including the problems associated with random protein orientations. Similar conclusions were reached by Seong and Choi.[20] Recently, Cahill and coworkers carried out a comparative study of various surface coatings for protein and antibody microarrays, all involving random orientations.[21] A major conclusion was that a PEG coating with epoxy termination was found to give best results for antibodies. This finding is consistent with the arguments presented in section 2 and Figure 4.1: the use of a nonfouling starting surface not only minimizes background adsorption, but also optimizes the local chemical environment for the immobilized protein molecule to maintain its activity and accessibility. Kusnezow et al.[22] and Guilleaume et al.[23] carried out systematic comparisons of various surface coatings

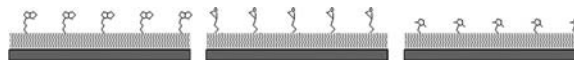


FIGURE 4.3 Functionalized polyether brush surfaces for protein immobilization: biotin (left), epoxy (middle), and NHS (right). The light-blue region represents a hydrated polyether film.

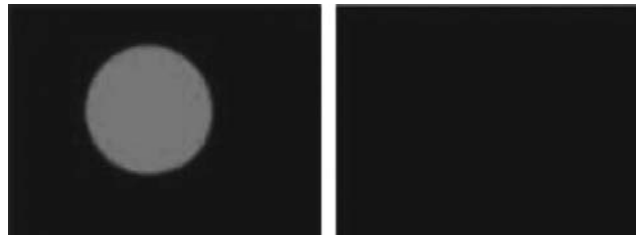


FIGURE 4.4 The specific adsorption of Cy3-labeled streptavidin to the biotin–polyether brush surface (left). The surface remains resistant to the nonspecific adsorption of fibrinogen (right).

(including most commercially available ones) for protein and antibody immobilization. Most of these coatings are generated from silane based chemistry and are not inherently inert toward protein adsorption.

We have systematically developed surfaces for protein immobilization based on the inert starting surface: a high-density polyether brush whose nonfouling property is demonstrated in Figure 4.2. Figure 4.3 shows some of the examples where a certain percentage of alcohol functional groups on the surface of the polyether brush are converted to biotin, epoxy, or NHS groups. Here the biotin surface is used for immobilization based on the specific biotin-streptavidin chemistry, while the epoxy or –NHS terminated surfaces are used for covalent attachment to $–NH_2$ groups on the surface of protein molecules. In the case of epoxy or –NHS terminated surfaces, the remaining active sites on the surface after protein immobilization can be easily titrated by small molecules containing $–NH_2$ groups. A common feature of all these surfaces is the exceptionally low background of the PEG backbone. As an example, Figure 4.4 shows the biotin/polyether brush surface specifically adsorbs streptavidin but remains completely inert to other protein molecules, such as the "sticky" fibrinogen.

CONTROLLED PROTEIN ORIENTATION AND ACTIVITY ON SURFACES

The advantage of controlled protein orientation over random orientation is easily understood from the cartoon in Figure 4.5. In the oriented approach, the site for the adsorption of a protein molecule can be engineered specifically to a domain remote

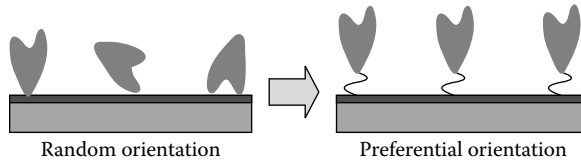


FIGURE 4.5 Schematic illustration of protein immobilization with and without orientation control.

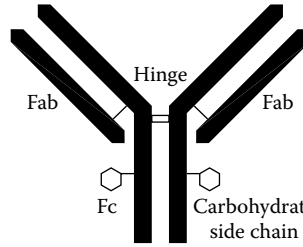


FIGURE 4.6 Schematic illustration of an antibody molecule.

from the active site. This approach not only makes the functional domain easily accessible but also minimizes protein-surface interaction through domains other than the activated region. Seong and Choi[20] recently summarized various approaches for the immobilization of protein molecules with controlled orientation, while Kusnezow and Hoheisel[19] focused their discussions on controlling the adsorption of antibodies.

Antibodies

Three approaches have been demonstrated for the immobilization of antibodies with controlled orientation. Figure 4.6 illustrates schematically a typical antibody molecule which consists of two heavy chains and two light chains linked together by disulfide bonds. Orientation control can be achieved via selective interaction with (a) specific regions on the antibody molecule; (b) the carbohydrate side chains in the Fc region; and (c) the disulfide bond in the hinge region or other free –SH groups.

The first relies on surface immobilized protein molecules that recognize specific domains on antibodies. Oriented antibodies may be obtained using immobilized Protein A or G, which binds to the Fc portion of antibodies.[24,25] Similarly, protein L is known to bind to a specific repeated homologous domain on the light chain.[26] There are two disadvantages associated with these approaches: (a) they are applicable only to a subset of immunoglobins with high affinity for proteins A, G, or L; and (b) the surface density of antibodies is usually low due to the low densities of surface immobilized protein molecules (A, G, or L) with the correct orientation.

The second approach uses recognition or special chemical modification of carbohydrate residuals in the Fc regions on antibodies. Peluso et al. used biotinylated antibodies on streptavidin coated surfaces to achieve orientation control via chemical modification of the glycosylation sites in the Fc region of IgG to attach biotin units.[27] Galactose residuals can be partially oxidized to give aldehyde functionality, which can covalently attach to surface hydrazide groups, as demonstrated by Turkova et al.[28]

The third approach relies on surface activity toward thiol (-SH) groups on antibody molecules. While free thiol groups are present at selected locations on antibody molecules,[29] they can also be generated by chemically reducing the disulfide bridges, e.g., one of the inter-heavy chain disulfide bonds in the hinge region (see Figure 4.6). The –SH group can covalently bond to a surface –SH group via disulfide bond or to a maleimido group in a cross linker molecule.[22,30,31] One concern with

chemical reduction of disulfide bonds is the disturbance to the structure (and thus activity) of antibody molecules.

The advantages of oriented over randomly adsorbed antibody molecules have been generally observed. Peluso et al. found that oriented IgGs immobilized via biotin modification of the glycol region with a long spacer showed higher activity in antibody-antigen binding than those with short spacers or those with random orientation.[27] The difference in spacer length is likely a result of nonspecific interaction between the immobilized antibody and the inherently "sticky" surface coating. Anderson et al. showed that oriented antibodies on a protein A–adsorbed surface possessed better sensitivity over randomly oriented antibodies in immunoassays.[32] Vijayendran and Leckband compared the activities of an anti-TNT antibody immobilized via random orientation on an amine active surface to those immobilized with controlled orientations via carbohydrate side chains, via recognition of the Fc region by surface adsorbed protein G, or via biotin modification of the Fc region and adsorption on streptavidin surfaces.[33] While the surface densities of oriented antibodies are lower than that of random orientation, much higher sensitivity is seen for oriented antibodies via the carbohydrate or protein G strategy. Interestingly, the biotin approach did not show enhanced antibody activity, probably due to disturbance to the antibody molecular structure by the chemical modification step.

There are also reports of mixed results on the performance of oriented vs. randomly adsorbed antibodies, depending on the specific surfaces used.[19,34,35,36] A fundamental deficiency common to most of the surfaces used is the lack of nonfouling properties of the coatings used. Besides the designed specific interaction for orientation control, the intrinsic stickiness of the surfaces may introduce undesirable and nonspecific interactions between the immobilized antibody molecules and the surface, leading to mixed results in different studies. This again calls for the use of a nonfouling starting surface.

Fusion Proteins and Peptides

Various fusion proteins, including protein–protein, protein–mRNA, and protein–cDNA, can be used to immobilize the protein of interest with controlled orientation. Mrksich and coworkers demonstrated a fusion protein approach in which serine–estarase cutinase is used as anchor to bind and covalently attach to a surface phosphate ligand.[37] They successfully immobilized cutinase–calmodulin fusion proteins on a self-assembled monolayer covered surface and showed the activity of oriented calmodulins in the binding of target calcineurin. Other fusion proteins,[38] such as those with glutathione S-transferase (GST), are also commonly obtained from recombinant technology and can be used for the immobilization of oriented proteins of interest. Weng et al. produced oriented protein microarrays based on mRNA-protein or mRNA-peptide hybrids.[39] The oriented protein on the surface exhibited exceptional high sensitivity. A similar strategy by Kurz et al. used cDNA-protein hybrids.[40]

Other approaches use chemical tagging of peptides. Raines and coworkers achieved site-specific protein immobilization based on the strong binding of a chemically modified S-peptide (with terminal azide group) to ribonuclease S', and the

Staudinger ligation reaction in which the azide and a surface phosphinothioester group react to form an amide bond.[41] They demonstrated that protein molecules immobilized on the surface with uniform and controlled orientation possessed higher enzymatic activity than those of random orientation. Mrksich and coworkers used peptide–cyclopentadiene conjugate to covalently attach peptides with controlled orientation to benzoquinone groups via the Diels-Alder reaction.[4] These authors used an OEG terminated monolayer as a nonfouling starting surface.

Poly-histidine Tagged Proteins

Perhaps the most general approach for the immobilization of oriented protein molecules is the use of recombinant tags, particularly poly histidine (His-tag). This strategy originates from immobilized metal ion affinity chromatography (IMAC)[42] and has been applied to protein immobilization.[43–46] There are a number of advantages of developing IMAC into a general strategy for the fabrication of protein microarrays. The generation of His-tag to either the C-terminus or N-terminus is the most commonly used method in recombinant protein technology. Unlike other fusion protein strategies, the His-tag approach for purification can be applied not only to proteins in native states, but also to those under denaturing conditions or to small peptides. When applied to protein microarray technology, this strategy effectively combines the steps of purification and immobilization, provided that the surface coating is inert otherwise. Thus, the labor intensive purification process required for most other strategies may be eliminated in the His-tag approach. In addition, unlike chemical modification in other methods, a His-tag generally does not interfere with the structure or function of proteins and does not affect the secretion, compartmentalization, or folding of fusion proteins within cells.[47,48] The anchoring bond is highly stable and reversibility occurs only in the presence of high concentration of competing ligands, such as imidazole. Most studies to date used nitrilotriacetic acid (NTA) or iminodiacetic acid (IDA) as a chelating group to bind bivalent metal ions on the surface. Recently, Johnson and Martin suggested an alternative, a macrocycle triazacyclononane, which showed improved long-term stability as compared to NTA.[49]

To satisfy all five requirements set forth in section 2, we have developed surface chemistry for protein immobilization via the His-tag on an otherwise "zero" background PEG coating. We demonstrated this approach using the high-density polyether film, whose excellent nonfouling property is illustrated in Figure 4.2. Our approach is illustrated in Figure 4.7. A high-density PEG coating is first formed on a silicon or

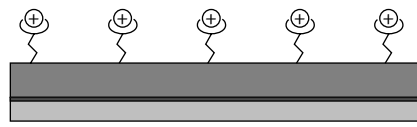


FIGURE 4.7 The polyether/Cu^{2+} surface for the immobilization of poly-His tagged protein molecules.

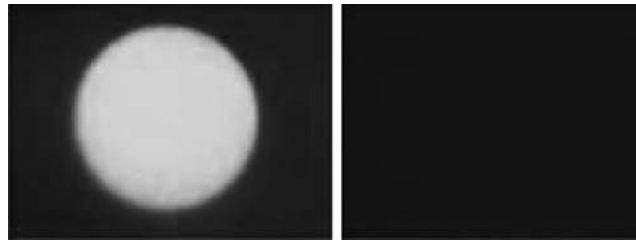


FIGURE 4.8 The left image shows the specific adsorption of 6xHis tagged green fluorescent protein (GFP) on the chelated Cu^{2+}/polyether surface. For comparison, the surface resists the nonspecific adsorption of other protein molecules, e.g., GFP without His tag (right). The spot diameter is ~0.2 mm.

glass surface. The exposed alcohol groups on the surface of the PEG coating is used to link chelating groups and the binding of Cu^{2+} ions.[5] Because there are inherent problems with IMAC (NTA/His-tag), such as leaching and protein dissociation[50,51] we used metal Cu instead of Ni to provide more robust binding to the poly His-tag. The resulting Cu^{2+}- PEG surface is shown to specifically bind 6x-histidine-tagged protein molecules, but otherwise retains its inertness towards nonspecific protein adsorption as demonstrated for green fluorescent protein (GFP) in Figure 4.8.

Except for the His-tag on the N- or C-terminus, each immobilized protein molecule stays away from and minimizes its interaction with the surface due to the repulsive nature of the PEG environment. As a result, there is minimal disturbance to the native conformation of the protein. Both the inertness of the chemical surrounding and the controlled orientation should contribute to an ideal environment for the immobilized protein molecule to retain its native conformation and activity.

We have compared the enzymatic activities of the 6x-histidine tagged Sta IV in the solution phase with those immobilized with controlled orientation on the Cu^{2+}-PEG surface or with random orientations on surfaces.[6] The sulfotransferases refer to an entire family of enzymes of detoxication that catalyzes the transfer of the sulfuryl group, SO_3^-, from adenosine 3'-phosphate 5'-phosphosulfate (PAPS) to a wide range of xenobiotics, such as phenols, alcohols and amines, etc. This model system is chosen because the mechanism and substrate specificity for this family of enzymes have been well characterized.[5] We characterize enzyme kinetics using the method of Beckmann who showed that Sta IV catalyzes the sulfation of a fluorescent compound, resorufin, to its nonfluorescent derivative.[52] Thus, we can simply follow the catalytic reaction in the time domain by recording fluorescence decay of the reactant.

Figure 4.9 shows fluorescence decay data for the sulfo transfer reaction catalyzed by 6xHis-Sta IV immobilized on different surfaces.[6] The first surface (A) is an epoxy-functionalized silane monolayer from 3-glycidyoxypropyl trimethoxysilane on a native oxide terminated silicon surface. The second surface (B) is the multiarmed poly(ethylene glycol) monolayer covered Si activated by disuccinimidyl carbonate (DSC) (Figure 4.9B). Both surfaces are reactive towards $-NH_2$ functional groups on protein molecules for covalent attachment. Since there are many lysine residuals on the protein molecules, each sulfotransferase can be immobilized with a variety of

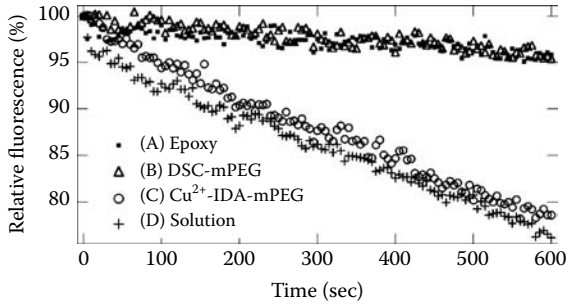


FIGURE 4.9 Fluorescence decay from resorufin during the reaction of sulfuryl from PAPS, catalyzed by 6xHis-Sta IV immobilized on the surface with random (A & B) and controlled orientations (C), or in the solution (D).[6]

orientations on these two surfaces. The third surface (C) is the Cu^{2+}-polyether surface for the specific binding to the poly-His tag. The slope of each decay curve measures the enzyme activity. The enzymatic activities of randomly oriented Sta IV molecules (A & B) are 5 to 6 times lower than that of the oriented sample (C). Within experimental uncertainty, the enzymatic activity of oriented 6xHis-Sta IV on the Cu^{2+}-polyether surface (C) is the same as that of enzyme molecules in the solution phase (D). We have also carried out similar comparisons for alkaline phosphatase. Both experiments are summarized in Figure 4.10.

These results establish the critical importance of controlling the orientation of immobilized molecules in protein microarray technology. While oriented protein molecules faithfully reflect activities of solution phase proteins, those with random orientations do not. In the case of randomly oriented enzyme molecules, the active sites on certain population on the surface are not accessible. The possible presence of multiple attachment sites on each protein molecule may also affect its conformation. We conclude that controlling the orientation of immobilized protein molecules and designing an ideal local chemical environment on the surface are both essential for quantitative applications of the protein microarray technology.

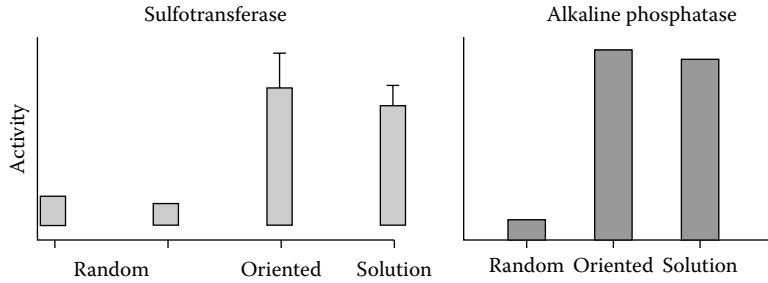


FIGURE 4.10 A comparison of enzyme activities in the solution phase or with random or controlled orientations on solid surfaces for sulfotransferase (left) and alkaline phosphatase (right).

SURFACE PROCESSES AND SPOT MORPHOLOGY: RINGS

In addition to controlling the local chemical environment for the immobilized protein molecules, we must also address new chemical/physical processes introduced by the use of small sample volumes in protein immobilization. Protein microarrays are usually made with robotic spotters, which deposit nanoliter to sub-nanoliter size droplets of protein solutions on a solid surface. The use for small sample volume is in fact one of the main attractions of protein microarray technology. After incubation and washing off excess solution, the microarray is used for probe-target interaction and the result is most commonly detected via fluorescence imaging. A survey of protein microarray literature shows that one of the major reasons for poor reproducibility is nonuniform spot profile. In particular, spots on a protein, peptide, or antibody microarray often exhibit ring-like structures (including donut and solar-eclipse shapes). Despite their common occurrence, the mechanism for ring formation in protein microarrays is not understood. Formation of ring structures is well documented for thin films deposited on solid surfaces by the evaporation of a solution or suspension of a wide variety of materials,[53–55] the most commonly seen rings being coffee stains.[56] However, the mechanism for generating these ring structures all involve drying and cannot be responsible for the ring structure seen in protein microarrays where the spots are kept hydrated.

A typical example of the ring structure is shown in Figure 4.11a for an antibody spot on an epoxy terminated glass slide.[62] After the deposition of nanoliter droplets of antibody solutions on the epoxy slide, we kept the sample in an environment with controlled humidity and confirmed using optical microscope that the size of each droplet on the surface did not change during incubation. A close examination of the

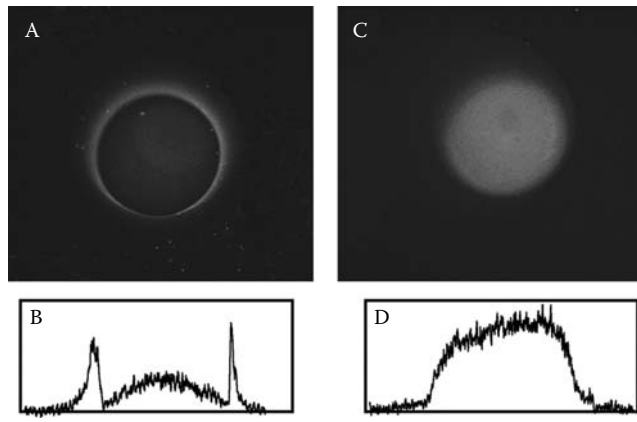


FIGURE 4.11 (a) & (c): Fluorescence microscope images of antibody spots immobilized on an epoxy functionalized glass slide. In (A), a diluted antibody solution (1:500) was used directly while in (C) a small amount detergent (0.006% triton X-100) was added to the diluted antibody solution. Panels (B) & (D) are cross-sectional profiles of images (A) & (C), respectively. The spot diameter is ~0.2 mm.[62]

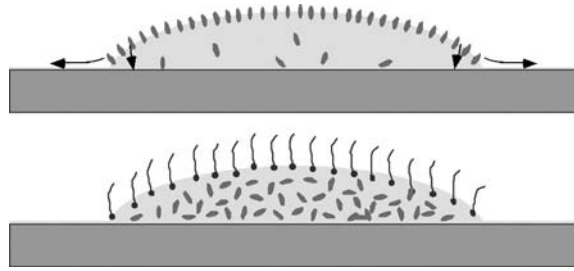


FIGURE 4.12 Schematic illustrations of nanoliter droplets (light blue) on a solid surface (gray) with protein molecules in red and detergent molecules in dark blue. The arrows indicate regions of enhanced reaction rate for protein adsorption.

morphology of the ring structure in Figure 4.11a, particularly the cross-sectional profile of fluorescence intensity in Figure 4.11b, provides clue. Within the ring, the fluorescence intensity peaks at the center and gradually decreases towards the boundary of the spot. Immediately outside the boundary, the concentration of immobilized antibody rises rapidly then decays with increasing distance from the boundary. To form such a concentration profile, protein molecules must be transported to the boundary of the droplet. Because the droplet remains stationary (no expansion or contraction) during incubation, we believe that transport of protein molecules occurs at the air-liquid interface.

The proposed mechanism is shown schematically in Figure 4.12 (upper panel). Protein molecules are known to preferentially accumulate at air/water interfaces.[57,58] Because the surface area to volume ratio scales with the inverse of droplet size, the equilibrium between solution phase protein and adsorbed protein at the air/liquid interface should greatly shift to the latter as the size of the droplet decreases from macroscopic to the nanoliter and sub nanoliter scale. This effect provides an efficient mechanism for transporting protein molecules to the perimeter of the droplet, thus giving rise to a high concentration of protein molecules at the boundary of the spot. Depending on the surface hydrophilicity and the contact angle, the accumulation of protein molecules at the boundary may result in either "donut" or "solar-eclipse' shapes. On one hand, if diffusion outside the spot boundary is not important, the enhanced probability of interacting with the surface near the perimeter via transport through the solution results in a donut shape. On the other hand, diffusion of protein molecules accumulated at the boundary to area outside the spot accounts for a solar-eclipse profile.

We can eliminate the ring structure by adding competitive surfactants to displace protein molecules at the air/water interface. Figure 4.11C shows fluorescence microscope image of the antibody spot obtained with a small amount of detergent (0.006% triton X-100) added to the antibody solution, under otherwise identical conditions as in Figure 4.11A. Instead of the ring, we now observe nearly uniform intensity inside the spot, with negligible intensity outside the boundary (see also cross-sectional profile in Figure 4.11D). The integrated intensity of the spot in Figure 4.11C is two times that of the total intensity in Figure 4.11A. In the absence of competitive surfactants, the accumulation of protein molecules at the air/water interface and the

perimeter of the spot results in a depletion of protein concentration within the nanoliter droplet and, thus, a decreased immobilization efficiency. When protein molecules are displaced from the air/water interface by competitive surfactants, the concentration of protein solution in contact with the solid surface is the same as concentration in bulk sample. As a result, the immobilization efficiency is now directly related to protein concentration. This is also critical to the quantitative use of protein microarrays. The role of detergent is illustrated schematically in the lower panel of Figure 4.12.

To further verify the mechanism of ring formation, we use a model system: the immobilization of 6x histidine tagged green fluorescent protein (6xHis-GFP) on polyether coated glass slides with controlled density of chelated surface Cu^{2+} ions. These surfaces are commercially available (MicroSurfaces, Inc., Minneapolis, USA) and are similar to that described in a previous publication.[5] The advantage of this system can be realized from the fact that intrinsic fluorescence is detected only when GFP is active under fully hydrated conditions and any ring formation mechanism due to drying can be completely eliminated. The reaction between 6xHis tags and surface Cu^{2+} sites is facile and highly selective. There is no protein adsorption in the absence of surface Cu^{2+} or poly-His tags. Except for activated surface sites with chelated Cu^{2+} ions, other area on the polyether coating is repulsive towards protein adsorption. With increasing concentration of surface active sites, the surface becomes less repulsive, resulting in a shift in equilibrium toward adsorbed protein on the solid surface. Figure 4.13 shows fluorescence microscope images of 6xHis-GFP immobilized on the surface with different concentrations of surface Cu^{2+} as determined by X-day photoelectron spectroscopy: (A) 2.6×10^{13}/cm;[2] (B) 4.9×10^{13}/cm2; (C) 1.2×10^{14}/cm2; and (D) 2.0×10^{14}/cm.[2] As expected, the efficiency of protein immobilization (fluorescence intensity) within the spot increases as the density of surface reactive sites increases (bottom panel). For [Cu^{2+}] less than ~$5 \times 10^{13}/cm^2$ (Panels A & B), fluorescence intensity inside the spot is less than that at or immediately outside the boundary and the ring structure is observed. At higher [Cu^{2+}] (panels C & D), the spot morphology becomes uniform. The finding of such a transformation in spot morphology illustrates the central role of kinetics in protein immobilization. For protein molecules within the nanoliter droplet, immobilization onto the surface and transport via the air/water interface to the spot boundary are two competing kinetic processes. In the 6xHis-GFP example shown here, transport dominates for low [Cu^{2+}] while surface immobilization reaction kinetics wins at higher surface active site densities. Thus, designing surface chemistry for a facile immobilization reaction is critical in ensuring uniform spot profiles.

The improvement of spot morphology by the addition of detergent has been observed earlier by Kusnezow et al.[19] and can be explained by the transport model presented above. Delehanty and Ligler observed similar improvement in spot morphology in protein microarrays by the addition of BSA into the printing buffer.[59] The more hydrophobic BSA molecule is expected to preferentially accumulate on the surface of the nanoliter droplet, thus resulting in a similar surfactant effect.

Note that the above addresses the unique role of interfaces (air–liquid and liquid–solid) in the immobilization step due to the small volume of liquid used. We should

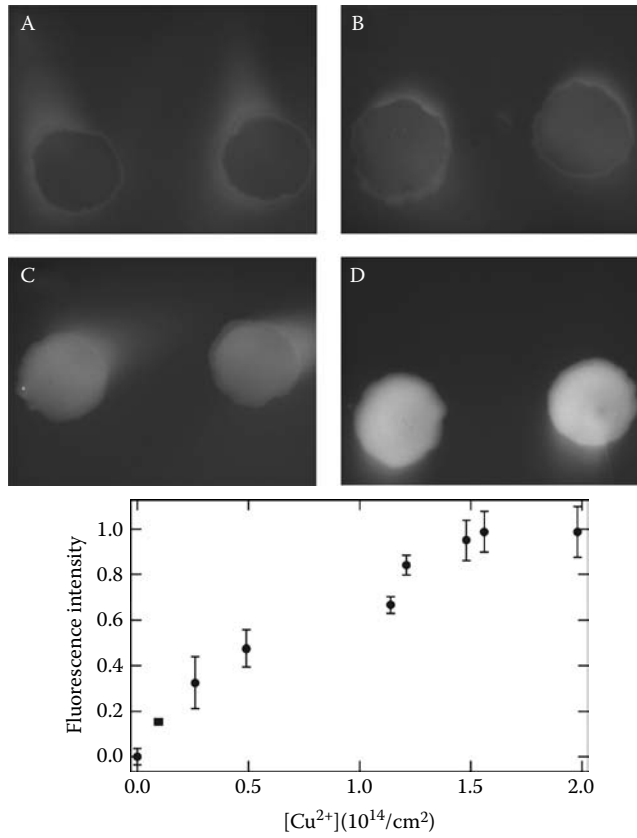


FIGURE 4.13 Fluorescence microscope images of 6xHis tagged GFP immobilized on Cu^{2+} chelated polyether/glass surfaces with different surface Cu^{2+} concentrations: (A) 2.6×10^{13}/cm;^{2} (B) 4.9×10^{13}/cm;^{2} (C) 1.2×10^{14}/cm;^{2} and (D) 2.0×10^{14}/cm.^{2} Nanoliter droplets of crude lysate solution (4 mg/ml) containing 6xHis-GFP and 10% glycerol were deposited onto the glass slide by the robotic spotter. Each slide was incubated at room temperature for 10 minutes, rinsed quickly with PBS containing 0.01% Tween-20 three times. The slide was covered with the buffer solution and imaged under the fluorescence microscope (excitation wavelength ~488 nm). The lower panel shows the fluorescence intensity within the spot as a function of surface $[Cu^{2+}]$ concentration.[62]

expect similar size-dependent effects in the assaying step, i.e., the incubation of a small droplet of sample solution with immobilized spots. This issue has been carefully analyzed by Kusnezow and co-workers recently.[60,61]

SUMMARY

We have demonstrated the importance of surface chemistry in protein microarray development. Compared to DNA microarrays, the demand on surface chemistry in protein microarray fabrication is much more stringent. Here, the surface chemistry

is not just about anchoring a protein molecule to the surface, but more about providing low background, controlling protein orientation and minimizing disturbance to the 3-D structure. In addition, the use of small volumes of protein solution in the fabrication of protein microarrays introduces new complications due to the large surface/volume ratio of nanoliter droplets and the partitioning of protein molecules among the solution phase, the liquid–air interface, and the liquid–solid interface. Meeting these challenges requires a fundamental understanding of protein-surface interaction chemistry and the changes in kinetics and equilibrium due to space confinement in nanoliter droplets. The great variation in the chemical and physical properties of protein molecules also necessitates custom-designing unique surface chemistry for difference classes of protein and antibody molecules.

ACKNOWLEDGMENTS

Financial supports from the National Institute of Health and the National Science Foundation in the form of SBIR grants to MicroSurfaces are acknowledged.

REFERENCES

1. A selected list of reviews and commentaries published before 2004:
   (a) Abbott, A., *Nature*, 402, 715, 1999.
   (b) Emili, A.Q. and Cagney, G. Cagney, *Nature Biotech.*, 18, 393, 2002.
   (c) Pandey, A. and Mann, M., *Nature*, 405, 837, 2000.
   (d) Walter, G. et al., *Curr Opin Microbiol*, 3, 298, 2000.
   (e) Blohm, D. H. and Guiseppi-Elie, A., *Curr. Opin. Biotech.*, 12, 41, 2001.
   (f) Cahill, D. J., *J. Immun. Methods*, 250, 81, 2001.
   (g) Kodadek, T., *Chem. & Biol.*, 8, 105, 2001.
   (h) Fung, E. T. et al., *Curr. Opin. Biotech.*, 12, 65, 2001.
   (i) Haab, B. B., *Curr. Opin. Drug Discov. Devel*., 4, 116, 2001.
   (j) Jenkins, R. E. and Pennington, S. R., *Proteomics*, 1, 13, 2001.
   (k) MacBeath, G., *Nature Biotechnol.*, 19, 828, 2001.
   (l) Reineke, U., Volkmer-Engert, R., and Schneider-Mergener, J., *Curr. Opin. Biotechnol.*, 12, 59, 2001.
   (m) Taussig, M.J., *Comp. Funct. Genom.*, 2, 298, 2001.
   (n) Tomlinson, I. M. and Holt, L. J., *Genom. Biol.,* 2, 1004, 2001.
   (o) Wilson, D.S. and Nock, S., *Curr. Opin. Chem. Biol.*, 6, 81, 2001.
   (p) Zhou, H. et al., *Trends Biotechnol*., 19, S34, 2001.
   (q) Zhu, H. and Snyder, M., *Curr. Opin. Biol.*, 5, 40, 2001.
   (r) Mirzabekov, A. and Kolchinsky, A., *Curr. Opin. Chem. Biol.*, 6, 70, 2001.
   (s) Mitchell, P., *Nature Biotechnol.,* 20, 225, 2002.
   (t) Abbott, A., *Nature*, 415, 112, 2002.
   (u) Braun, P. and LaBaer, J., *Trends Biotechnol.*, 21, 383, 2003.
   (v) Cutler, P., *Proteomics*, 3, 3, 2003.
   (w) Forman, J. E., Suseno, A. D., and Wagner, P., *Methods Enzymol*., 361, 530, 2003.
   (x) Tyers, M. and Mann, M., *Nature*, 422, 193, 2003.
   (y) Gershon, D., *Nature*, 424, 581, 2003.
   (z) Hanash, S., *Nature*, 422, 226, 2003.

(aa) Wilson, D.S. and Nock, S., *Angew. Chem. Int. Ed. Engl.,* 42, 494, 2003.
(ba) Wingren, C. et al., *Nature Biotechnol.,* 21, 223, 2003.
(ca) Zhu, H., Bilgin, M. and Snyder, M., *Annu. Rev. Biochem.,* 72, 783, 2003.
(da) Cutler, P., *Proteomics*, 3, 3, 2003.
(ea) Barry, R. and Soloviev, M., *Proteomics*, 4, 3717, 2004.

2. Hanash, S., ed. "Special Issue: Protein microarrays" *Proteomics*, 3, 11, 2003 (all papers therein).
3. (a) Prime, K. L. and Whitesides, G.M., *J. Am. Chem. Soc.,* 115, 10714, 1993.
   (b) Harder, P. et al., *J. Phys. Chem.,* 102, 426, 1998.
4. (a) Houseman, B.T. et al., *Nature Biotech.,* 20, 270, 2002.
   (b) Houseman, B.T. and Mrksich, M., *Chem. & Biol.,* 9, 443, 2002.
5. Cha, T.-W. et al., *Proteomics,* 4, 1965, 2004.
6. Cha, T.-W., Guo, A., and Zhu, X.-Y., *Proteomics,* 5, 416, 2005.
7. (a) Groll, J. et al., *Biomacromolecules,* 6, 956, 2005.
   (b) Groll, J. et al., *Langmuir*, 21, 3076, 2005.
8. http://proteinlsides.com
9. Ostuni, E. et al., *Langmuir*, 17, 5605, 2001.
10. Harris, J.M. and Zalipsky, S., *Poly(ethlylene glycol) Chemistry and Biological Applications,* Plenum Press, New York, 1992.
11. (a) Chen, C.S. et al., *Science,* 276, 1425, 1997.
    (b) Whitesides, G.M. et al., *Annu. Rev. Biomed. Eng.,* 3, 335, 2001.
    (c) Qian, X. et al., *Anal. Chem.,* 74, 1805, 2002.
    (d) Ostuni, E. et al., *Langmuir,* 17, 5605, 2001.
12. Asay, D.B. and Kim, S.H., *J. Phys. Chem. B,* 109, 16760, 2005.
13. Ge, H., *Nucleic Acid Res.,* 28, e3, 2002.
14. Bussow, K. et al., *Nucleic Acid Res.,* 26, 5007, 1998.
15. Mendoza, L.G. et al., *Biotechniques*, 27, 778, 1999.
16. Lueking, A. et al., *Anal. Biochem.*, 270*,* 103, 1999.
17. MacBeath, G. and Schreiber, S.L., *Science*, 289, 1760, 2000.
18. Zhu, H. et al., *Nature Genetics*, 26, 283, 2000.
19. Kusnezow, W. and Hoheisel, J.D., *J. Mole. Recognit.,* 16, 165, 2003.
20. Seong, S.-Y. and Choi, C.-Y., *Proteomics,* 3, 2176, 2003.
21. Angenendt, P. et al*., Chromatog. A*, 1009, 97, 2003.
22. Kusnezow, W. et al., *Proteomics,* 3, 254, 2003.
23. Guilleaume, B. et al., *Proteomics* 5, 4705, 2005.
24. Kanno, S. et al., *J. Biotech.,* 76, 207, 2000.
25. Turkova, J. et al. *J. Chromatogr. B.,* 722*,* 11, 1999.
26. Wikstrom, M., Forsen, S., and Drakenberg, T., *Eur. J. Biochem.,* 235, 543, 1996.
27. Peluso, P. et al., *Anal. Biochem.,* 312, 113, 2003.
28. Turkova, J. et al., *J. Chromatogr.,* 597, 19, 1992.
29. Zhang, W. and Czupryn, M.J., *Biotech. Prog.,* 18, 509, 2002.
30. Rowe, C.A. et al., *Anal. Chem.,* 71, 433, 1999.
31. Karyakin, A.A. et al., *Anal. Chem.*, 72, 3805, 2000.
32. Anderson, G.P. et al., *Biosens. Bioelectron.,* 12, 329, 1997.
33. Vijayendran, R.A. and Leckband, D.E., *Anal. Chem*., 73, 471, 2001.
34. Shriver-Lake, L.C. et al., *Biosens. Bioelectron.,* 12, 1101, 1997.
35. Nisnevitch, M. et al., *J. Chromatogr. B Biomed. Sci. Applic.*, 738, 217, 2000.
36. Nisnevitch, M. and Firer, M.A., *J. Biochem. Biophys. Meth.,* 49, 467, 2001.
37. Hodneland, C.D. et al., *Proc. Natl. Acad. Sci. USA*, 99, 5048, 2002.
38. Braun, P., et al., *Proc. Natl. Acad. Sci. USA*, 99, 2654, 2002.

39. Weng, S. et al., *Proteomics*, 2, 48, 2002.
40. Kurz, M. et al., *Chembiochem.,* 2, 666, 2001.
41. Soellner, M.B. et al., *J. Am. Chem. Soc.* 125, 11790, 2003.
42. Porath, J. et al., *Nature*, 258, 598, 1975.
43. Keller, T.A. et al., *Superamol. Sci.*, 2, 155, 1995.
44. Sigal, G.B. et al., *Anal. Chem.,* 68, 490, 1996.
45. Schmid, E.L. et al., *Anal. Chem.,* 69, 1979, 1997.
46. Zhu, H. et al., *Science*, 293, 2101, 2001.
47. Hochuli, E. et al., *J. Chromatography*, 411, 177, 1987.
48. Hochuli, E., *Biologically Active Molecules*, Schlunegger, U., Ed., Springer-Verlag, Berlin, 1989.
49. Johnson, D.L. and Martin, L.L., *J. Am. Chem. Soc.,* 127, 2018, 2005.
50. Jiang, W. et al., *Anal. Biochem*., 255, 47, 1998.
51. Chaga, G.S., *J. Biochem. Biophys. Methods*, 49, 313, 2001.
52. Beckmann, J.D., *Anal. Biochem*., 197, 408, 1991.
53. Schenning, A.P.H.J. et al., *J. Am. Chem. Soc.,* 118, 8549, 1996.
54. Hahm, J. and Sibener, S.J., *Langmuir,* 16, 4766, 2000.
55. Ohara, P.C., Heath, J.R., and Gelbart, W.M., *Angew. Chem. Int. Ed. Engl*., 36, 1078, 1997.
56. Deegan, R.D. et al., *Nature,* 389, 827, 1997.
57. Clark, D.C. et al., *Faraday Discuss*., 98, 253, 1994.
58. Mackie, A.R. et al., *J. Coll. Interf. Sci*., 210, 157, 1999.
59. Delehanty, J.B. and Ligler, F.S., *Biotechniques,* 34, 380, 2003.
60. Konstantin, V.K., Kusnezow, W., and Langowski, J., *J. Chem. Phys.,* 6, 111, 2006.
61. Kusnezow, W. et al., *Proteomics,* 6, 794, 2006.
62. Deng, Y. et al., *J. Am. Chem. Soc.,* 128, 2768, 2006.

5 Fabrication of Sol-Gel-Derived Protein Microarrays for Diagnostics and Screening

Nicholas Rupcich and John D. Brennan

CONTENTS

INTRODUCTION

In recent decades analytical science has witnessed a rise in the utility of immobilized biomolecules for sensing applications. Biological recognition elements provide unsurpassed selectivity and specificity that is difficult to reproduce synthetically. Advances in biochemistry, molecular biology, and immunochemistry have allowed for a rapid expansion in the range of biological recognition elements used in the field of biosensing and solid-phase assays; with uses spanning the selective extraction, delivery, separation, conversion and detection of numerous target analytes. The employment of biomolecules such as enzymes, antibodies, DNA and membrane-bound receptors, and more complex biological entities such as organelles, microorganisms, animal and plant cells or tissues in these applications has typically relied heavily on their successful immobilization onto or within a suitable transducer surface.

One area that has emerged as a result of the success in protein immobilization is microarrays. These devices provide a facile route to allow miniaturization of conventional assays, which has been a general trend in biomedical research. Microarrays consist of spatially ordered elements, usually less than 300 μm in diameter, that are deposited or synthesized for the purpose of performing biochemical reactions in a parallel and high-throughput fashion. This format allows for true parallelism, miniaturization, multiplexing and automation, all key features that could not be achieved with earlier technologies. Together, these features lead to microscale assays that reduce reagent consumption, minimize reaction volumes, increase sample concentrations and accelerate reaction kinetics.

DNA microarray technology was the first format to use biomolecule immobilization in arrays of ordered spots, and emerged owing to both the success of the human-genome sequencing project and the relative ease with which DNA could be immobilized.[1,2] However, the realization that genetic information could not provide sufficient insight into the understanding of complex cellular networks, as well as the missing relationship between mRNA and protein abundance,[3,4] eventually led to the development of comparable technology for the analysis of proteins.[5,6] Initially, antibodies, being natural protein binding moieties, were immobilized in an ordered fashion on a solid support to create antibody microarrays;[5] and in parallel, protein microarray technology evolved for the study of protein interactions and modifications.[6,7] Although such arrays are envisaged to become a valuable tool for tasks such as the characterization of enzyme kinetics,[8,9] antibody specificity,[10,11] and the elucidation of gene function,[12,13] many limitations of the technology are still unsolved and prevent protein microarray technology from reaching its full potential. These limitations include the generation of protein libraries in large quantity, the conservation of protein function during immobilization, particularly for labile proteins such as membrane-bound receptors, and the need for high levels of immobilized protein to obtain sufficient sensitivity for detection and quantitation of binding interactions.

Methods used to immobilize biomolecules onto inorganic, organic or polymeric surfaces have typically been based on physical adsorption,[14] covalent binding to surfaces,[15] entrapment in semi-permeable membranes[16] and microencapsulation into polymer microspheres and hydrogels.[17,18] However, such techniques are not generic and in most cases can be used only for a limited range of biomolecules or applications. Additionally, problems related to leaching and desorption,[14] denaturation, and the orientational control of the biomolecule often result in the need for substantial optimization of the immobilization protocol each time a new biological species is used, making such methods time-consuming and labor-intensive.[19]

To meet the requirements for preparation of robust protein microarrays, several approaches have been proposed which can be broadly divided into three major groups. The first comprises spotting onto two-dimensional (2-D) plain glass slides which are activated with a variety of coupling chemistries such as aldehyde, epoxy or carboxylic esters. Slides with these surfaces bind proteins and antibodies either by electrostatic interactions or through the formation of covalent bonds. Although they offer several advantages, such as a strong attachment combined with low variation, they suffer from rapid evaporation of the liquid environment as well as close protein surface contact, which may affect protein three-dimensional structure. An alternative is formation of

arrays on three-dimensional (3-D) gel or membrane-coated surfaces, such as polyacrylamide,[20,21] agarose[22] and nitrocellulose.[23] These surfaces bind proteins mainly by physical adsorption and are expected to be the most favorable with regard to the preservation of native protein conformation. However, large variations in signal intensity and lack of orientational control are a disadvantage of these surfaces.[24] The third approach is a hybrid of the aforementioned methods, and includes spotting onto dendrimer or avidin-coated slides, which display a supramolecular structure on their surface yet are not formally 3-D layers. These surfaces have higher surface areas and binding capacities than conventional 2-D surfaces, but often require recombinant or labeled proteins to allow binding by affinity interactions.

An alternative route for bio-immobilization involves the entrapment of biological components into inorganic silicate matrixes formed by a low temperature sol-gel processing method.[25–28] Entrapment does not rely on either covalent or affinity-based interactions with the substrate, eliminating the need for derivatization of the protein and the potential for improper orientation of the biomolecule.[29] Sol-gel-derived microspots are inherently three-dimensional, thus allowing for higher protein loading than can be obtained from a 2-D monolayer. Additionally, multiple proteins can be simultaneously entrapped within the sol-gel matrix, permitting the use of coupled reactions from immobilized protein systems.[30,31] Finally, sol-gel materials can be used to entrap a variety of native proteins including membrane-bound proteins,[32–34] suggesting that microarraying of these clinically relevant species should be possible by this method. A potential disadvantage of sol-gel based microarrays is that they are likely to be amenable only to studies of protein–small molecule interactions, since it is not likely that large species such as proteins can enter the glass to interact with the entrapped protein. Even so, such microarrays should find use in areas such as small molecule screening (i.e., drug screening), multianalyte biosensing and metabolic profiling.

SOL-GEL-BASED BIOMOLECULE IMMOBILIZATION

Protein encapsulation via the sol-gel method involves forming a mesoporous silica network around the protein via polymerization of suitable silane precursors. The nanometer-scale pores allow analytes to diffuse freely in and out of the matrix while retaining the entrapped protein. While the earliest reports of protein entrapment in sol-gel-derived glasses appeared in the 1950s,[35] it was not until Braun and coworkers published a seminal paper in 1990 describing the entrapment of proteins in alkoxysilane derived glasses that the field began to bloom.[25] Since then, an enormous amount of work has been published describing the entrapment of a variety of biological species including enzymes, antibodies, regulatory proteins, membrane-bound proteins, nucleic acids and even whole cells into a range of sol-gel-derived nano-composite materials.[27,29,36]

Figure 5.1 shows a typical process to produce a protein-doped silica material. The formation of sol-gel-derived materials begins with the hydrolysis of a suitable silane precursor to form an aqueous sol. At present, the most common precursors are tetraalkoxysilanes, such as tetraethylorthosilicate (TEOS) or tetramethylorthosilicate (TMOS); however, it is possible to include several mono-, di-, and tri-substituted alkoxysilanes that incorporate alkyl, aryl, amino, carboxyl, thiol, or other functional groups to provide specific properties to the sol-gel material.[27,29] Hydrolysis of

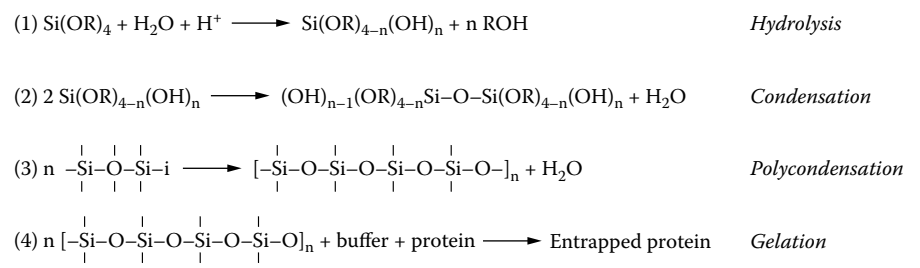


FIGURE 5.1 The sol-gel process for formation of protein-doped silica from tetraalkoxysilane precursors. Note: the bonds to Si are denoting further Si-O bonds. (From Brennan, J.D. Using Intrinsic Fluorescence to Investigate Proteins Entrapped in Sol-Gel Derived Materials. *Appl. Spectrosc*., 53, 106A–121A, 1999. With permission.)

the precursor can be achieved by either acid or base catalysis to form the sol. The hydrolyzed precursor is then mixed with a buffered aqueous solution containing the biomolecule of interest, along with any additives (polymers, osmolytes, templating agents, or any other material modifiers). The sudden shift in pH from either low or high values to the physiological range results in a rapid polycondensation of the silane and gelation of the material. The speed of the reaction is dependent upon the pH and ionic strength of the solution as well as the presence of polymerization catalysts; as a result gelation times can range from seconds to days. As the silica network ages over time, further cross-linking of the material occurs and the entrapped water and alcohol (from alkoxysilane hydrolysis) begin to evaporate, resulting in material shrinkage and a reduction in the submicrometer pore diameters (Figure 5.1).

While the majority of sol-gel based entrapment studies use TEOS or TMOS as the silane precursor, there are drawbacks to these precursors when used for the entrapment of biomolecules. The most important of these is the liberation of alcohol (ethanol or methanol) during the hydrolysis of these precursors, which can lead to rapid denaturation of the entrapped proteins. The fabrication of materials from these precursors also requires separate hydrolysis and condensation steps at different pH values. In addition, the resulting materials undergo excessive shrinkage and cracking as they evolve over time, which can hinder their use in applications requiring long-term protein stability.

To overcome these disadvantages there has been an effort to develop more biocompatible silane precursors, including (a) sodium silicate, a colloidal silica precursor[37] and (b) diglyceryl silane (DGS), a newly developed silane precursor that releases the protein stabilizer and humectant glycerol as the by-product of hydrolysis.[38] Recently synthesized silane precursors also exist that have covalently tethered sugars that can retain entrapped water, reduce shrinkage and cracking and ultimately help stabilize entrapped biomolecules.[39] Key advantages of such materials include the removal of alcohol as a hydrolysis by-product, the ability to process materials at neutral pH, and, in the case of DGS and sugar silanes, the presence of protein stabilizing species.

An advantage of protein-doped silicate materials is that it is possible to cast the protein-doped liquid sol in a variety of configurations prior to gelation. Formats can

include monolithic blocks or columns, powders, thin films, fibers and, as discussed here, pin-printed microarrays. All of these configurations provide different levels of performance based on the parameters of protein loading, desired response time, sensitivity and detection limits, and the ability to interface the material to commonly used analytical devices. The versatility in terms of formatting leads to the ability to use such materials for a variety of applications, including: selective coatings for optical and electrochemical biosensors; stationary phases for affinity chromatography; immunoadsorbent and solid-phase extraction media; solid-phase biocatalysts; controlled release agents; unique matrices for biophysical studies, and media for fabrication of protein microarrays.[28] In this article we review the various aspects of fabricating sol-gel microarrays and the applications of sol-gel-based microarrays, highlighting novel aspects of this approach, particularly for "multicomponent" protein arrays.

FABRICATION OF SOL-GEL-DERIVED MICROARRAYS

While there are a number of methods for fabricating microarrays, including stamping, pin-printing and ink-jet deposition, we chose to use the pin-printing method owing to the ease of adapting this method for printing sols, the ability to control spot sizes by simply changing the pin or printing speed, and the potential for using multiple pins in parallel to accelerate array fabrication. Although the sol-gel route provides significant potential as a method for preparation of pin-printed protein microarrays, it was necessary to address several issues in order to develop a robust fabrication method. For example, the pin-printing of solutions that are undergoing changes in viscosity and cross-linking prior to gelation may result in irreproducible spot sizes or even clogging of the pins if the gelation time is too fast. The spots, once printed, must remain adhered to the substrate and resist cracking as a result of analyte introduction and washing cycles. Furthermore, the entrapped biomolecule must remain functional and accessible but must also resist leaching from the microspot. The effects of variables such as the surface chemistry of the substrate, the nature of the sol-gel precursor, the type and level of buffer, the water-to-silane ratio, the pH of the sol and the presence of the protein stabilizing agent glycerol on the properties of the resulting microarray were first examined by Rupcich et al.[40]

As expected, the printability of the material was dramatically affected by the gelation time of the sol. While in some cases it was possible to print 100 spot arrays using solutions with gelation times as short as 10 min, optimal printing without clogging of pins requires gelation times of at least 20 min, although longer gelation times were generally used for printing of arrays. The choice of precursor, printing pH, buffer type and ionic strength as well as the use of small molecule or polymer additives also affected both the gelation time and the cracking/adhesion of spots. Four different silica precursors, including tetraethlyorthosilicate (TEOS), sodium silicate (SS), monosorbitol silane (MSS) and diglyceryl silane (DGS), were investigated when printed onto three different surfaces: bare glass slides, aminopropylsilane-coated slides, and epoxy-coated slides. Printability studies demonstrated that a sol formed by mixing 100 m*M* Tris:HCl buffer (pH 8.0) containing the protein of interest with a sodium silicate solution (2.8 g per 10 ml of H_2O, first brought to pH 4.0 using Dowex 50x8-100 cation exchange resin and then filtered) provided the best

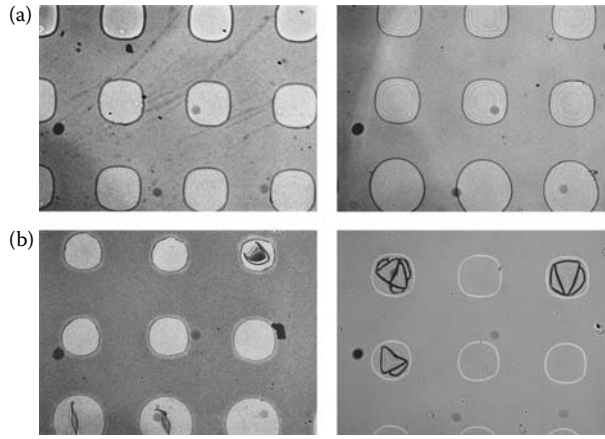


FIGURE 5.2 Sodium-silicate-derived spots pin-printed onto epoxy-coated slides, before (left) and after (right) washing with aqueous buffer solution. (a) spots containing no glycerol, (b) spots containing 25% (v/v) glycerol in the original sol. (From Rupcich, N., et al., Optimization of sol-gel formulations and surface treatment for the development of pin-printed protein microarrays, *Chem. Mater.*, 15, 1803–1811, 2003. With permission.)

printing performance based on gelation times and pin-clogging. The uniformity of the printed spots and adherence of spots to the substrate was enhanced with increasing slide surface hydrophobicity and was best on epoxy derivatized slides. The addition of glycerol improved the viscosity and increased gelation times; however, its presence was detrimental to microspot integrity, as it resulted in cracking and poor adhesion of spots, as shown in Figure 5.2.[40]

To illustrate the three-dimensionality of the sol-gel-derived arrays, our group also examined images of the pin-printed spots side-on (Figure 5.3). Based on volumetric

FIGURE 5.3 Profile images of sodium-silicate-derived microarray spots printed on epoxy-derivatized surfaces. (a) Microspots containing both anti-fluorescein antibody and 25% glycerol, (b) horizontal profile of microspots containing 25% glycerol (no protein). Spots are ~100 µm diameter, interspot spacings are 500 µm, images are 0.8 × 1.1 mm. (From Rupcich, N., et al., Optimization of sol-gel formulations and surface treatment for the development of pin-printed protein microarrays, *Chem. Mater.*, 15, 1803–1811, 2003. With permission.)

calculations it was estimated that the three-dimensional nature of the sol-gel spots allowed for a 50-fold enhancement in protein loading relative to an immobilization of a close-packed monolayer of an antibody. The increase in loading was confirmed by pin-printing a fluorescein-loaded antifluorescein antibody solution directly onto Super-Aldehyde slides substrates to form a monolayer, and a sodium silicate-based solution containing an identical antibody concentration to form sol-gel microarray spots. Based on the fluorescence intensity of spots in the two samples, it was determined that there was over 100 times more antibody in the sol-gel-derived arrays — and an enhanced signal to background ratio of 300:1 vs. 7:1 relative to the covalently bound monolayer system.[40]

To demonstrate the activity and selectivity of the entrapped antifluorescein antibody, a 10×10 sodium silicate-based microarray was produced that contained all necessary controls, including the antifluorescein antibody (target protein), a blank sample consisting of only sodium silicate with buffer, a positive fluorescence control containing entrapped fluorescein dextran (70,000 MW), and a selectivity control consisting of entrapped anti-dansyl antibodies. These controls ensure that fluorescence emission from spots in the microarray following doping of the microarray with fluorescein and washing is solely due to the activity and selectivity of the antifluorescein antibody and is not due to nonspecific adsorption of fluorescein to the sol-gel surface or nonselective binding to the antibody. Figure 5.4 illustrates (a) the initial fluorescence of the array prior to adding fluorescein; (b), the fluorescence response after adding fluorescein over the entire array and (c) the pattern of fluorescence following washing to remove unbound fluorescein. The data clearly show the presence of all printed spots (Panel b), showing that no spots wash off the surface, and demonstrate that the fluorescein does not bind nonselectively to the array spots and that the antibody remains active in the array.

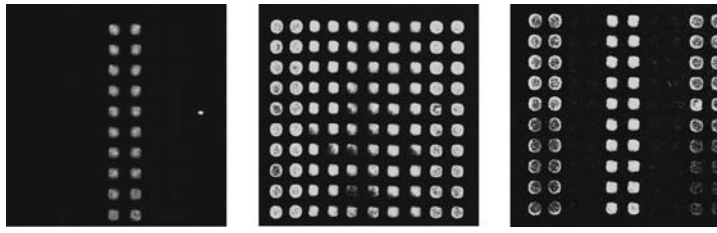


FIGURE 5.4 Sodium-silicate-based antibody array. Columns 1, 2, 9 and 10 contain anti-fluorescein antibody, columns 3 and 4 are blanks (sodium silicate only), columns 5 and 6 contain fluorescein dextran and columns 7 and 8 contain anti-dansyl antibody. (a) Fluorescence image before doping with fluorescein; (b) fluorescence image after adding fluorescein; (c) fluorescence after washing to remove unbound fluorescein. Spot sizes are ~100 _m in diameter, separation is 150 μm, image area is 1.6×1.6 mm. (From Rupcich, N., et al., Optimization of sol-gel formulations and surface treatment for the development of pin-printed protein microarrays, *Chem. Mater.*, 15, 1803–1811, 2003. With permission.)

One of the advantages of the microarray format is the ability to use the method to perform high-throughput screens for optimal sol-gel material formulations that lead to maximum protein activity, as demonstrated in pioneering work by Cho et al.[41] In one example, over 900 biodegradable polymer formulations were prepared in a microarray format and assayed to determine the optimal composition to maintain the viability of entrapped keratinocyte growth factor (KGF). The formulations were based on varying molecular weights of polylactic acid (2K to 300K Da) and additives including polyglycolic acid and the surfactant sodium bis(ethylhexyl) sulfosuccinate (AOT). Total polymer content within a given formulation was maintained at 3% by weight and each sample contained 15 ppm KGF in either phosphate or Tris buffer. The intrinsic fluorescence emission spectrum of KGF within each array spot was assessed to determine the extent of the protein denaturation. Of the 900 samples, only 6 formulations produced KGF emission spectra equivalent to the native spectrum of the protein in buffer and remained stable for over one month when stored at 4°C.

In a second example, over 600 silica formulations were examined in microarray format to find compositions that maximized the signaling capability of entrapped antifluorescein antibody activity. In this case, TMOS-derived materials containing varying amounts of the additives aminopropyltriethoxysilane (APTES), Nafion, polyethyleneimine (PEI, 70 kDa), polyethyleneoxide (PEO, 100 kDa) and dextran (25 kDa) were pin-printing on plain glass slides and the fluorescence intensity of the entrapped antibody was measured after addition of fluorescein and washing. In this case, over 80% of the formulations demonstrated some level of fluorescein binding. However, complex formulations, such as 95% TMOS, 4% APTES, 1% Nafion that was mixed in a 1:1 volume ratio with 20 m*M* Tris buffer at pH 6.2 provided optimal binding.[41] This composition would not be predicted to be optimal, and shows the utility of using the array method for screening materials to provide optimal protein performance.

The examples presented above provide interesting insights into the use of the sol-gel method for array fabrication. In the case of unstable proteins, screening many compositions allows for the identification of materials that provide proteins with the correct conformation, even in cases where less than 1% of the compositions tested provided a useful immobilized protein. In the case of robust proteins, such as antibodies, a large fraction of compositions lead to good activity; in this case it is then possible to choose an ideal composition on the basis of other criteria, such as ease of printing or long-term protein stability. This demonstrates the versatility of the sol-gel approach in that it can be modified to suit a wide range of biomolecules.

Another advantage of the use of a sol-gel-based material for fabrication of microarrays is the ability to utilize unconventional formats for preparing arrays. As an example, Bright's group has shown that sol-gel-based microarrays can be deposited into micromachined microwells[42] or pin-printed onto the surface of planar light emitting diodes[43] (LEDs) to create self-contained chemical sensors. Figure 5.5 illustrates a schematic of the micromachined LED used to create a portable and inexpensive oxygen sensor. In this system, wells of either 250 or 500 μm diameter were drilled into the flat surface of the LED, followed by filling of the

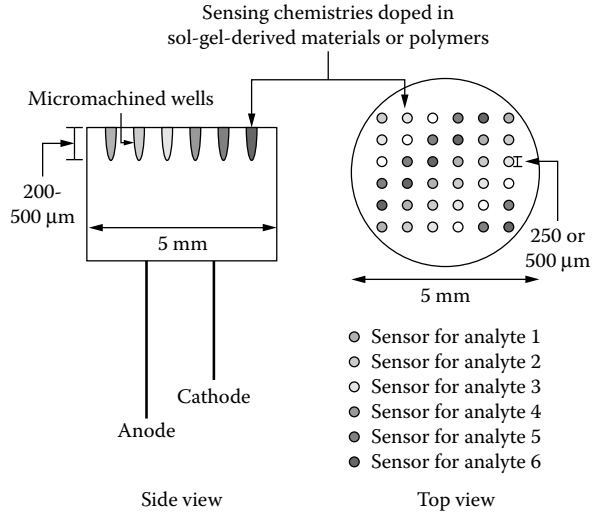


FIGURE 5.5 Simplified schematic of an optical sensor array integrated into a LED light source. (From Cho, E.J. and Bright, F.V., Optical sensor array and integrated light source, *Anal. Chem.*, 73, 3289–3293, 2001. With permission.)

wells with a TEOS-based sol containing the O_2-responsive luminophore (tris(4,7-diphenyl-1,10-phenanthroline)ruthenium(II) $[Ru(dpp)_3]^{2+}$. The self-contained LED sensors provided reversible signaling to alternating streams of N_2 and O_2 gas and had good reproducibility, boding well for the advancement of such devices for remote sample analysis.

SOL-GEL-DERIVED ENZYME MICROARRAYS

The first example of enzyme entrapment in a sol-gel microarray format was performed by Cho et al.[44] They developed stable and robust biosensors for detecting glucose and O_2, based on the immobilization of the enzyme glucose oxidase (GOx) and the oxygen sensitive dye tris(4,7′-diphenyl-1,10′-phenanathroline)ruthenium(II) chloride pentahydrate ($[Ru(dpp)_3]Cl_2 \bullet 5H_2O$) in TMOS-derived silica materials. Polyethylene glycol (PEG), Pluronic 104 (P104) and sorbital were added to the TMOS to help produce crack-free spots with extended gelation times to avoiding pin clogging. Arrays were formed by pin-printing $[Ru(dpp)_3]^{2+}$ onto either glass slides or onto the surface of a planar LED. This was followed by either spin casting a second layer of sol-gel material containing glucose oxidase over the existing microarray or by overprinting a second layer of GOx-doped silica over the microarray to form a layered microarray element. In this system, consumption of O_2 by the GOx catalyzed oxidation of glucose leads to a reduction in quenching of the luminophore, and a corresponding increase in fluorescence intensity.

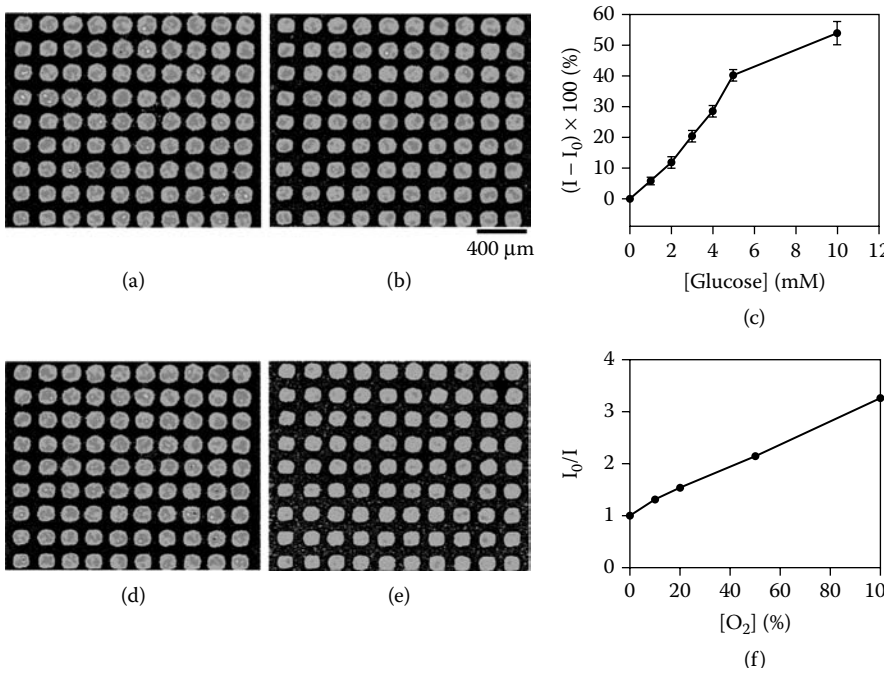


FIGURE 5.6 Glucose and O_2 sensing on layered microarrays. Arrays elements contained an oxygen sensitive $[Ru(dpp)_3]^{2+}$ dye and glucose oxidase, and remained sensitive to either glucose exposure (Panels A–C) or O_2 saturated buffer (Panels D–F). (From Cho, E.J. and Bright, F.V., Pin-printed biosensor arrays for simultaneous detection of glucose and O_2. *Anal. Chem*., 74, 6177–7184, 2002. With permission.)

Figure 5.6 summarizes the response characteristics of their array produced by spin-coating GOx-doped silica over the existing O_2–sensitive array. Panels (A) and (B) show the array in response to air-saturated buffer containing no glucose and air-saturated buffer containing 10 m*M* glucose, respectively, while Panel (C) shows the relative increase in intensity as a function of glucose concentration. Panels (D) and (E) show the O_2-dependent response to N_2 and O_2 saturated buffer, respectively, while Panel (F) shows the Stern-Volmer response of the fluorescence as a function of $[O_2]$.

The success of this demonstration highlights the potential for creating layered samples with sol-gel microarrays, thus making use of the third dimension, something that is difficult to do with other technologies. The use of the overlayering method also provides a useful route to allow entrapment of small molecules, such as the $[Ru(dpp)_3]^{2+}$ luminophore, without leaching, thus allowing the formation of protein arrays with incorporated signaling elements.

Park and Clark demonstrated numerous examples of sol-gel-derived protein arrays based on millimeter-scale sol-gel elements placed within microwells constructed from PDMS on a microscope slide.[45] Among the enzymes used were

numerous hydrolases as well as co-entrapped GOx and horseradish peroxidase (HRP). Using a methyltrimethoxysilane and polyvinyl alcohol sol containing the protein of interest, they pipetted 5 μl volumes of solution into 1.8 mm diameter wells created by puncturing a 1.7-mm-thick PDMS film which was attached to a glass microscope slide. Absorbance-based measurements were used in conjunction with the common indicator dye bromothymol blue to measure pH variance due to the various hydrolysis reactions. Park et al. were able to closely correlate the solution-based activity of 20 different hydrolases from various sources to that of the entrapped enzyme array assay, as well as measure the inhibition of active hydrolases with the inhibitor chymostatin.

Figure 5.7 illustrates the size of the arrays constructed and the formation of a colored dye solution due to GOx/HRP activity. This example demonstrates another advantage of sol-gel-based array fabrication; the ability to form array elements that contain multiple proteins. This has significant implications in terms of performing coupled enzyme reactions on arrays, which is often useful for generating signals from enzymatic reactions.

Our group reported on the further development of protein microarrays based on the co-immobilization of multiple components within a single pin-printed sol-gel array element.[46] Two different enzyme-based systems were pin-printed using sodium silicate as the silane precursor: (a) a coupled two-enzyme reaction involving glucose oxidase and horseradish peroxidase along with the fluorogenic reagent Amplex Red, allowing fluorimetric detection of glucose, and (b) the co-immobilization of urease with fluorescein-labeled dextran to detect the hydrolysis of urea based on a pH-induced change in fluorescein emission intensity as a result of the production of ammonium carbonate. Using an epifluorescence microscope for array imaging, it was possible to follow the time-dependent changes in intensity from the array, as shown in Figure 5.8 for the GOx/HRP system. An advantage of using the array format was that all selectivity controls as well as positive and negative fluorescence controls could be included to alleviate the potential for false signaling.

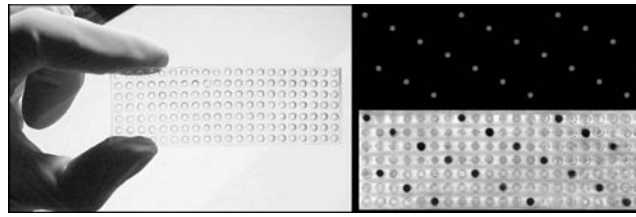


FIGURE 5.7 Arrays made by Park and Clark. The wells within the PDMS slide coating contained enzymes entrapped in alkoxysilane derived gels. The darker spots in the array on the right indicate glucose oxidase and horseradish peroxidase activity using 4-aminoantipyrine and p-hydroxybenzene sulfonate as dye components for the absorbance based assay. (From Park, C.B. and Clark, D.S., Sol-gel encapsulated enzyme arrays for high throughput screening of biocatalytic activity, *Biotechnol. Bioeng*., 78, 229–235, 2002. With permission.)

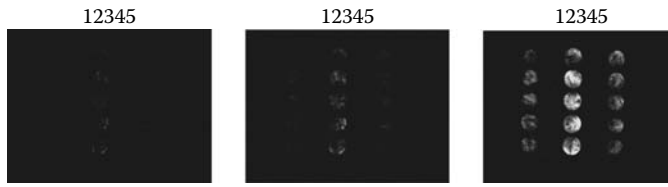


FIGURE 5.8 5 × 5 microarray of glucose oxidase/horseradish peroxidase co-immobilized in sol-gel-derived glass. Columns 1 and 5 contain GOx/HRP co-immobilized with Amplex Red (coupled reaction site), column 2 contains only buffer and Amplex Red and acts as a negative control, column 3 contains GOx/HRP and glucose along with partially reacted Amplex Red, and acts as a positive control. Column 4 contains only GOx and Amplex Red and serves as a negative control. The first panel shows the array before the addition of glucose (only column 3 is fluorescent owing to the presence of resorufin). The middle panel shows the array 1 min after addition of glucose and the third panel shows the array 12 min after glucose addition, showing the time dependence of the enzyme catalyzed reaction. All spots are 100 mm wide. (From Rupcich, N. and Brennan, J.D., Coupled enzyme reaction microarrays based on pin-printing of sol-gel biomaterials, *Anal. Chim. Acta*, 500, 3–12, 2003. With permission.)

An advantage of being able to perform time-dependent imaging studies on the arrays was the ability to extract both enzyme kinetic data and inhibition constants. Table 5.1 summarizes the results of the kinetic experiments performed for both GOx/HRP and urease (co-entrapped with fluorescein-dextran) in solution, bulk silica materials and microarrays. As shown in Table 5.1, the K_M values for

TABLE 5.1
Kinetic Parameters for Substrate Turnover and Enzyme Inhibition for Free and Entrapped Enzymes and for Enzyme Microarrays

	GOx/HRP		Urease/FD		
	K_m(μM)	K_{cat} (S-1)	K_m (μM)	K_{cat} (S-1)	K_I
Solution	103 ± 9	$9 \pm 1 \times 10^5$	1.3 ± 0.2	78 ± 2	48–85[a]
Entrapped enzyme in plate reader	188 ± 4	$1.9 \pm 0.3 \times 10^5$	2.35 ± 0.03	1.33 ± 0.02	54 ± 2
Microarray	58 ± 3	$4.9 \pm 0.3 \times 10^4$	1.9 ± 0.1	1.1 ± 0.1	62 ± 7

[a] The range of K_I values is due to enzyme activity fluctuations at different pH values (5.5 to 8).

Source: From Rupcich, N. and Brennan, J.D., Coupled enzyme reaction microarrays based on pin-printing of sol-gel biomaterials, *Anal. Chim. Acta*, 500, 3–12, 2003. With permission.

entrapped enzymes were in all cases within a factor of two of the value in solution and are in good agreement with the literature values. On the other hand, k_{cat} values were significantly lowered upon entrapment, with the value for the entrapped protein being up to 70-fold lower than in solution. Decreases in the catalytic rate constant for entrapped enzymes has been reported by several groups[47–50] including our own,[51] and is expected based on the tortuous path that must be taken to allow diffusion of small molecules through the porous network of the silica.[52] The data show that (a) concentration dependent fluorescence responses can be obtained on a microarray; (b) "reagentless" assays can be done conveniently on an array; and (c) entrapped enzymes on an array follow Michaelis–Menten kinetics. It was also demonstrated that inhibition constants (K_I) for small molecule inhibitors could be obtained (for urease), based on changes in enzyme kinetic constants in the presence of various inhibitor concentrations. In this case, the K_I values were within error of the solution values, demonstrating the potential of sol-gel-based microarrays as a format for inhibitor screening.

The use of co-entrapped enzymes for the development of multianalyte sensor arrays for renal clinical analytes was demonstrated by Doong's group using TMOS-based sol-gel formulations.[53] In this work, relatively large wells of 600 μm diameter and 10 μl volume were used to form a multianalyte sensor to measure conversion of glucose, urea, creatinine and uric acid. Using sensing systems similar to those of reported above,[46] GOx and HRP were co-entrapped with Amplex Red to measure glucose and urease was co-entrapped with fluorescein dextran to measure urea levels. In addition, the enzyme uricase was coupled to HRP and Amplex Red to measure uric acid levels and creatinine deaminase was coupled with fluorescein dextran to measure creatinine conversion. The array was able to accurately detect the four analytes when present in fetal calf serum, showing the potential for utilizing sol-gel microarrays for clinical applications.

Doong et al. followed this work by using either acetylcholinesterase (AChE) or urease co-entrapped with fluorescein-dextran[54] and rhodamine-labeled dextran to detect the activity and inhibition of the enzymes. The use of two probes, one pH sensitive and the other not, provided a means to perform ratiometric intensity measurements to overcome problems with photobleaching or leaching of the dyes. In these arrays a PVA/glycerol/TMOS composite material was utilized to allow printing of crack-free spots. The array-based sensor was used for detection of acetylcholine using AChE, and for the trace detection of the metal ions Cd(II), Cu(II), and Hg(II) based on inhibition of urease activity.[55]

SOL-GEL-DERIVED MEMBRANE PROTEIN MICROARRAYS

One of the key criteria for the development of new drugs is their ability to modulate the target of interest without causing cytotoxic side effects. The standard method for assessing both metabolism and toxicity of drugs and their products is to determine their interaction with cytochromes P450, which are the primary liver enzymes responsible for clearance of drugs from the body. Compounds that are metabolized to cytotoxic products as well as compounds that inhibit normal P450 function need

to be identified early in the drug development process at a rate that is commensurate with the rate of high-throughput screening.

A particular challenge in developing cytochrome P450 assays is the fact that the P450 complex involves a series of enzymes (cytochrome P450-3A4, cytochrome b5, NADPH reductase) that are present in the membrane of microsomes. Thus, special precautions need to be taken to ensure the viability of the membrane-associated enzymes during the array fabrication step. In recent years, several reports have emerged describing the use of sol-gel methods for entrapment of a wide range of membrane-bound proteins,[49] including bacteriorhodopsin-ATP synthase,[50] the acetylcholine and dopamine receptors,[51] and photosystem I.[52] Thus, it was expected that sol-gelderived materials may provide a route to fabricate a P450 microarray.

Clark's group developed microarrays containing baculosomes of cytochrome P450 enzymes and demonstrated the coupling of the arrays to cell-based screening to develop a method for evaluating prodrug toxicity.[53] Their metabolizing enzyme toxicology assay chip (MetaChip™) integrates the high-throughput, metabolite-generating capability of P450 catalysis with human cell-based screening on a microarray platform, allowing for rapid and inexpensive assessment of metabolism and toxicity. As shown in Figure 5.9, a methyltrimethoxysilane (MTMS)-derived sol-gel microarray was first produced that contained either one or both of the human P450 isoforms CYP3A4 and CYP2B6 and a regeneration system (glucose-6-phosphate and glucose-6-phosphate dehydrogenase). The MTMS sol solution was prepared by sonicating 250 µl MTMS with 100 µl of 5 m*M* HCl for 10 minutes. The second component is a monolayer of human MCF7 breast cancer cells within a chamber slide. The application of a 60 nl solution of lead compound (prodrug) is applied to the 30 nl sol-gel spots by using a microarrayer, in order to catalyze the release of active metabolites. The cancer cell monolayer was then stamped onto the sol-gel array and incubated for 6 hours at 37°C. Following incubation the cell layer was removed and the cells were stained using a live/dead test kit to determine the percentage of dead cells by using a microarray scanner.

It was clearly shown that the CYP-containing arrays could convert the nontoxic pro-drug cyclophosphamide into the cytotoxic chemotherapeutic drug 4-hydroxycyclophosphamide, as indicated by site specific cell death on the overlaid cancer cell slide. Controls showed that less than 13% cell death was obtained in spots containing no P450 in the presence of CP. The sensitivity of the MetaChip was compared to P450 solution reactions with CP as well as 5-fluoro-1-(tetrahydro-2-furfuryl)-uracil (Tegafur) and acetaminophen, which yield the cytotoxic compounds 5-fluorouracil and N-acetyl-p-benzoquinone-imine, respectively. Figure 5.10 illustrates the cytotoxicity results for each of the three prodrugs and the correlation to solution assays.

An important aspect of this work was the extension of sol-gelderived microarray technology to membrane-bound proteins, and the pharmacologically important P450 family of enzymes in particular. Coupling the entrapped P450s with a cell-based cytotoxicity test demonstrates a clever manipulation of standard microarray readout methods. Use of the live/dead cell system works in part because the metabolized compound remains in the array element and thus can

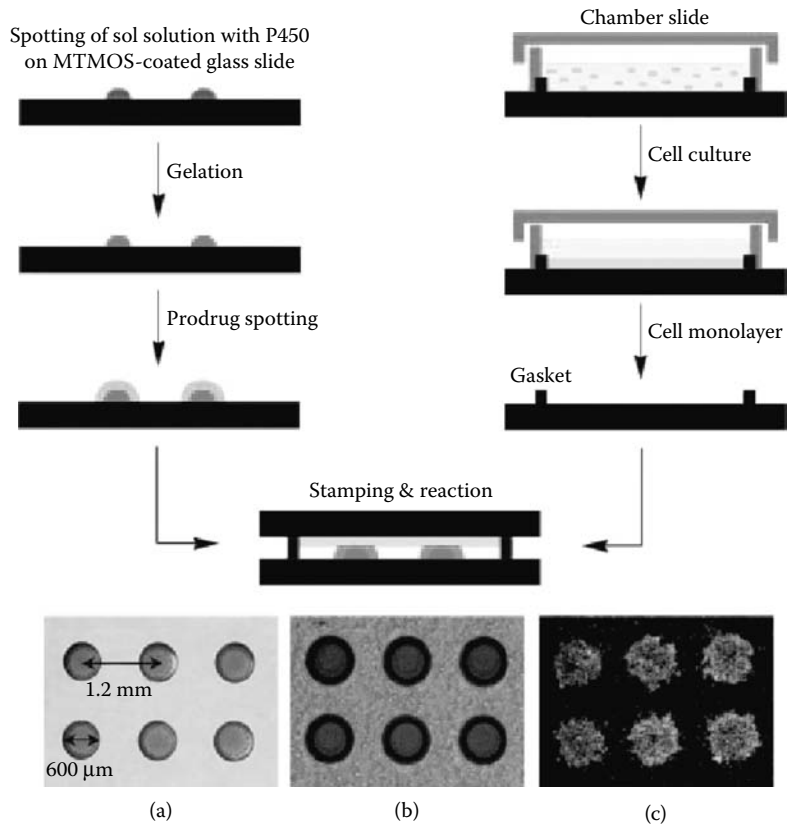


FIGURE 5.9 Schematic of MetaChip™ platform. Shown are: (A) 30 nL P450 sol-gel spots; (B) 30 nL sol-gel spots with 60 nL of prodrug solution after being stamped by MCF7 cell monolayer and; (C) the MCF7 cell monolayer after removal from the sol-gel array and staining. (From Lee, M.-Y., et al., Metabolizing enzyme toxicology assay chip (MetaChip) for high-throughput microscale toxicity analyses. *Proc. Natl. Acad. Sci. USA*, 102, 983–987, 2005. With permission.)

be blotted onto the cell bed. This would not likely be the case had adsorption or covalent attachment of the P450 to the substrate surface been done, since the product could simply diffuse away. The versatility of the sol-gel method is also highlighted by the fact that hydrophobic materials derived from MTMS worked well for the P450 entrapment, which is usually not the case for more polar soluble proteins. This demonstrates the flexibility that commercially available silane precursors can provide to fabricate materials that are specifically designed to stabilize a particular class of protein. The success of this assay also bodes well for the eventual use of sol-gel-derived microarrays use for other important membrane proteins like GPCRs or nuclear receptors, which have recently been microarrayed using other formats.[61–63]

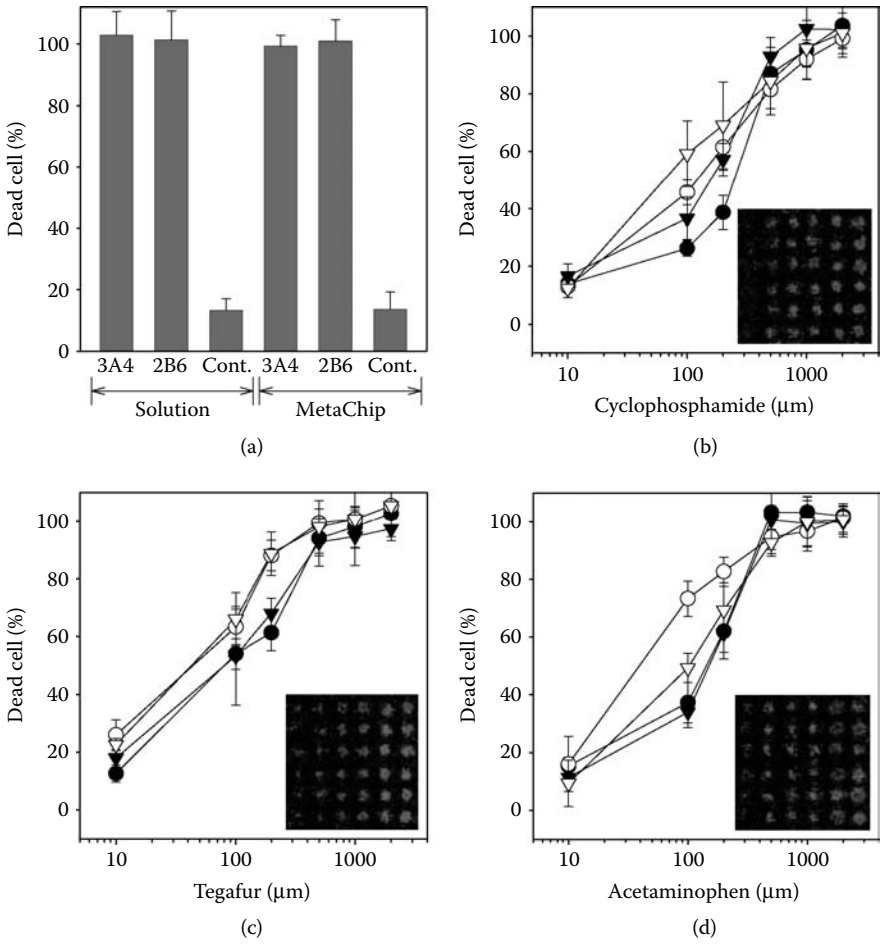


FIGURE 5.10 Comparison of cytotoxicity results for the MetaChip and solution-phase reactions. (A) Cytotoxicity of P450-activated CP for solution and sol-gel incubations. Control incubations consisted of all system components, except for a P450 isoform. (B) Effect of CP concentration on the cytotoxicity of MCF7 breast cancer cell for: 3A4 solution (●), 3A4 sol-gel (○), 2B6 solution (▼), and 2B6 sol-gel (▽). (C) Effect of Tegafur concentration on the cytotoxicity of MCF7 breast cancer cell cells for: 1A2 solution (●), 1A2 sol-gel (○), 3A4 solution (▼), and 3A4 sol-gel (▽). (D) Effect of acetaminophen concentration on the cytotoxicity of MCF7 breast cancer cell cells for: 3A4 solution (●), 3A4 sol-gel (○), 2B6 solution (▼), and 2B6 sol-gel (▽). In B–D, images from the array scanner are presented. In each 6 × 6 array segment, the columns represent different concentrations of spotted compounds (from left to right: 10, 100, 200, 500, 1000 and 2000 μM), and the rows represent replicates. (From Lee, M.-Y., et al., Metabolizing enzyme toxicology assay chip (MetaChip) for high-throughput microscale toxicity analyses. *Proc. Natl. Acad. Sci. USA*, 102, 983–987, 2005. With permission.)

KINASE-SUBSTRATE MICROARRAYS FOR SCREENING APPLICATIONS

Protein phosphorylation by kinases is an important mechanism in several intracellular processes and signaling cascades. The family of human protein kinases consists of over 500 members, of which only a fraction have yet been characterized.[64,65] Following G protein-coupled receptors (GPCRs), kinases are currently the most important target family of proteins for drug discovery, due to their involvement in therapeutic areas such as cancer,[66] inflammation[67] and diabetes.[68]

With pharmaceutical compound libraries surpassing the size of one million chemicals, there has arisen a need for the development of high-throughput assays that use minimal volumes of reagents. Thus, traditional techniques used to identify kinase substrates such as genetic screens, yeast two-hybrid approaches and biochemical purifications have become overly laborious and unreliable.[69] Kinase arrays provide a means of screening hundreds of miniaturized samples in parallel, allowing for relatively fast, easy and cheap determination of kinase action on numerous substrates at once. Zhu et al. used protein arrays to determine phosphorylation activity for 119 of the 122 known *Saccharomyces cereevisiae* kinases on 17 different substrates by using PDMS microwells as a solid support for immobilization.[70] Similarly to protein chips, arrays of immobilized peptides can be used to determine preferred sequences for phosphorylation by a kinase.[71–73] In this case, the peptide arrays can be incubated with the kinase of interest and [γ-32P]ATP and the levels of phosphorylation can be determined by phosphoimaging. Alternatively, mass spectrometry can be used to monitor kinase activity on the surface of peptide chips, avoiding the need for labeled reagents and simplifying assay formatting.[74]

Unfortunately, current peptide and protein chip strategies have several limitations. The first is the unwanted adsorption of soluble proteins, which can often compete with detection of protein-substrate interactions, leading to higher background levels of signal.[75] Secondly, only a fraction of the immobilized proteins are competent to participate in binding interactions since many of them are immobilized in inaccessible orientations or are denatured to some extent, both of which compromise their ability to interact with substrates.[76] A consequence of these limitations is that most immobilization procedures are not well suited for quantitative assays of protein-substrate interactions. Therefore protein and peptide chips have only been used in a surveying manner to generate a set of "hits," which are then evaluated and validated using more tried and tested solution-based assays.

Recently our group reported on a kinase microarray based on the co-immobilization of both kinase and substrate components within a single pin-printed sol-gel microarray element and used the arrays for nanovolume inhibition assays.[77] Using the α-catalytic subunit of cAMP dependent protein kinase (PKA) and the peptide substrate kemptide as a model system, the ability to monitor both phosphorylation and inhibition was demonstrated with Pro-Q Diamond™ dye[78] as an endpoint indicator of phosphorylation.

Compatibility of the stain with sol-gel materials as well as phosphoprotein detection limit and linearity were demonstrated using a β-casein concentration gradient pin-printed in an array format. Our experiments exhibited the ability to selectively

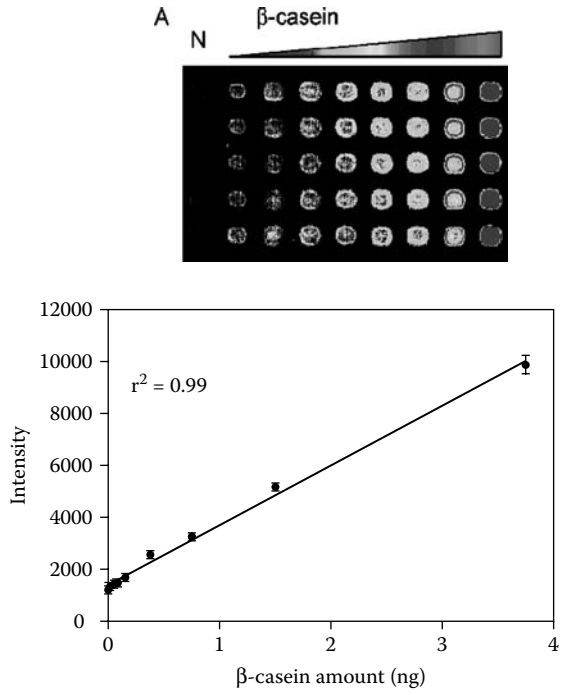


FIGURE 5.11 Linearity of phosphoprotein detection with Pro-Q Diamond dye within a sol-gel-derived microarray. Panel A shows the fluorescence intensity of the protein gradient on the array. Panel B shows the correlation between signal intensity and amount of protein. (From Rupcich, N. et al., Nanovolume kinase inhibition assay using a sol-gel-derived multi-component microarray, *Anal. Chem*., 77, 8013–8019, 2005. With permission.) See color insert following page 236.

detect phosphoproteins over nonphosphorylated controls and the ability to detect β-casein over a 500-fold concentration range (Figure 5.11). Limits of detection for β-casein were 7.5 pg and the detectable signal remained linear up to 3.75 ng of protein per array spot, which compared well to the original report on the Pro-Q concentration response, which claimed detection limits of 2 to 10 pg for three different phosphorylated peptides and a linear range of 130-fold.[78]

To demonstrate the utility of the co-immobilized kinase-substrate system for quantitative inhibition assays, 14 × 5 arrays were printed to determine IC_{50} values for the two PKA inhibitors, H7 and H89. The arrays contained three types of samples: (a) 10% w/v BSA as a negative control; (b) 50 μ*M* β-casein as a positive control and; (c) 12 columns of co–immobilized PKA and kemptide as test spots. Once printed, each of the columns in the arrays were overprinted with either buffer (two control columns) or one of twelve inhibitor concentrations in a gradient which straddled the respective literature IC_{50} value of the specific inhibitor. In all cases overprinting delivered approximately 0.6 nl of solution per array spot. The overprinted arrays were incubated for 30 minutes with the inhibitors, followed by

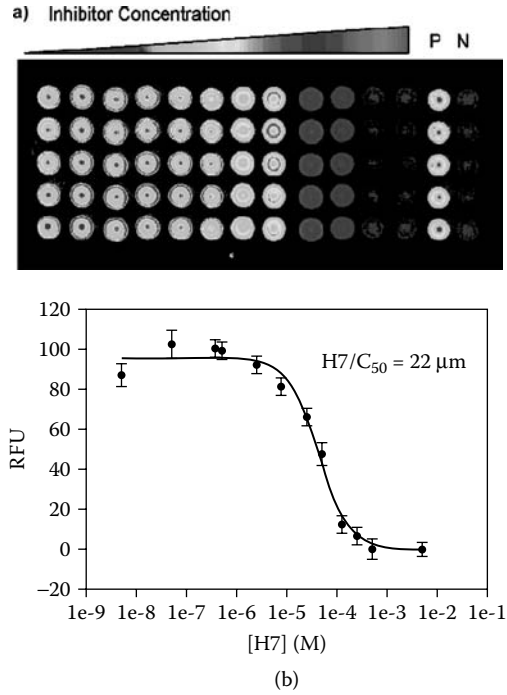


FIGURE 5.12 (a) H7 IC_{50} assay performed on a PKA/kemptide array. Inhibitor concentration increases from left to right, resulting in decreased fluorescence intensity due to inhibition of the phosphorylation reaction. N is the BSA negative control, P is the β-casein positive control. (b) IC_{50} curve generated from the H7 inhibition assay. Background signals from the negative control sample were subtracted and the data was normalized to the maximum intensity obtained in the absence of inhibitor. (From Rupcich, N. et al., Nanovolume kinase inhibition assay using a sol-gel-derived multi-component microarray, *Anal. Chem.*, 77, 8013–8019, 2005. With permission.) See color insert following page 236.

overprinting of each column with a solution of 50 µ*M* ATP containing the respective inhibitor at the concentration previously exposed to that column to avoid dilution effects. The reaction was allowed to ensue for 30 minutes prior to blocking, staining and imaging. Figure 5.12 shows the resultant array image and IC_{50} plot for the inhibitor H7. The experimental results provided IC_{50} values of 44 µ*M* and 55 n*M* for H7 and H89 and K_I values of 22 ± 3 µ*M* and 28 ± 4 n*M*, respectively, which compare well to literature K_I values of 8.3 µ*M* and 48 n*M*.[78,80]

This nanovolume array–based assay has significant potential as a tool for secondary screening or detailed inhibition studies. The four primary advantages to this method are (a) ease of sample manipulation owing to the co-immobilized enzyme and substrate solid-phase assay format; (b) very rapid sample analysis due to high parallelization; (c) significantly reduced reagent volumes; and (d) the ability to perform multiplexed assays which can examine numerous kinases, substrates and

inhibitors at once. The presence of the protein/substrate in the solid phase allows for rapid staining and washing steps that would not be possible in a solution-phase assay or with other immobilization strategies (due to the requirement of sufficient substrate/enzyme mobility for activity); while the reduction in reagent volumes using the array-based assay minimizes both the protein and substrate/inhibitor volumes drastically. The ability to rapidly detect inhibition in single-point assays suggests that the array method may be amenable to high-throughput compound screening, while the accurate determination of IC_{50} values demonstrates the utility of this method for secondary screening of hits found in a primary screen.

CONCLUSIONS AND FUTURE DIRECTIONS

An emerging method for the preparation of protein microarrays is their entrapment within sol-gel-derived microspots that can be pin-printed onto planar surfaces. The use of a sol-gel-based entrapment method for the immobilization of proteins within a microarray has several potential advantages over conventional adsorption, covalent linkage or hydrogel-based methods. Entrapment eliminates the need for protein derivatization, the use of recombinant proteins or affinity capture agents. The three-dimensional nature of sol-gel microspots provides higher protein loading capacity within a biocompatible matrix in addition to allowing the simultaneous co-entrapment of multiple proteins. While sol-gel-based microarrays are likely amenable only to studies of protein-small molecule interactions, their potential for small molecule screening (i.e., drug screening), multianalyte biosensing, and metabolic profiling is exceptional.

The field of sol-gel-derived protein microarrays has produced substantial promise, although this method of microarray fabrication is still relatively undeveloped. The future of this research area can be expanded in numerous directions, which include (a) the improvement of biocompatible silica-based materials; (b) expansion of the number and types of biological targets that can be used for array formation and; (c) scale-up from proof-of-concept to high-throughput, multiplexed analysis of real samples. It is likely that the most important advancements in sol-gel technology will arise due to further bridging of the gap between working in bulk sol-gel materials and working in the nanoscale. Current sol-gel microarraying techniques are based primarily on pin-printing of materials, and thus formulation stability, gelation behavior and biocompatibility remain the largest hurdles to success of this method. Materials need to have adequate working times to ensure ease of pin-printing, while printed spots need to be uniform, crack-free and resistant to overprinting or washing steps. In addition to material optimization, the exploration of novel assay formats can be explored. While Bright's example of layered materials[44] and Clark's MetaChip[60] demonstrate novel approaches that accentuate the advantages of sol-gel entrapment, new studies based on these examples could yield new strategies based on immobilization and assay development in three dimensions.

At this point, numerous proteins have been demonstrated to be compatible with sol-gel entrapment in a microarray format, and in time the targets will undoubtedly become more clinically relevant and/or reveal novel biological information with regard to function or drug inhibition. The successes demonstrated in bulk sol-gel

materials with membrane receptors[58] and whole cells[72] suggest the possibility of using the sol-gel approach to create membrane receptor or cell-based microarrays. In addition to expanding upon the types of biomolecules that can be entrapped, the work to date illustrates the ability to co-immobilize several targets within a given array element. Thus, it is possible to examine metabolism of a given substrate by a cascade of related proteins.

Linked to the issue of materials optimization is the ability to extend the sol-gel-based array format toward large-scale, high-throughput assays. While certain protein families, for instance kinases, may be active within a given formulation, the particular silane precursor or additive used in one instance may not be compatible with other targets. Thus, significant effort may be needed to identify suitable formulations that retain protein activity and are amenable to pin-printing. Given the diversity of available sol-gel precursors and additives, along with the ability to optimize the material directly in an array format, it is likely that suitable materials can be found for almost any biomolecule, thus highlighting the dexterity of the sol-gel method.

REFERENCES

1. Schena, M. et al., Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, 270, 467–470, 1995.
2. Pease, A.C., et al., Light–generated oligonucleotide arrays for rapid DNA sequence analysis, *Proc. Natl. Acad. Sci. U.S.A.*, 91, 5022–5026, 1994.
3. Anderson, L. and Seilhamer, J., A comparison of selected mRNA and protein abundances in human liver, *Electrophoresis*, 18, 533–537, 1997.
4. Griffin, T.J., et al., Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae, *Mol. Cell. Proteomics*, 1, 323–333, 2002.
5. Macbeath, G. and Schreiber, S.L., Printing proteins as microarrays for high–throughput function determination, *Science*, 289, 1760–1763, 2000.
6. Zhu, H., et al., Global analysis of protein activities using proteome chips, *Science*, 293, 2101–2105, 2001.
7. Ptacek, J., et al., Global analysis of protein phosphorylation in yeast, *Nature*, 438, 679–684, 2005.
8. Angenendt, P., et al., Subnanoliter enzymatic assays on microarrays, *Proteomics*, 5, 420–425, 2005.
9. Kramer, A., et al., Identification of barley CK2alpha targets by using the protein microarray technology, *Phytochemistry*, 65, 1777–1784, 2004.
10. Bacarese–Hamilton, T., et al., Protein microarray technology for unraveling the antibody specificity repertoire against microbial proteomes, *Curr. Opin. Mol. Ther.*, 5, 278–284, 2003.
11. Lueking, A., et al., Protein microarrays for gene expression and antibody screening, *Anal. Biochem.*, 270, 103–111, 1999.
12. Glokler, J. and Angenendt, P., Protein and antibody microarray technology, *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, 797, 229–240, 2003.
13. Kersten, B., et al., Large-scale plant proteomics, *Plant Mol. Biol.*, 48, 133–141, 2002.

14. Andrade, J.D., Hlady, V., and Wei A.P., Adsorption of complex proteins at interfaces, *Pure Appl. Chem.*, 64, 1777–1781, 1992.
15. Weetall, H.H., Preparation of immobilized proteins covalently coupled through silane coupling agents to inorganic supports, *Appl. Biochem. Biotechnol.*, 41, 157–188, 1993.
16. Doretti, L., Ferrara, D., and Lora, S., Enzyme-entrapping membranes for biosensors obtained by radiation-induced polymerization, *Biosens. Bioelectron.*, 8, 443–450, 1993.
17. O'Driscoll, K.F., Techniques of enzyme entrapment in gels, *Meth. Enzymol*. 44, 169–183, 1976.
18. Scouten, W.H., A survey of enzyme coupling techniques, *Methods Enzymol.*, 135, 30–65, 1987.
19. Lu, B. et al., Oriented immobilization of antibodies and its applications in immunoassays and immunosensors, *Analyst*, 121, 29R–32R, 1996.
20. Arenkov, P. et al., Protein microchips: Use for immunoassay and enzymatic reactions, *Anal. Biochem*., 278, 123–131, 2000.
21. Rubina, A.Y. et al., Hydrogel-based protein microchips: Manufacturing, properties, and applications, *Biotechniques*, 34, 1008–1014, 1016–1020, 1022, 2003.
22. Afanassiev, V. et al., Preparation of DNA and protein micro arrays on glass slides coated with an agarose film, *Nucleic Acids Res*., 28, E66, 2000.
23. Angenendt, P. et al., Toward optimized antibody microarrays: a comparison of current microarray support materials, *Anal. Biochem.*, 309, 253–260, 2002.
24. Angenendt, P. et al., Next generation of protein microarray support materials: Evaluation for protein and antibody microarray applications, *J. Chromatogr. A*, 1009, 97–104, 2003.
25. Braun, S. et al., Biochemically active sol-gel-glasses: The trapping of enzymes, *Mat. Lett.*, 10, 1–5, 1990.
26. Ellerby, L.M. et al., Encapsulation of proteins in transparent porous silicate glasses prepared by the sol-gel method, *Science*, 255, 1113–1115, 1992.
27. Brennan, J.D., Using Intrinsic Fluorescence to Investigate Proteins Entrapped in Sol-Gel Derived Materials. *Appl. Spectrosc.*, 53, 106A–121A, 1999.
28. Jin, W. and Brennan, J.D., Properties and applications of proteins entrapped in sol-gel derived silica, *Anal. Chim. Acta,* 461, 1–36, 2002.
29. Gill, I., Bio-doped nanocomposite polymers: Sol-gel bioencapsulates, *Chem. Mater.* 13, 3404–3421, 2001.
30. Gill; I. and Ballesteros, A. Encapsulation of biologicals within silicate, siloxane, and hybrid sol-gel polymers: An efficient and generic approach, *J. Am. Chem. Soc.* 120, 8587–8598, 1998.
31. Obert, R. and Dave, B.C., Enzymatic conversion of carbon dioxide to methanol: enhanced methanol production in silica sol-gel matrices, *J. Am. Chem. Soc.*, 121, 12192–12193, 1999.
32. Wu, S. et al., Bacteriorhodopsin encapsulated in transparent sol-gel glass: A new material. *Chem. Mater.*, 5, 115–120. 1993.
33. Weetall, H., et al., Bacteriorhodopsin immobilized in sol–gel glass, *Biochim. Biophys. Acta*, 1142, 211–213, 1993.
34. Besanger, T.R. and Brennan, J.D., Ion sensing and inhibition studies using the transmembrane ion-channel peptide gramicidin A entrapped in sol-gel derived silica, *Anal. Chem.*, 75, 1094–1101, 2003.

35. Dickey, F.H., Specific adsorption, *J. Phys. Chem*., 58, 695–707, 1955.
36. Avnir, D. et al., Recent bio-applications of sol-gel materials, *J. Mater. Chem.*, 16, 1013–1030, 2006.
37. Bhatia, R.B., et al., Aqueous sol-gel process for protein encapsulation, *Chem. Mater.*, 12, 2434–2441, 2000.
38. Brook, M.A. et al., Proteins entrapped in silica monoliths prepared from glyceroxysilanes, *J. Sol-Gel Sci. Technol.*, 31, 343–348, 2004.
39. Cruz-Aguado, J.A. et al., Ultrasensitive ATP detection using firefly luciferase entrapped in sugar-modified sol-gel derived silica, *J. Am. Chem. Soc.*, 126, 6878–6879, 2004.
40. Rupcich, N. et al., Optimization of sol-gel formulations and surface treatment for the development of pin-printed protein microarrays, *Chem. Mater.*, 15, 1803–1811, 2003.
41. Cho, E.J. et al., Tools to rapidly produce and screen biodegradable polymer and sol-gel-derived xerogel formulations, *Appl. Spectrosc*., 56, 1385–1389, 2002.
42. Cho, E.J. and Bright, F.V., Optical sensor array and integrated light source, *Anal. Chem*., 73, 3289–3293, 2001.
43. Cho, E.J. and Bright, F.V., Cho, Eun Jeong; Bright, Frank V. Integrated chemical sensor array platform based on a light emitting diode, xerogel-derived sensor elements, and high-speed pin printing, *Anal. Chim. Acta*, 470, 101–110, 2002.
44. Cho, E.J. and Bright, F.V., Pin-printed biosensor arrays for simultaneous detection of glucose and O_2. *Anal. Chem*., 74, 6177–7184, 2002.
45. Park, C.B. and Clark, D.S., Sol-gel encapsulated enzyme arrays for high throughput screening of biocatalytic activity, *Biotechnol. Bioeng.*, 78, 229–235, 2002.
46. Rupcich, N. and Brennan, J.D., Coupled enzyme reaction microarrays based on pin-printing of sol-gel biomaterials, *Anal. Chim. Acta*, 500, 3–12, 2003.
47. Williams, A.K. and Hupp, J.T., Sol-gel-encapsulated alcohol dehydrogenase as a versatile, environmentally stabilized sensor for alcohols and aldehydes. *J. Am. Chem. Soc.*, 120, 4366–4371, 1998.
48. Badjic, J.D. and Kostic, N.M., Effects of encapsulation in sol-gel silica glass on esterase activity, conformational stability, and unfolding of bovine carbonic anhydrase II, *Chem. Mater.*, 11, 3671–3679, 1999.
49. Yamanaka, S.A. et al., Nicotinamide adenine dinucleotide phosphate fluorescence and absorption monitoring of enzymic activity in silicate sol-gels for chemical sensing applications, *J. Am. Chem. Soc.*, 117, 9095–9096, 1995.
50. Yamanaka, S.A. et al., Enzymic activity of oxalate oxidase and kinetic measurements by optical methods in transparent sol-gel monoliths. *J. Sol-Gel Sci. Technol.*, 7, 117–121, 1996.
51. Besanger, T.R. et al., Screening of inhibitors using enzymes entrapped in sol-gel derived materials, *Anal. Chem.*, 75, 2382–2391, 2003.
52. Zheng, L. et al., Measurement of fluorescence from tryptophan to probe the environment and reaction kinetics within protein-doped sol-gel-derived glass monoliths, *Anal. Chem.*, 69, 3940–3949, 1997.
53. Tsai, H.-C. and Doong, R.-A., Simultaneous determination of renal clinical analytes in serum using hydrolase- and oxidase-encapsulated optical array biosensors, *Anal. Biochem.*, 334, 183–192, 2004.

54. Gulcev, M.D. et al., Reagentless pH-based biosensing using a fluorescently-labeled dextran co-entrapped with a hydrolytic enzyme in sol-gel derived nanocomposite films, *Anal. Chim. Acta*, 457, 47–59, 2002.
55. Tsai, H.-C. and Doong, R.-A., Simultaneous determination of pH, urea, acetylcholine and heavy metals using array-based enzymatic optical biosensor, *Biosens. Bioelectron.*, 20, 1796–1804, 2005.
56. Besanger, T.R. and Brennan, J.D., Entrapment of membrane proteins in sol-gel derived silica, *J. Sol-Gel Sci. Technol*., 40, 209–225, 2006.
57. Luo, T.-J.M. et al., Photo-induced proton gradients and ATP biosynthesis produced by vesicles encapsulated in a silica matrix. *Nature Mater.*, 4, 220–224, 2005.
58. Besanger, T.R. et al., Entrapment of highly active membrane-bound receptors in macroporous sol-gel derived materials, *Anal. Chem.*, 76, 6470–6475, 2004.
59. O'Neill, H. and Greenbaum, E., Spectroscopy and photochemistry of spinach photosystem I entrapped and stabilized in a hybrid organosilicate glass. *Chem. Mater.*, 17, 2654–2661, 2005.
60. Lee, M.-Y. et al., Metabolizing enzyme toxicology assay chip (MetaChip) for high-throughput microscale toxicity analyses. *Proc. Natl. Acad. Sci. U.S.A.*, 102, 983–987, 2005.
61. Fang, Y. et al., Membrane protein microarrays, *J. Am. Chem. Soc.*, 124, 2394–2395, 2002.
62. Hong, Y. et al., Functional GPCR Microarrays, *J. Am. Chem. Soc*., 127, 15350–15351, 2005.
63. Hong, Y. et al., G-protein-coupled receptor mircoarrays for multiplexed compound screening, J. *Biomol. Screen.*, 11, 435–438, 2006.
64. Manning, G. et al., The protein kinase complement of the human genome, *Science*, 298, 1912–1934, 2002.
65. Kostich, M. et al., Human members of the eukaryotic protein kinase family, *Genome Biol.*, 3, 1–12, 2002.
66. Dancey, J. and Sausville, E.A., Issues and progress with protein kinase inhibitors for cancer treatment, *Nat. Rev. Drug Discov.*, 2, 296–313, 2003.
67. Adams, J.L. et al., p38 MAP kinase: molecular target for the inhibition of pro-inflammatory cytokines, *Prog. Med. Chem.*, 38, 1–60, 2001.
68. Yousif, M.H. et al., The role of tyrosine kinase–mediated pathways in diabetes–induced alterations in responsiveness of rat carotid artery, *Auton. Autacoid Pharmacol.*, 25, 69–78, 2005.
69. Manning, B.D. and Cantley, L.C., Hitting the target: Emerging technologies in the search for kinase substrates, *Sci. STKE*, 2002, PE49, 2002.
70. Zhu, H. et al., Analysis of yeast protein kinases using protein chips, *Nat. Genet.*, 26, 283–289, 2000.
71. Reimer, U. et al., Peptide arrays: from macro to micro, *Curr. Opin. Biotechnol.*, 13, 315–320, 2002.
72. Houseman, B.T. et al., Peptide chips for the quantitative evaluation of protein kinase activity, *Nat. Biotechnol.*, 20, 270–274, 2002.
73. Schutkowski, M. et al., Automated synthesis: high-content peptide microarrays for deciphering kinase specificity and biology, *Angew. Chem. Int. Ed. Engl.*, 43, 2671–2674, 2004.

74. Min, D.H. et al., Profiling kinase activities by using a peptide chip and mass spectrometry, *Angew Chem. Int. Ed. Engl.*, 43, 5973–5977, 2004.
75. Williams, R.A. and Blanch, H.W., Covalent immobilization of protein monolayers for biosensor applications, *Biosens. Bioelectron.*, 9, 159–167, 1994.
76. Vijayendran, R.A. and Leckband, D.E., A quantitative assessment of heterogeneity for surface–immobilized proteins, *Anal. Chem.*, 73, 471–480, 2001.
77. Rupcich, N. et al., Nanovolume kinase inhibition assay using a sol-gel derived multicomponent microarray, *Anal. Chem.*, 77, 8013–8019, 2005.
78. Martin, K. et al., Quantitative analysis of protein phosphorylation status and protein kinase activity on microarrays using a novel fluorescent phosphorylation sensor dye, *Proteomics*, 3, 1244–1255, 2003.
79. Tamanini, A. et al., Adenosine 3′:5′–monophosphate–dependent protein kinase from human placenta: Characterization of the catalytic subunit, *Enzyme*, 45, 97–108, 1991.
80. Lin, Q. et al., Effects of protein kinase A activation on the responses of primate spinothalamic tract neurons to mechanical stimuli, *J. Neurophysiol.*, 88, 214, 2002.
81. Livage, J. et al., Encapsulation of biomolecules in silica gels, J. Phys.: Cond. Matter, 13, R673–R691, 2001.

6 Printing and QC of Functional Protein Microarrays

Fang X. Zhou, Dee Shen, and Barry Schweitzer

CONTENTS

INTRODUCTION

The human genome project has catalyzed the development of new large-scale approaches to addressing biological questions. Over the last decade, for example, the use of DNA microarrays has become a routine approach for simultaneously analyzing the expression of thousands of genes. Functional protein arrays (microarrays with immobilized functional proteins) are an extension of DNA microarrays, although the manufacture of protein microarrays presents additional challenges in areas such as content generation, printing, functional immobilization, and detection. The uses of protein microarrays, which are reviewed elsewhere,[1,2] cover an impressive range of applications, from probing molecular interactions for protein functional characterization to optimization of drug–protein interactions, and from profiling of enzyme substrates to profiling immune response in various diseases.

Functional protein microarrays clearly have the potential to make significant contributions to both basic and applied research. This chapter reviews the current state of microarray printing technology as well as a discussion of the quality control that is required to assure a product that can be used to develop meaningful insights and discovery in biology.

MICROARRAY PRINTING TECHNOLOGIES

Manufacturing a microarray by printing involves delivering a small volume of (typically) many samples onto a solid surface in a reproducible and spatially addressable fashion. The volume of dispensed liquid is typically in the nanoliter to picoliter range. Two commercially available technologies that have been utilized for printing protein microarrays are noncontact ink-jet printing and contact pin-transfer. Given the requirement to array large numbers of different proteins, contact printing is currently the most suitable choice for the manufacture of functional protein arrays, although noncontact printing of certain types of these arrays is certainly possible. Other recently described approaches for protein microarray manufacture includes a laser transfer technique,[3] microfabricated fountain pens for high-density array construction,[4] as well as a novel affinity contact printing procedure employing a multiuse stamp.[5] Cooks' group at Purdue University recently described an exciting proof-of-concept using electrospray ionization of a protein mixture followed by mass ion separation and sequential soft landing deposition onto a surface to create a protein array.[6] While promising, these technologies face many challenges before they can be commercialized, including improving print speed as well as addressing protein quantity, identity and functionality.

Noncontact ink-jet printing is derived from the ink-jet printing industries. The fundamental principle of this technology involves the application of force to create a rapidly move liquid stream, which then passes through a small orifice. When samples pass through the orifice, the stream achieves sufficient velocity to overcome surface tension and a droplet is ejected from the print head onto the surface. The most widely used ink-jet technology for printing microarrays is piezoelectric. Some of the commercially available piezoelectric ink-jet microarraying instruments include those from Perkin Elmer (Wellesley, MA, USA), GeSim (Germany), and MicroFab (Plano, Texas, USA). Typical piezoelectric dispensers can create drops in the picoliter range and with coefficient of variations (CV) of 3~7%.[7] However, the main difficulties in implementing this technology include intermittent dispensing caused by gas bubbles and tip clogging due to the small size of the orifice and its dependence on surface tension. Because of these and other engineering limitations, the number of samples that these instruments can dispense in a reasonable amount of time is relatively low; consequently, the use of these instruments for manufacture of protein microarrays has been generally limited to products containing <100 proteins, typically antibodies.

Contact pin printing technology involves using a rigid pin to transfer liquid from a source plate of samples to a precise destination on a solid surface. Dipping the pins into the samples results in a small volume of liquid either on the tips of the pins or drawn up into a reservoir within the pins. The pins are then tapped

onto the slide surface to deposit the liquid. The typical printing volume is in the high picoliter to low nanoliter range. Pin printing was initially carried out using solid pins (V&P Scientific, San Diego, CA, USA), and later other pin variations (split or quill) were developed to permit the printing of multiple arrays with a single loading (Harvard BioScience, Holliston, MA, USA; ArrayIt, Sunnyvale, CA, USA; Incyte/Stanford, Palo Alto, CA, USA). Pin-based printing has the advantage of being relatively simple and inexpensive. However, pin-to-pin variation can be higher (CV of 10~25%) than ink-jet dispensing due to variations in pin geometry and surface chemistry.[8] In addition, pins can also clog or can be deformed or wear over time.

The power of microarraying technology has been demonstrated primarily in manufacturing DNA microarrays.[9,10] Unlike DNA, however, proteins must maintain a chemically fragile three-dimensional structure in order to preserve functionality. The production of protein microarrays requires careful consideration of the printing environment to maintain the quality and functionality of the proteins on the arrays. As mentioned above, a major consideration when choosing the type of printing technology to use in printing protein microarrays is throughput. In general, a noncontact printing method is best when the number of samples is small and the number of replicates is high. Commercially available noncontact printers have fewer dispensers (typically 1~4), and they are typically designed to load a large volume to dispense thousands of replicates with a single aspiration. To produce arrays consisting of thousands of different proteins, a contact printing system equipped with 48~64 pins is significantly faster and thus more adept at printing larger sets of samples. Even with this large number of pins, the time to print 100 slides containing a few replicates of thousands of proteins can last about 10~15 hours. Consequently, the contact printing system has to be located in an environment that is temperature and humidity controlled in order to protect proteins in source plates and on arrays while they are being printed.

CHALLENGES IN MANUFACTURING HIGH-QUALITY PROTEIN MICROARRAYS

Variables in Protein Microarrays Manufacturing

Protein microarrays have moved from simple forms that were designed to show proof-of-concept to commercial products that contain thousands of human proteins. There are still considerable challenges in manufacturing high-quality protein microarrays that suit the needs of the various applications for which they are used. The quality of functional protein microarrays can be viewed by their content (the number, diversity, annotation, and activity of proteins) and performance (the minimal detection limit, dynamic range, and reproducibility of the assay). While high-content microarrays have been produced in some academic laboratories, the performance of these arrays is limited by lack of a robust, controlled manufacturing process. In contrast, most of the high-performance protein microarrays that are currently available commercially include only antibodies or a few hundred proteins, limiting their applications.

As the demand for more protein content on arrays grows, high costs of manufacturing could become a barrier to commercialization and customer adoption. To achieve high quality at reduced costs, three interdependent processes have to be optimized, including protein content generation, surface chemistry development or selection, and microarray printing. In this chapter, we will discuss printing commercial-quality high-density functional protein microarrays and the quality control procedures required to achieve optimal application performance.

Assuming that the source of protein and the surface chemistry used for manufacturing microarrays are fixed, the primary factor that determines the performance of a protein microarray is the amount of protein delivered to the surface. The goal is to produce identical spots on each microarray in a batch as well as between batches. This requires the process to control the dispensing volume (in subnanoliters) and the size of the protein spots (micrometers in diameter) on the substrate within very tight specifications. The process involves parallel deposition of diverse proteins, highly complicated mass transport phenomena, and surface chemistry. In addition, interactions between various protein solutions, surfaces, and environmental factors make the dispensing volume and spot size difficult to predict and control. Therefore, dedicated resources as well as expertise in protein chemistry, surface chemistry, and engineering are required to develop a reliable manufacturing and quality control process. We will focus our discussion on four major factors (protein solutions, microarray slides, pins, and environment) that have to be closely controlled in contact printing. Other methods or variables, such as microarray design, robotic capability, and human factors, also affect the quality in manufacturing but will not be discussed here.

Protein Solutions

Protein solutions are one of most complicated and least discussed quality variables in functional protein microarray manufacturing. In a high-throughput protein production process, purified proteins are stored frozen in microtiter plates and used later to create microarrays. Specific buffers are required to purify proteins to achieve optimal recovery and protein activity. Certain components in the buffers are needed to stabilize protein structure and to maintain their function during storage. However, the same buffers have properties such as surface tension and viscosity that influence the printing performance during production. It is technically difficult or at least noneconomical to exchange buffers of thousands of proteins in order to obtain ideal spot intensity and/or morphology on a microarray. Consequently, the effects of each buffer component on the entire manufacturing process have to be considered during the selection and/or optimization of the buffer; compromises often have to be made when there is a conflict. For example, 10 to 50% of glycerol is included in many protein buffers to protect proteins during storage. Glycerol not only affects the hydration of proteins but also changes the viscosity, surface tension, and hydrophobicity of the buffers, which in turn have an effect on manufacturing results. Another example is nonionic detergents that are often required for protein solubility and structural stability. Although these detergents are typically used at very low (<0.1%) concentrations, their effects on

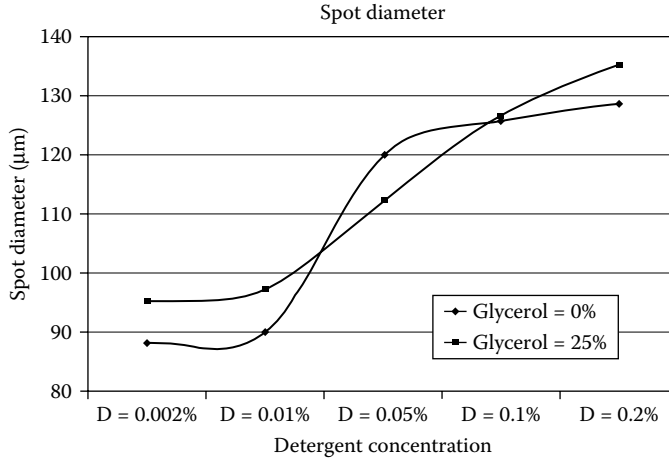


FIGURE 6.1 Buffers containing fluorescent protein and varying amounts of glycerol and detergent were printed with the same set of pins on the same slide. The spot diameter is measured directly using fluorescent scanning and automated spot finding software.

spotting performance can be complex. In Figure 6.1, a protein sample in different buffers was printed using a contact printing method on the same substrate. Changing the detergent concentration from 0.01 to 0.2% can change the spot diameter by 40 microns (or >40%). Furthermore, such changes are nonlinear and affected by the presence of glycerol.

Microarray Slides

Surface chemistry development for protein microarrays is discussed by other authors in this book. While much technological development on microarray substrates has been centered on immobilization and functional activity in applications, less has been done to address quality and consistency in the manufacturing of high-density protein microarrays. Because the amount of liquid delivered to the substrate surface depends on the interaction between the protein solution and surface chemistry, the choice of a surface chemistry has to satisfy both application and printing needs. In addition, any variation in chemical composition or physical structure of the surface can cause defective microarrays with varying spot size, morphology, and protein function. In fact, we have observed batch-to-batch, slide-to-slide, as well as intra-slide surface variations on virtually all commercial slides that we have tested.

One of the surface properties that affect production of high-density protein microarrays is wettability, which can be measured by various contact angle instruments. The contact angle of water on the slide correlates with surface hydrophobicity. In general, a lower contact-angle surface produces larger spots given the same drop of water. Protein solutions are much more complex than water, and therefore it is

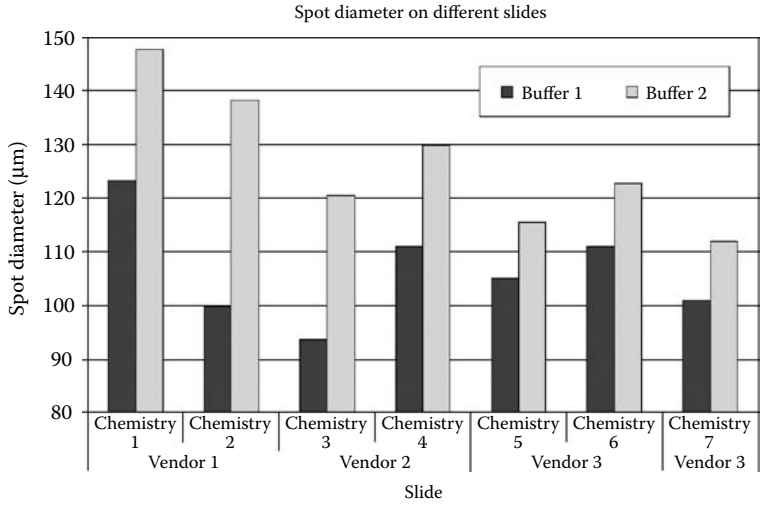


FIGURE 6.2 Two protein buffers containing fluorescent protein were printed on a number of microarray slides from 4 vendors. The spot diameter varies among slides, and the variation depends on the buffer.

not surprising that their behaviors do not always correlate with surface chemistry. As shown in Figure 6.2, two buffers exhibit different spot diameters when printed on a number of different surfaces. Despite its limited power for predicting spot size, contact angle measurements can still be very useful for detecting gross variations in surface quality.

Contact angle methods measure an area of several millimeters in diameter. Microscopic variations are much more difficult to measure routinely and can affect protein immobilization, conformational stability, and functional availability of protein domains. More sophisticated methods, such as atomic force microscopy (Figure 6.3), can measure variations in nanometers but are less suitable for application in manufacturing. Atomic composition can be measured on a surface by X-ray Photoelectron Spectroscopy (XPS). Differences in composition and coating thickness between lots of slides of the same chemistry can be significant (see example in Table 6.1). Furthermore, the application performance of microarrays made on these lots is considerably different (Figure 6.4). As discussed later in the chapter, protein microarrays should always be tested in functional applications to ensure quality.

Printing Pins

Spot size, uptake volume, content carryover, durability, and consistency are some of the characteristics of pins that need to be taken into account when employing contact printing for making protein microarrays. On high-content microarrays (>15,000 features/slide), a spot diameter of 150 microns or smaller is desired. Because the spot size is highly dependent on the protein solution and slide chemistry, testing with

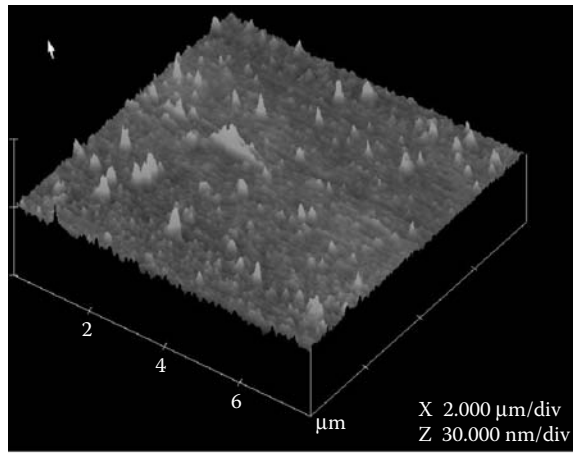


FIGURE 6.3 The surface topology of a slide was analyzed with Atomic Force Microscopy. The surface roughness is clear in the image.

production materials is necessary. The amount of solution deposited on a surface also depends on the volume taken up by the pin. Several manufacturers provide pins that are suitable for protein microarrays. For example, TeleChem International (Sunnyvale, CA) has three series of Stealth pins that produce a range of spot sizes (62.5 to 600 µm diameter) and sample uptake volumes (0.5 to 12.5 µl). A high-density protein microarray typically has spots of about 0.5 nl or less and requires sub-microliters of uptake volume in a pin. Since there is always some sample left in the pin after printing each set of proteins, pins have to be washed extensively between samples to avoid carryover of contents. The protein carryover property of the pins has to be examined with real protein samples to establish acceptance criteria of pins and to develop adequate pin wash protocols (see quality control example later in the chapter).

One of the goals in making high-content microarrays is to produce thousands of consistent spots, which are made by separate pins. Not all pins are identical, however,

TABLE 6.1
The Atomic Composition of Two Lots of the Same Slide Chemistry Was Measured by X-ray Photoelectron Spectroscopy (XPS), Which Shows Surface Variation

	C1s	N1s	O1s	Na1s
Lot 1	51.0	2.6	37.1	2.3
Lot 2	42.7	0.5	42.6	2.5

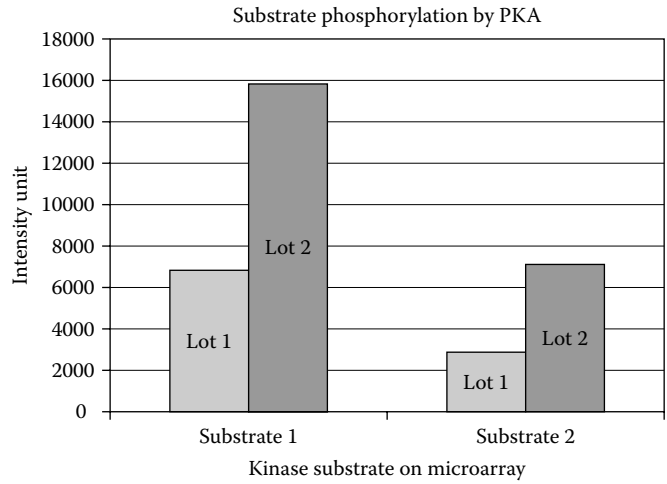


FIGURE 6.4 Two lots of the same slide chemistry were tested for substrate phosphorylation performance. Identical substrates were printed on the slides and assayed with PKA at the same time.

and they wear out or age at different rates during production. Common issues that arise with used pins are deformation of the printing end and changes in surface properties, such as surface energy or roughness. Either of these defects can result in missing, irregular-sized, or irregular-shaped spots and loss in feature signals due to insufficient volume delivered. Some pin defects are obvious under a microscope, and some are not. Occasionally, brand new pins may not perform to the required specifications, and must therefore be conditioned or broken-in before use in production. More on quality control of printing pins is discussed later in the chapter.

Environment

Protein microarray facilities that lack sufficient environmental control on temperature, humidity, and air quality may produce inconsistent products and also compromise protein integrity and function. While proteins are normally stored in freezers immediately after purification, slides are not always protected from the environment. Because many microarray substrates include active functional groups, storage of slides in an environmentally-controlled location is recommended to reduce uncertainty in quality. In one experiment, slides from the same lot were stored in different conditions for 4 days and then printed with fluorescently labeled proteins. As shown in Figure 6.5, protein retention on the surface varied with the buffer as well as with the storage conditions. In addition to temperature and humidity, pollution (e.g., ozone) in the atmosphere may also cause deterioration in slide performance.

As in most manufacturing facilities, precise control of temperature, humidity, and air particles in the printing room is a must if consistent quality is desired. A cold

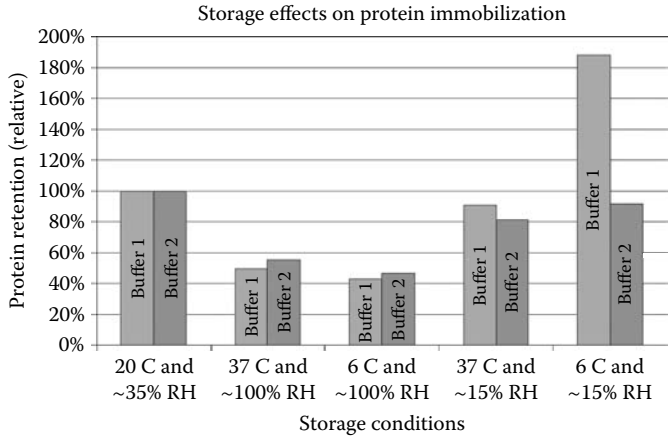


FIGURE 6.5 Slides were stored under different environments for 3 days and printed with the same fluorescent protein at the same time. The fluorescent signals from each slide were measured before and after they were washed. Signals are shown relative to the one stored in the lab (20°C and 35% RH).

room (4 to 8°C) is essential to protect proteins both in the source plates and on the slides during the batch printing process, which lasts hours for high-content microarrays. Because surface tension and viscosity of solutions are temperature-dependent, it is important to develop all printing methods at the same temperature used in production and to maintain the temperature throughout each run. The humidity in the printing room or chamber has to be optimized for each solution/slide combination because wettability of a slide depends on the moisture in the air and on the slide surface. A proper humidity level also helps to prevent excess evaporation of samples during the time of printing. Air particles are often a major problem if the printing room is not clean; misshaped, merged, and missing spots will occur with increased frequency if there are dust particles on the slide surface or in the printing pin or nozzle. Figure 6.6 shows a magnified image of a pin catching a small piece of fiber, which results in noncircular spots on a microarray. Ideally, microarray production is carried out in a clean-room environment where air is HEPA-filtered and surfaces are regularly cleaned.

QUALITY CONTROL IN PROTEIN MICROARRAY MANUFACTURING

Preprinting Quality Control

Because variations in buffers, printing pins, and slides can lead to defective microarrays, it is imperative to perform preprinting quality control in order to detect and reduce such variations in materials. Careful research is required to relate a material's

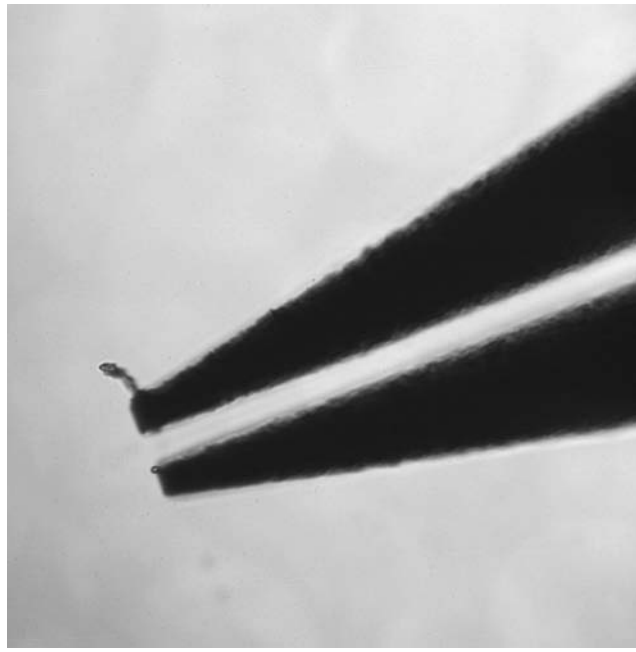


FIGURE 6.6 A typical pin for printing microarray is shown catching a piece of fiber, which can cause irregular and inconsistent spot morphology.

measurable physical and/or chemical properties to relevant microarray quality parameters, such as spot diameter. Once the relationship is found, a quality control step is added to ensure the material meets specific acceptance criteria before it is used in production. The acceptance criteria are determined based on 1) the sensitivity of the quality parameter to the measurable properties and 2) product quality tolerance (or allowed variation in the quality parameter). For example, the concentration of buffer components can be measured by light absorbance, densitometry or enzymatic assays, and the surface quality of microarray slides can be evaluated by contact angle measurements or fluorescent scanning.

Some material variables are unknown, difficult to measure, and may not affect quality parameters independently. For instance, precise physical dimensions and surface properties of printing pins play critical roles in spot quality on microarrays but are extremely difficult to measure directly. In such cases, preprinting tests are required to detect nonconformances and to perform corrective actions prior to production. Before each production printing in our facility, the arrayer and supporting components are tested to ensure that production specifications are met. Quantitative pin QC criteria were developed to assess reproducibility of spot volume, spot size and morphology, and sample carryover. Pins not meeting QC criteria are cleaned, or replaced and retested. One test that we typically run is to measure the decay of signal for each pin while printing 256 replicates after a single loading (Figure 6.7).

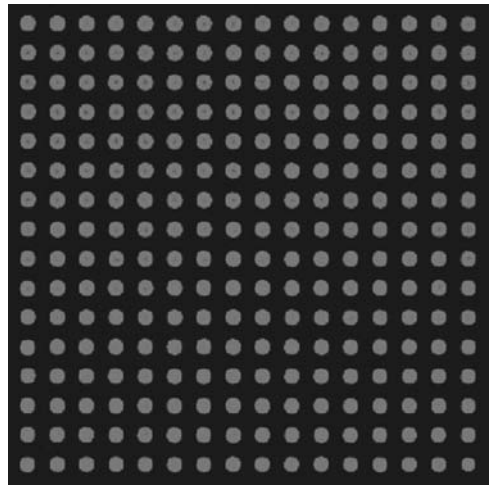


FIGURE 6.7 A subarray image of the same fluorescent sample printed by a single pin is shown to show the consistency in spot intensity and size. All pins are tested and compared to ensure consistency before production runs.

This test ensures that the arrayer can dispense duplicates per sample on 100 slides required in our product layout (shown in the production section) without redipping. The individual pin CV is usually less than 10% and spot size is approximately 130 μm. While comparing droplets of sample dispensed by different pins, the variability is somewhat greater, sometimes around 15%.

Another quality assurance step that is taken prior to initiating a production run is to test the cleaning/washing components of the system. After each sample is dispensed on the arrays, the pins are cleaned by dipping into distilled water, and then into a washing solution. Cleaning is also facilitated by activation of an ultrasonic transducer in the cleaning bath. Finally, residual washing buffer is removed from the pins by a vacuum. This procedure is repeated several times to minimize sample carryover. The efficacy of the procedure is tested by first dispensing Alexa Fluor® 647/Alexa Fluor® 555-labeled antibody in a series of spots, and then dispensing buffer alone in second series of spots. Analysis of the fluorescent signals should indicate that carryover is less than 1 part in 5000 if the system is performing adequately.

Production

As mentioned above, the production of protein microarrays at Invitrogen is done in a cold room. The humidity level of the printing environment is maintained by the dehumidification unit of the cold room and a humidifier controlled by the arrayer. The room is also equipped with HEPA filters and is regularly checked for particle counts to ensure a clean room environment. In addition, the production area is limited only to arrayer operators, who follow clean room operational procedures.

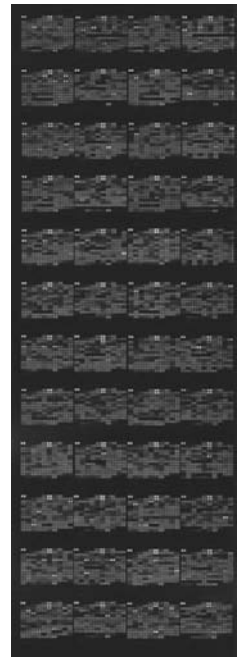
After the arrayer has been calibrated and tested, and bar-coded protein source plates and slides are loaded onto the arrayer, the production process is completely automated. Figure 6.8 shows an example of the printing layout of one of our products, ProtoArray™ Human Protein Array v3.0, a protein array containing approximately 5000 different human proteins. This array is designed to accommodate 19,200 spots with a 220 μm pitch printed in 48, 20 × 20 spot subarrays (4400 μm^2 each). A 100 μm gap exists between adjacent subarrays. Each subarray contains a number of controls (e.g., a gradient of BSA as a negative control) and calibration spots (i.e., a gradient of GST is used for generating a standard curve for post-printing QC as discussed below). All proteins are printed in duplicate.

Post-printing Quality Control

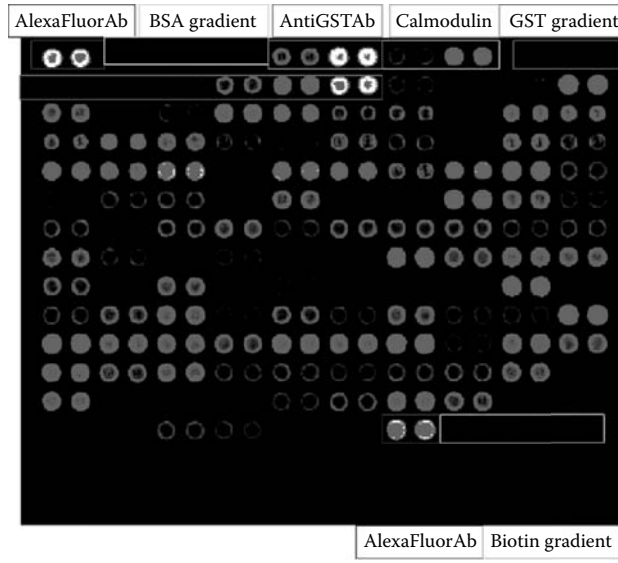
Every effort should be made before and during the production of microarrays to ensure the best quality. Post-printing quality control is necessary before protein microarrays can be provided to customers. Two types of quality measurements are routinely carried out at Invitrogen. One measurement is the consistency of proteins printed within and between arrays, and the other is a test of their functional performance in specific applications.

After printing, all arrays are visually inspected for scratches, fibers, and other obvious defects. The second step of the post-printing QC process consists of a more detailed analysis of each spot on the array. In our protein manufacturing process, each protein is tagged with an epitope (e.g., GST); consequently, QC can be accomplished by using a labeled antibody that is directed against this epitope. A typical fluorescent image obtained with this QC step is shown in Figure 6.8. This procedure measures the variability in spot intensity and morphology, the number of missing spots, and the presence of controls. Another objective of this QC process is to determine how much material is deposited on each spot. Every array, therefore, is printed with a dilution series of known quantities of a protein containing the epitope tag (e.g., purified GST) that is used to generate a standard curve. This procedure enables the signal intensities for each spot to be converted into the amount of protein deposited. Data acquired from two arrays from the beginning, middle and the end of a printing run are also used to determine the reproducibility of the manufacturing procedure.

The final step of the quality control process is to ensure that the products will perform as needed for specific applications. One common application of these arrays is to use them to probe for protein–protein interactions. In this application, customers use a recommended procedure to probe the array with their protein and then detect interactions using streptavidin or antibodies labeled with a fluor, preferably Alexa Fluor 647. Consequently, arrays from each print lot are probed with a protein, calmodulin kinase, that is biotinylated and that also contains a V5 epitope tag, and detection is carried out with Alexa Fluor 647–labeled streptavidin or Alexa Fluor 647–labeled anti-V5 antibody. Appropriate interactions with control elements in each subarray such as an anti-biotin antibody and calmodulin must be observed before the lot is released as product. A representative image of these interactions is shown in Figure 6.9.

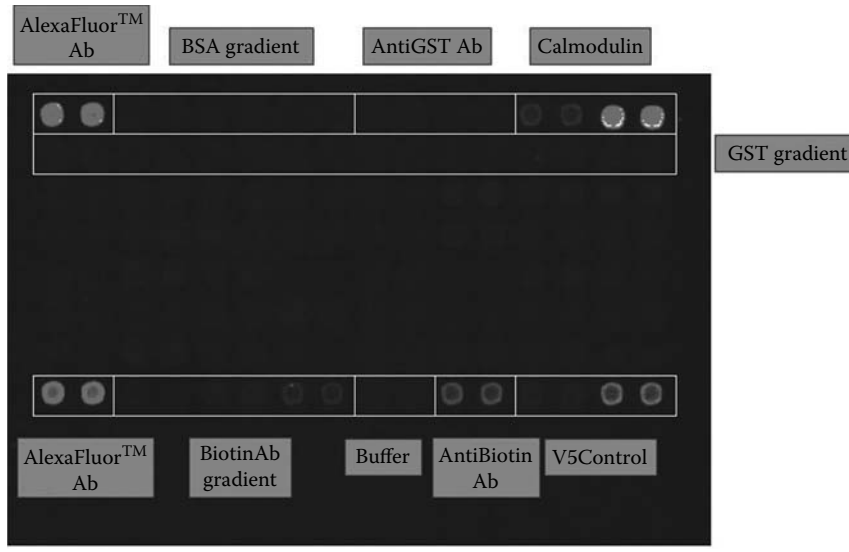


(a)

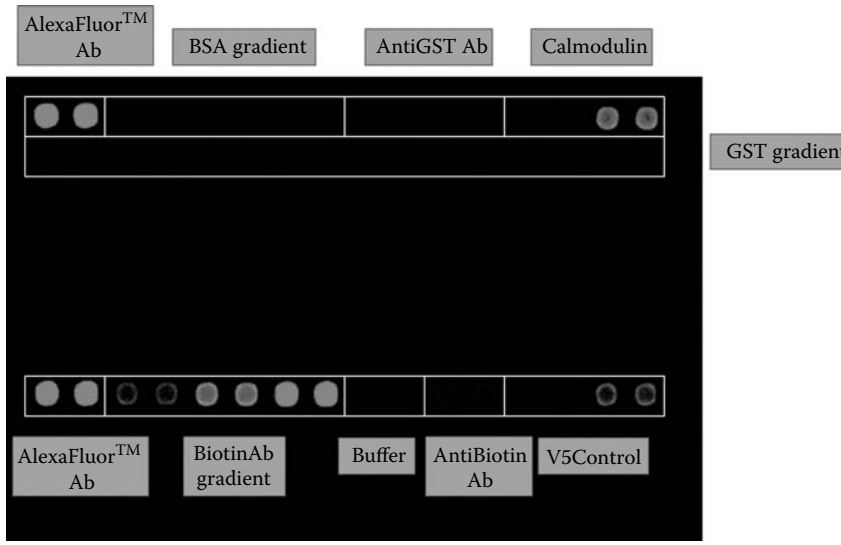


(b)

FIGURE 6.8 The protein array was probed with an anti-GST antibody followed by an AlexFluor 647 labeled secondary antibody. 8A is an image of the entire array and 8B is one of the 48 subarrays. Control proteins are included in every subarray and shown in the boxes. See color insert following page 236.



(a)



(b)

FIGURE 6.9 Protein arrays from every batch are tested for functionality. Calmodulin kinase was used as a probe to detect its interaction with Calmodulin printed on each array. (a) is a subarray image of the probed array detected with Alex Fluor 647 labeled anti-V5 antibody, and (b) detected with Alex Fluor 647 labeled streptavidin. See color insert following page 236.

CONCLUSIONS

An increasing number of researchers are benefiting from the commercial availability of high-content protein microarrays. The manufacturing process and quality control are some of the major challenges in delivering affordable, high-quality functional protein microarrays. As shown in this chapter, significant progress has been made in controlling various manufacturing factors that have improved the consistency and functionality of these innovative products. The future will likely see further advancement in content generation, surface chemistry, and microarray manufacturing technologies.

REFERENCES

1. Schweitzer, B., Predki, P., and Snyder, M., Microarrays to characterize protein interaction on a whole-proteome scale, *Proteomics*, 3, 2190, 2003.
2. Zhou, F.X., Bonin, J., and Predki, P.F., Development of functional protein microarrays for drug discovery: Progress and challenges, *Comb. Chem. High Throughput Screen*, 6, 539, 2004.
3. Delehanty, J.B. and Ligler, F.S., Method for printing functional protein microarrays, *BioTechniques*, 34, 380, 2004.
4. Ringeisen, B.R. et al., Picoliter-scale protein microarrays by laser direct write, *Biotechnol. Prog.*,18, 1126, 2002.
5. Reese M.O., Van et al., Microfabricated fountain pens for high-density DNA arrays, *Genome Res.*,13, 2348, 2003.
6. Ouyang, Z. et al., Preparing protein microarrays by soft-landing of mass-selected ions, *Science*, 301, 1351, 2003.
7. Englert, D., Production of microarrays on porous substrates using noncontact piezoelectric dispensing, in *Microarray BioChip Technology*, Schena, Ed., Natick, MA: Eaton Publishing Company, pp. 231–246, 2000.
8. Zhang, W., Shmulevich, I., and Astola, J., *Microarray Quality Control*, Hoboken, NJ: John Wiley & Sons, 2004.
9. Schena, M. et al., Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, 270, 467, 1995.
10. Heller, R.A. et al., Discovery and analysis of inflammatory disease-related genes using cDNA microarrays, *Proc. Natl. Acad. Sci. USA*, 94, 2150, 1997.

7 Protein Engineering for Surface Attachment

Aparna Girish, Grace Y. J. Chen, and Shao Q. Yao

CONTENTS

INTRODUCTION

Protein microarrays have emerged as a powerful tool in the high-throughput analysis of protein abundance and function.[1–3] One of the key concerns in the fabrication of functional protein microarrays is the method of immobilization, which to a large extent determines whether or not an immobilized protein retains its native biological

function.[4] Currently, there are two categories of immobilization methods used in a protein microarray, either random methods of immobilization or methods that allow site-specific orientation of proteins.[5] This chapter give a summary of the different methods that have been employed to site specifically label and attach proteins onto a glass slide to generate the corresponding protein microarray. Some of the methods will be elaborated in details.

STRATEGIES FOR IMMOBILIZATION

The different methods that have been employed to engineer modified proteins can be grouped into those that introduce (a) peptide/polypeptide affinity tags (b) small molecule tags, and (c) genetic tags, all of which have been used successfully to immobilize proteins onto modified glass surfaces, giving rise to the corresponding protein microarray.

Peptide/Polypeptide-Based Immobilization

Recombinant proteins that are fused to a desired peptide/protein affinity tag at either terminus of proteins can be easily produced in standard molecular biology labs. With the corresponding affinity partner coated onto slides, protein immobilization can be easily achieved. In essence, this is similar to the bead-based affinity chromatography of proteins using affinity tags, e.g., Glutathione-S-transferase (GST) or polyHistidine tags.

As one of the earliest protein microarrays generated by site-specific immobilization methods, Snyder's group developed the so-called "yeast proteome array" by making use of the specific interaction between poly Histidine tags, expressed at the end of recombinant proteins, and Ni-NTA ligands.[6] The authors were able to show that signal intensities from $(His)_6$-tagged proteins spotted onto Ni-NTA slides were 10 times higher than those spotted onto aldehyde slides. They also showed that the majority of the immobilized proteins retained their biological activity. However, the binding between Ni-NTA and $(His)_6$-tagged proteins is not very stable, often susceptible to interference by many commonly used chemicals and salts,[7] making this immobilization method incompatible with a variety of protein-screening assays. The advantage of this method over other methods, however, is that the affinity tag is a small peptide, thus possibly causing minimal effect to the target protein.

In a different approach, Mrksich and coworkers captured cutinase-fused proteins onto self-assembled monolayers (SAMs) of alkanethiolates coated on a gold surface. By using active site-directed phosphonate ligands, the authors were able to achieve site-specific and covalent immobilization of the cutinase fusion proteins.[8] Cutinase is a 22 kDa serine esterase that forms a site-specific covalent adduct with phosphonate ligands. It was shown through SPR that calmodulin–cutinase fusion was captured successfully and irreversibly, and that calcineurin–calmodulin interactions could occur favorably and specifically on the surface. However, the phosphonate ligand might have cross reactivity towards other proteases, esterases and lipases from the crude cell lysate.[9] Kindermann et al. successfully developed a site-specific method to covalently immobilize hAGT-fused proteins onto modified glass surfaces.[8] A hAGT mutant that specifically catalyses the transfer of O_6-benzylguanine

to its own cysteine residue could be fused to either terminus of desired proteins. The authors showed the immobilization of GST-hAGT fusions were possible on O_6-benzylguanine-coated carboxymethylated dextran chips. However, it was noted that the fusion tags on both the above strategies are bulky, and thus may affect the biological property of the fused proteins.

Camarero et al. described the use of Expressed Protein Ligation (EPL) to generate functionally active proteins possessing a C-terminal thioester handle, and subsequently immobilized them onto a cysteine-modified glass slide, generating the corresponding protein array.[9] Choi et al. devised an alternative strategy using DNA surfaces by exploiting the GAL4 DNA binding domain to generate fusion proteins for immobilization onto slides coated with the target dsDNA sequence (that binds with the GAL4 domain selectively, with a low dissociation constant in the nanomolar range).[10] In a recent development, Tirell et al. made use of leucine zipper domains to immobilize proteins onto microarrays.[11] They fused the ZE domain (43 amino acids) to the desired proteins, and captured it on ZR-coated slides. The ZE/ZR dimer was based on the original design by Vinson et al.,[12] who showed the heterodimerization affinity was around 10^{-15} *M*. By incorporating an unnatural amino acid into the ZR domain that could be photo-cross-linked to modified glass surfaces, the authors were able to achieve covalent immobilization. Two model proteins, GST and EGFP, were spotted and shown to have higher spot intensities than controls without ZE domain.

Small Molecule-Based Immobilization

Similar to peptide/protein-based immobilization methods, small molecule-based approaches usually require the target proteins be genetically engineered, then modified with small molecules, e.g., biotin and its conjugates, for subsequent immobilization onto appropriately coated glass surfaces. For example, Walsh et al. used Sfp phosphopantetheinyl transferase to mediate site-specific covalent immobilization of target proteins fused to a peptide carrier protein (PCP) which was originally excised from a nonribosomal peptide synthetase (NRPS).[13] Over the past few years, we have explored intein-based protein modification methods and successfully used them to immobilize proteins onto a protein microarray.[14–18] In the following sections, we will elaborate these intein-mediated strategies in more details.

Intein-Mediated Biotinylation Strategies to Generate Protein Microarray

By taking advantage of the extremely high affinity between biotin and avidin/streptavidin ($K_d \sim 10^{-15}$ *M*), we developed intein-mediated approaches to express recombinant proteins, which can then be site-specifically biotinylated at the C terminus. The resulting proteins are therefore suitable for protein microarray generation (Figure 7.1).[17,18] Intein-mediated protein expression, originally developed for easy and effective purification of fusion proteins on chitin columns,[19] had previously been used to modify proteins with a number of chemical tags.[20] Our biotinylation strategies may be carried out either *in vitro*, *in vivo*, or in a cell-free expression system (Methods A, B, and C in Figure 7.1, respectively).

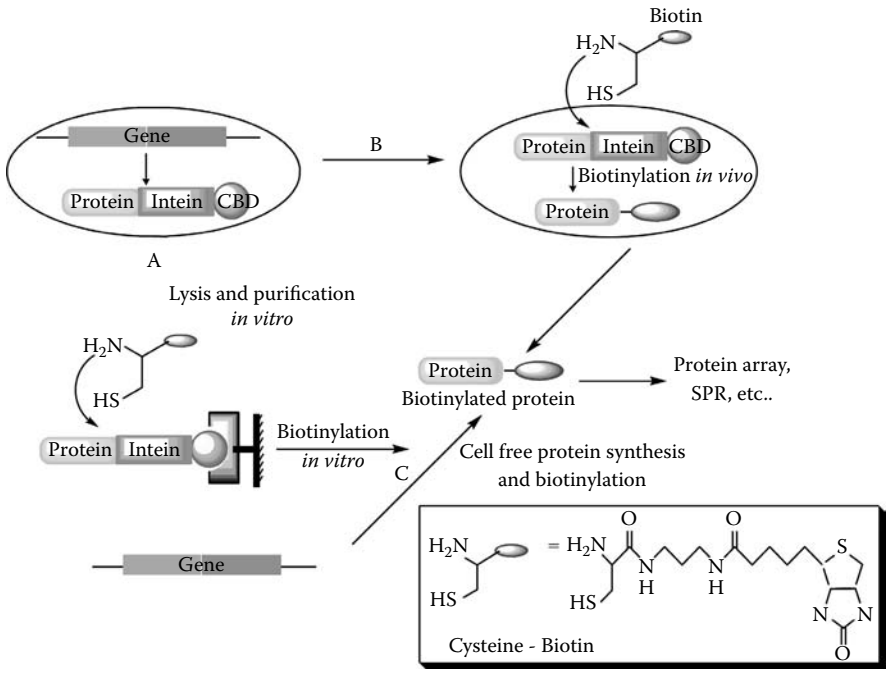


FIGURE 7.1 Three intein-mediated protein biotinylation strategies: (A) *in vitro* biotinylation of column-bound proteins; (B) *in vivo* biotinylation in live cells; (C) cell-free biotinylation of proteins. See color insert following page 236.

Key Aspects of Intein-Mediated Biotinylation Strategies

1. Proteins are site-specifically biotinylated at their C-termini, leading to their subsequent immobilization on avidin-functionalized surfaces in a uniform orientation.
2. Biotin is a small molecule (0.24 kDa), thus minimizing the potential perturbation to the protein's native biological activity.
3. Various formats are applicable with the intein-mediated, protein biotinylation strategies (*in vitro*, *in vivo* or cell-free), thus allowing easy access to desired biotinylated proteins from crude cellular lysates (or mixtures of unpurified proteins) for subsequent protein immobilization and microarray generation.
4. Avidin is an extremely stable protein, making it an excellent candidate for slide derivatization and immobilization.
5. Each avidin/streptavidin molecule can bind rapidly and almost irreversibly up to four molecules of biotin, thus doing away with the long incubation time which alternative methods typically need for the critical immobilization step.
6. Avidin also acts as a molecular layer that minimizes nonspecific binding of proteins to the slide surface, thereby eliminating blocking procedures and minimizing background signals in downstream screenings.

Protein Biotinylation *In Vitro*

In our *in vitro* strategy (Method A; Figure 7.1), the protein of interest was fused through its C-terminus to an intein, which contains a chitin-binding domain as an affinity tag.[17,18] To biotinylate the protein *in vitro*, the host cell over-expressing the protein of interest was first lysed and the lysate containing the intein fusion protein was loaded onto a column packed with chitin beads. Following addition of a thiol-cleaving reagent (for example, cysteine-biotin; inset in Figure 7.1), the fusion protein underwent an on-column self-cleavage reaction, catalyzed by the fused intein, to generate a protein having a reactive -thioester group at its C-terminus. The thioester moiety was subsequently quenched by the thiol side-chain from the added cysteine-biotin, resulting in a thioester-linked intermediate that spontaneously rearranged to form a native peptide bond and generated the target protein which was site-specifically biotinylated at its C-terminus (see Protocol 1). We have shown that this strategy is capable of biotinylating a variety of proteins from different biological sources in 96-well formats,[17] making it possible for future high-throughput generation of a large number of biotinylated proteins needed in a protein microarray.

Protein Biotinylation *In Vivo*

We also successfully carried out the intein-mediated strategy to biotinylate proteins *in vivo* in both bacterial and mammalian systems.[17] Early attempts of *in vivo* protein biotinylation had relied on fusing proteins at the N- or C-termini with a 15 amino acid peptide, the Avitag™ (GLNDIFEAQKIEWHE) which was subsequently biotinylated by biotin ligase- a 35.5 kDa monomeric enzyme encoded by the birA gene in *E. coli*.[21] Biotin ligase catalyzes the transfer of biotin to the ε-amino group of a specific lysine residue within the Avitag, *in vitro* or *in vivo*. Unfortunately, *in vivo* biotinylation of proteins mediated by biotin ligase is often inefficient due to a limiting amount of biotin ligase in the cells (over expression of BirA in bacterial cells results in the formation of inclusion bodies) and is highly cytotoxic.[21] In our approach (Method B, Figure 7.1), the simple addition of the cell-permeable cysteine-biotin probe to the culture media containing cells expressing the target protein, followed by a brief incubation, resulted in a substantial biotinylation of the protein inside the cells. Further optimizations of the cell growth and *in vivo* biotinylation conditions led to an increased level (90 to 95%) of protein biotinylation in the cells. Following *in vivo* labeling the cells are lysed and the crude lysate can be directly spotted onto microarrays (Protocol 2). Endogenous nonbiotinylated proteins present in the cell lysate can be washed away in an efficient and highly-parallel fashion (thousands of different protein spots could be processed simultaneously on a single glass slide), so that protein purification and immobilization are essentially carried out in a single step to generate functional protein microarrays.[17] This is true because of the rare occurrence of naturally biotinylated proteins in the cell, and the highly specific and strong nature of biotin/avidin interaction, which can withstand extremely stringent washing/purification conditions otherwise impossible with other affinity tags.[16]

For both systems (*in vitro* and *in vivo*), apart from endogenous biotinylated proteins, the only other biotinylated protein was the target protein. We have found that the efficiency of intein-mediated protein biotinylation, both *in vitro* and *in vivo*,

depends greatly on the intein fused to target protein — as much as 2- to 10-fold improvement in protein biotinylation may be achieved by the simple switch in the intein used.[15]

Protein Biotinylation in a Cell-Free System

The intein approach has also been extended to a cell-free protein synthesis system (Method C; Figure 7.1). A cell-free system has many advantages[22] over both the *in vitro* and *in vivo* methods described. Potentially a large number of proteins could be simultaneously expressed in a matter of hours in 96- or 384-well formats using commercially available, cell-free protein translation systems. Cellular toxicities due to the over expression of certain proteins, possible degradation by endogenous proteases and formation of inclusion bodies by proteins can be all together avoided as well.

We recently reported another cell-free strategy which utilizes puromycin-containing small molecules to site-specifically biotinylate proteins at their C-termini (Figure 7.2).[16] Puromycin is an aminonucleoside antibiotic produced by *Streptomyces alboniger*. As puromycin resembles the 3' end of the aminoacyl-tRNA, it competes with the ribosomal protein synthesis by blocking the action of the peptidyl transferase, leading to inhibition of protein synthesis in both prokaryotic and eukaryotic ribosomes. It was previously found that, at low concentrations, puromycin and its analogs act as noninhibitors of the ribosomal protein synthesis and get incorporated at the C-terminus of the newly synthesized protein.[23] Our approach thus exploited a similar phenomenon for protein biotinylation. Using this newly developed method, we showed biotinylated proteins could be obtained in a matter of hours using plasmids or PCR products as DNA templates, and that this method is compatible with other high-throughput cloning/proteomics methods such as the Gateway® (Invitrogen, Carlsbad, CA) cloning strategy (Protocol 3).[15]

Immobilization of Biotinylated Proteins onto a Microarray

Following the expression of biotinylated proteins using the various approaches described above, the proteins could be spotted onto avidin-functionalized glass slides, and detected using specific analytes such as antibodies. Using optimized procedures (e.g., Protocols 4 and 5), we have successfully immobilized many different proteins onto avidin-functionalized slides. In most cases, we were able to retain sufficient functional activity of the immobilized proteins (Figure 7.3). Our studies also revealed that the interaction between the biotinylated protein and avidinfunctionalized slide was highly stable and able to withstand harsh treatments, including 1 *M* acetic acid at pH 3.3, 60°C water, and 4 *M* guanidium hydrochloride (Figure 7.4): no reduction in the intensity of printed protein signals was detected when probed with a fluorescein-labeled, anti glutathione-S-transferase (FITC-anti-GST) antibody.

Protein Immobilization via an N-Terminal Cysteine

In a separate but complementary method, the Ssp intein tag was used to generate N-terminal cysteine-containing proteins for site-specific immobilization onto thioester-functionalized glass slides by means of a highly specific chemical reaction

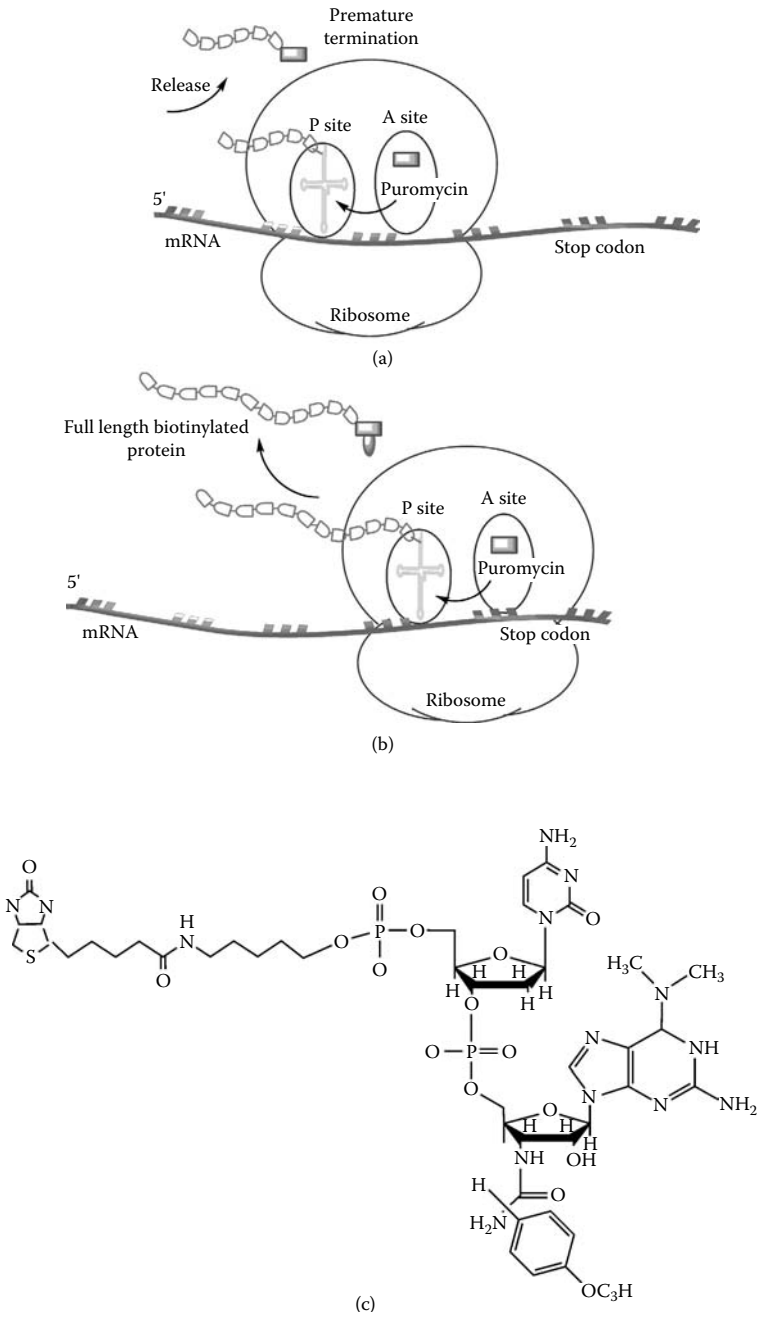


FIGURE 7.2 Puromycin-assisted protein biotinylation. (A) At a high concentration, puromycin binds nonspecifically to nascent protein, bringing about premature termination; (B) At a low concentration, puromycin binds to full length protein at the stop codon; (C) Structure of the 5'-biotin-dC-Puromycin used for protein biotinylation.

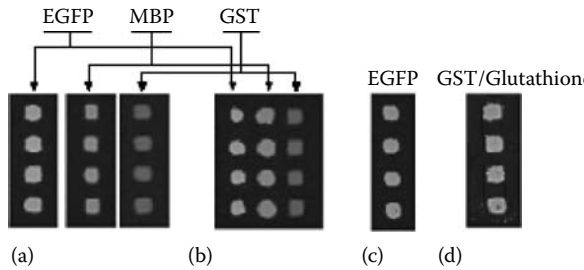


FIGURE 7.3 Site-specific immobilization of biotinylated, functionally active proteins onto avidin slides. (a) EGFP, MBP, and GST were individually detected with Cy3-anti-EGFP (green), Cy5-anti-MBP (red), and FITC-anti-GST (blue), respectively; (b) specific detection of all three proteins with a mixture containing all three antibodies; (c) fluorescence from the native EGFP; (d) specific binding between GST and its Cy3-labeled natural ligand, glutathione. No binding between glutathione and EGFP/MBP was observed.

known as native chemical ligation.[24,25] Terminal cysteine-containing proteins were generated using the pTWIN vectors (Figure 7.5). These vectors allow the expression of target proteins with the self-cleavable modified Ssp DnaB intein having a chitin binding domain fused at their N-termini. The recombinant protein was engineered by standard PCR-based methods and subsequently expressed to have a cysteine residue as its N-terminus, by simply inducing protein expression (the physiological pH induces complete cleavage of the intein from the fusion protein, generating an N-terminal cysteine). Following induction, cells were lysed and the crude cell lysate with the N-terminal cysteine-containing protein could be site-specifically immobilized onto thioester-functionalized slides via the chemoselective native chemical ligation reaction.[14,26–28] Only the terminal cysteine residue reacts with the thioester to form a stable peptide bond; other reactive side chains, including internal cysteines, do not react to form a stable product (Protocols 6 and 7).

For a trial study, two N-terminal cysteine-containing proteins, enhanced green fluorescent protein (EGFP) and GST, were generated and immobilized onto

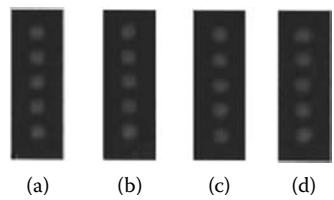


FIGURE 7.4 Biotinylated GST on avidin slides subjected to different washing conditions. (a) 30 min in 1 *M* acetic acid at pH 3.3, (b) 30 min in 60°C water, (c) 30 min in 4 *M* guanidimium hydrochloride, and (d) control slide with no treatment. Slides probed with FITC-anti-GST.

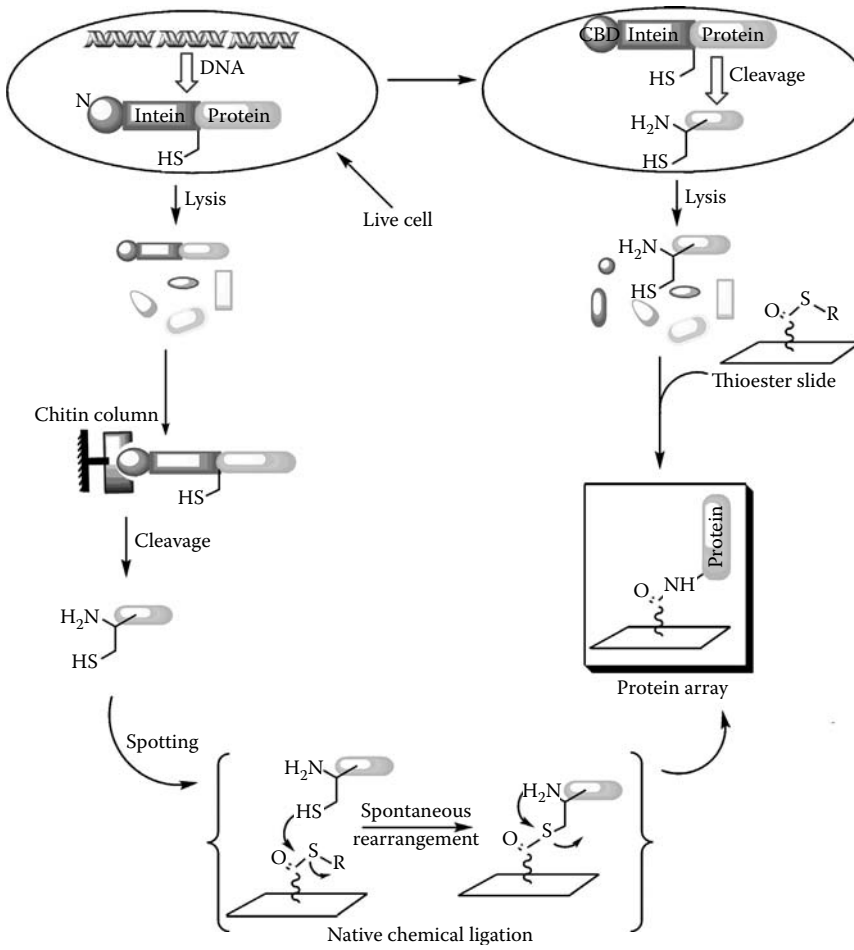


FIGURE 7.5 Site-specific immobilization of N-terminal cysteine-containing proteins using thioester-derivatized glass slides. The N-terminal cysteine-containing proteins were expressed using intein-fused proteins.

PEG-thioester functionalized glass slides (Figure 7.6). The immobilized proteins were successfully detected with specific antibodies conjugated with a fluorescent dye. They were shown to retain their biological activities. EGFP fluorescent intensity showed no significant decrease with prolonged storage and stringent wash conditions.

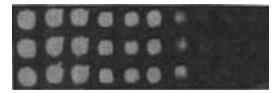


FIGURE 7.6 EGFP printed onto PEG-thioester-functionalized slides in decreasing protein concentrations from 1 mg/ml to 0.001 mg/ml with Cy5-labeled anti-EGFP.

Immobilization Using Genetic Tags

DNA microarrays can be fabricated more easily and enjoy longer shelf life than their protein counterparts.[29] Some scientists have come up with strategies that allow nucleic acid-mediated immobilization of proteins, because in addition to proving as robust tools of immobilization, they also provide unique addresses that can allow production, purification and immobilization of a library of proteins in a highly parallel fashion.

Weng et al. tethered *in vitro* translated proteins with their coding mRNAs, and subjected these assemblies on slides printed with complementary nucleotide sequences.[30] This strategy was shown to localize the protein conjugates to predefined "addresses" by simple hybridization. It was also demonstrated that the relative amount of immobilized proteins could be directly controlled by varying the concentration of the capture oligonucleotides spotted on the glass slide. This strategy, termed PROfusion™ technology, adopts traditional DNA microarray strategies for the provision of protein microarrays by self-assembly mediated by DNA hybridization. Along the same vein, Ramachandran et al. have developed an interesting strategy by immobilizing a variety of plasmids (cross-linked using ultraviolet light to psoralen-biotin) that code for target proteins together with a C-terminal GST epitope.[31] During the printing process, anti-GST antibodies were co-immobilized together with avidin and the biotinylated plasmids onto predefined locations on the array. Proteins were expressed by subjecting the array surface to *in vitro* transcription and translation, allowing each protein to be immobilized *in situ* through the GST tag. Cross reactivity between spots was shown to be negligible by using suitable spotting densities as well as other optimized conditions. The strategy, termed nucleic acid programmable protein array (NAPPA) enables long-term storage of the stable DNA microarrays, which can be readily converted, when required, into active protein microarrays.

PROTOCOLS

Protocol 1

In vitro protein biotinylation using intein-mediated strategy

1. Transform the plasmid into a suitable host strain that bears the T7 RNA polymerase gene under an inducible promoter; induce protein expression under optimal conditions.[a]
2. Harvest cells by centrifugation (5000 g, 15 min, 4°C). Discard supernatant. Store pellets at –20°C or immediately proceed to cell lysis.
3. Resuspend pellet from 1 liter culture into 50 ml cold lysis buffer (20 m*M* Tris-HCl pH 8.0, 500 m*M* NaCl, 1 m*M* EDTA, 0.1% Triton X-100).[b] Lyse cells by sonication or using French press.[c]
4. Clarify the lysate by centrifugation (20,000 g, 30 min, 4°C) and collect the supernatant.
5. Pack the desired volume of chitin beads into a column (3 ml of beads is sufficient for protein purification from 200 ml of culture).

6. Prior to loading of the crude cell lysate, pre-equilibrate the column with 10 column volumes of column buffer (20 m*M* Tris-HCl pH 8.0, 500 m*M* NaCl, and 1 m*M* EDTA) at 4°C.[d]
7. Load the clarified cell lysate onto the column at a flow rate of 0.5 ml/min.
8. Wash the column with ≥ 30 column volumes of column buffer at a flow rate of 2 ml/min to remove all traces of contaminating proteins.
9. Quickly flush the column with 3 bed volumes of column buffer containing 30 m*M* cysteine-biotin; stop the flow and incubate the column overnight at 4°C.[e]
10. Elute biotinylated target protein using column buffer or a specific buffer for long-term storage of proteins.
11. Desalt with a NAP-5 column, if necessary, before proceeding to spotting.

Protocol 2

In vivo protein biotinylation in bacterial cells

1. The initial steps of cloning, transformation and induction of protein expression are essentially the same as the *in vitro* based method.[16–17]
2. Following induction, add MESNA and cysteine-biotin to the induced bacterial culture to a final concentration of 30 m*M* and 3 m*M*, respectively; incubate at 4°C for 24 h with constant shaking.
3. Harvest cells by centrifugation (6000 g, 15 min, 4°C).
4. Wash the cell pellet at least twice with PBS to remove excessive unreacted cysteine-biotin.
5. Resuspend the cell pellet in 1 ml lysis buffer for protein extraction. Cell pellet can also be stored at –20°C without any significant degradation of the biotinylated protein.
6. Lyse the cells by sonication on ice at 50% duty, 20% power in 5 treatments of 30 sec each with 30 sec cooling interval or by using French press.
7. Remove cell debris by centrifugation (20,000 g, 20 min, and 4°C) and collect clarified cell lysate (supernatant).
8. The cell lysate can be spotted directly onto the avidin slide without any further treatment.

Protocol 3

Protein biotinylation in a cell-free system.

1. Prepare reaction solution in one 0.2-ml PCR tube from the Rapid translation system RTS 100 *E. coli* HY kit (i.e., 12 µl *E. coli* lysate, 10 µl reaction mix, 12 µl amino acids, 1 µl of 1 m*M* methionine, 5 µl of reconstitution buffer) on ice.
2. Add DNA template: 0.5 µg of plasmid DNA, 0.5 µg of linear template generated via a standard PCR reaction or 0.1 µg of linear template generated via two-step PCR containing gene of interest and appropriate T7 regulatory regions.[f]

3. Add 5′-biotin-dC-Pmn to a final concentration of 35 μ*M* and RNase-free deionized water to a final reaction volume of 50 μl.
4. Start the reaction at 30°C for 6 h in a thermal cycler.[g]
5. Remove reaction solution from the thermal cycler and store it at –20°C until further processing.
6. Use 5 μl of the reaction solution for western blot to confirm the presence of the biotin labeled target protein.
7. For downstream microarray application, proteins need to be desalted. Prepare the MicroSpin™ G-25 column by resuspending the resin in the column (vortexing gently).
8. Loosen the cap one-fourth turn and snap off the bottom closure.
9. Place the column in a 1.5-ml screw-cap microcentrifuge tube for support. Alternatively, cut the cap from a flip-top tube and use this tube for support.
10. Prespin the column for 1 min at 735 g.
11. Place the column in a new 1.5-ml tube and slowly apply the reaction solution to the center of the angled surface of the compacted resin bed, being careful not to disturb the resin. Careful application of the reaction solution to the center of the bed is essential for good separation. Do not allow any of the reaction solution to flow around the sides of the bed.
12. Spin the column for 2 min at 735 g and collect the desalted reaction solution at the bottom of the support tube.
13. Discard the column, and the reaction solution is ready for spotting onto avidin slides.

Protocol 4

Preparation of avidin-functionalized slides.

1. Clean glass slides in piranha solution for at least 2 h.[h]
2. Wash the slides copiously with deionized water, rinse with 95% ethanol, and finally dry the slides.
3. Soak the freshly clean slides in glycidyloxypropyl trimethoxysilane for 1 h.
4. Place the derivatized slides in a slide holder and wash two to three times with 95% ethanol.
5. Cure slides at 150°C for at least 2 h (overnight curing gives the same result). Rinse the slides with ethanol and dry.
6. Add 40–60 μl of 1 mg/ml avidin onto the slides, cover with cover slip and incubate for 30 min.[i]
7. Subsequently, wash the slides with deionized water in slide tray and dry the slides.
8. React the remaining epoxides by adding 2 m*M* aspartic acid onto the slides and covering with cover slip.
9. Finally, wash the slides with deionized water and dry them for spotting.

Protocol 5

Immobilization of biotinylated proteins onto avidin-functionalized slides.

1. Add 10 µl of the clarified cell lysate or reaction solution into source plate.
2. Spot the cell lysate or reaction solution onto the avidin-functionalized slides using an ESI SMA™ arrayer.
3. Incubate the spotted slides at room temperature for approx. 2 to 3 h.
4. Wash spotted slides with PBS for a few minutes before drying in air.
5. To visualize the immobilized proteins on the avidin slides, incubate the spotted slides with fluorescently labeled monoclonal antibody for 1 h.[j]
6. Wash the slides twice with PBST on an orbital shaker (each time for 15 min).
7. Finally, rinse the slides with distilled water to remove salt debris.
8. Dry slide and visualize spots with an ArrayWoRx microarray scanner.

Protocol 6

Preparation of thioester-derivatized slides.

1. Incubate epoxy-derivatized slides with 10 m*M* diamine-PEG for 30 min. Protocols for slide preparation have been described elsewhere.[24,25]
2. Wash slides with deionized water and place them in a solution of 180 m*M* succinic anhydride for 30 min and then in boiling water for 2 min.
3. Prepare NHS solution and incubate it with the slides for 3 h.
4. Rinse slides with deionized water and react overnight with a solution of benzylmercaptan.
5. Finally, wash the slides with deionized water and dry them for spotting.
6. Add 10 µl of the clarified cell lysate from Protocol 7 into source plate.
7. Spot the cell lysate onto the thioester-functionalized slides using an ESI SMA arrayer, and incubate for approx 10 min.
8. Wash spotted slides with PBS for a few minutes before drying in air.
9. Slides are now ready for detection by fluorescently labeled monoclonal antibody.[j]

Protocol 7

Expression/immobilization of N-terminal, cysteine-containing proteins.

1. Inoculate 2 ml of freshly grown transformed ER2566 cells into 200 ml of LB medium supplemented with 100 µg/ml ampicillin.
2. Incubate the culture at 37°C in a 250-rpm incubator shaker to an OD600 of approx 0.5 (about 3 h).
3. Add IPTG to a final concentration of 0.3 to 0.5 m*M* to induce fusion protein expression.
4. Incubate the culture overnight at room temperature on an orbital shaker. For optimization of *in vivo* cleavage of fusion protein, incubate the culture for at least 18 h before harvesting.[k]
5. Harvest cells by centrifugation (6000 g, 15 min, 4°C).

6. Discard supernatant and resuspend cell pellet in 5 ml lysis buffer.
7. Lyse bacterial cells by sonication on ice at 50% duty, 20% power, in three treatments of 30 s each with 30-s cooling interval.
8. Centrifuge the cell lysate at 20,000 g, 30 min, and 4°C.
9. Use clarified supernatant for direct spotting onto thioester glass slides

Protocol Notes

a. ER2566 is available from NEB for expression. Other strains [BL 21 (DE3) and its derivates] can also be used for expression.
b. Other nonionic detergents (0.1 to 0.2% Tween 20), protease inhibitors (PMSF, pepstatin, leupeptin) and reducing agents like 1 m*M* TCEP/TCCP can be included to stabilize target proteins. The presence of thiol reagents (β-mecaptoethanol, 1,4 dithiothreitol, cysteine) will cause premature cleavage of fusion protein, resulting in a loss of target protein before affinity purification. As such, thiol compounds should be strictly avoided in all steps to maximize target protein recovery.
c. Lysozyme binds and digest chitin and should be avoided during lysis. If alternate methods for lysis are not available, mild treatment with lysozyme (10 to 20 μg/ ml, 4°C, 1 h) can be used.
d. All purification steps should be carried out at 4°C to ensure stability of fusion protein.
e. Several factors like the amino acid residue at the cleavage site, duration, pH and temperature during cleavage may affect the cleavage efficiency and hence the final yield of protein. For proteins which do not cleave effectively, longer time (40 h) at a higher temperature (16 to 23°C) and pH (9.0) may be used.
f. Any vector or linear DNA to be used in combination with the Rapid Translation System must include the following elements and structural features: (1) target gene under control of T7 promoter located downstream of a ribosomal binding site (RBS) sequence, (2) distance between T7 promoter and start ATG should not exceed 100 base pairs, (3) distance between the RBS sequence and start ATG should be more than five to eight base pairs, and (4) T7 terminator sequence at the 3′ end of the gene. A two-step PCR protocol is recommended for incorporation of the 5′ and 3′ T7 regulatory regions into the linear template. The purity (OD260/280 = 1.7) of plasmids obtained from commercially available DNA preparation kits is sufficient for the use as template in the Rapid Translation System.
g. Optimal temperature for most protein synthesis is 30°C. However, lower temperatures may be used for proteins that tend to aggregate. Protein synthesis can proceed for up to 6 h, but the synthesis reaction is usually 90% complete after 4 h.
h. When handling the slides, care must be taken to ensure that the slide is kept clean at all times, and that nothing comes into contact with spotting surface. Dust especially may result in extraneous fluorescence and may affect the fluorescent readout when the slide is scanned. Also gloves, if

used, should be of the powder-free variety to ensure that the slides remain uncontaminated even after handling.

i. If there is sufficient reagent, it may be convenient to react both surfaces of the slides by placing them on slide racks in deep-well dishes. However, for expensive reagents, where it is preferable to utilize a conservative volume of the chemical, coverslips may be used. For a 22 × 60 mm coverslip, a 50-µl preparation is sufficient to allow for confluent coverage. Two methods may be used to apply the reagent on the surface. Either the reaction mix is first applied to the slide, and the coverslip is applied, or it could be applied to the coverslip and the slide may be inverted upon it. Both methods work equally well, but one ought to use the method that would allow production of a uniform spread of the reagent across the slide surface, without introducing any bubbles or voids between the coverslip and the slide (where the reagent does not come into contact with the slide surface). Coverslips may be slid off the slide once the reaction is complete, or be removed by vigorously shaking the slide in a water (or solvent) bath, until the coverslip slowly comes off.

j. Fluorescently labeled monoclonal antibody against target protein can be used to confirm successful immobilization of biotinylated proteins onto avidin slides. Some of these fluorescently labeled monoclonal antibodies may be commercially available, while others might require self labeling using fluorescent dye.

k. No further treatment of the clarified cell lysate was needed prior to spotting, since trace amounts of the cysteine-biotin probe and endogenous biotinylated protein (acetyl-CoA carboxylase) in the *E. coli* lysate did not seem to interfere with binding of the target protein to the avidin slide.

CONCLUSION AND FUTURE DIRECTIONS

While the most convenient method of immobilization is physical adsorption, it is known that it might render active sites inaccessible or even denature proteins.[32] Cha et al. experimentally documented that the activity of randomly immobilized enzyme was roughly 5 to 6 times lower than that of enzyme with controlled orientation on the microarray slides.[4] Therefore, site-specific immobilization which maintains native protein conformation is generally required to generate a functional protein microarray. To this end, however, it typically entails painstaking cloning and expression of individual protein before spotting on the surface of microarray slide. A facile site-specific and stable immobilization of all the proteins from a proteome in microarray format is currently a technical hurdle. In the light of these technical challenges, there is considerable room for new innovative strategies to immobilize proteins in the field of protein microarrays.

REFERENCES

1. Hu, Y., Uttamchandani, M., and Yao, S.Q. Microarray: A versatile platform for high-throughput functional proteomics, *Comb. Chem. High Throughput Screening*, 9, 203, 2006.
2. Uttamchandani, M., Wang, J., and Yao, S.Q., Protein and small molecule microarrays: Powerful tools for high-throughput proteomics, *Mol. BioSyst.*, 2, 58, 2006.

3. LaBaer, J. and Ramachandran, N., Protein microarrays as tools for functional proteomics, *Curr. Opin. Chem. Biol*., 9, 14, 2005.
4. Cha, T., Guo, A., and Zhu, X.Y., Enzymatic activity on a chip: The critical role of protein orientation, *Proteomics*, 5, 416, 2005.
5. Yeo, S.Y.D. et al., Strategies for immobilization of biomolecules in a microarray, *Comb. Chem. High Throughput Screening*, 7, 213, 2004.
6. Zhu, H. et al., Global analysis of protein activities using proteome chips, *Science*, 293, 2101, 2001.
7. Paborsky, L.R. et al., A nickel chelate microtiter late assay for six histidine-containing proteins, *Anal. Biochem*., 234, 60, 1996.
8. Hodneland, C.D. et al., Selective immobilization of proteins to self-assembled monolayers presenting active site-directed capture ligands, *Proc. Natl. Acad. Sci. USA*, 99, 5048, 2002.
9. Bjorkling, F. et al., Inhibition of lipases by phosphonates, *Bioorg. Med. Chem*., 2, 697, 1994.
8. Kindermann, N. et al., Covalent and selective immobilization of fusion proteins, *J. Am. Chem. Soc*., 125, 7810, 2003.
9. Camarero, J.A., Kwon, Y., and Coleman, M.A., Chemoselective attachment of biologically active proteins to surfaces by expressed protein ligation and its application for "Protein Chip" fabrication, *J. Am. Chem. Soc*., 126, 14730, 2004.
10. Choi, Y.S., Pack, S.P., and Yoo, Y.J., Development of a protein microarray using sequence-specific DNA binding domain on DNA chip surface, *Biochem. Biophys. Res. Commun.*, 329, 1315, 2005.
11. Zhang, K., Diehl, M.R.D., and Tirell, A., Artificial polypeptide scaffold for protein immobilization, *J. Am. Chem. Soc*., 127, 10136, 2005.
12. Moll, J.R. et al., Designed heterodimerizing leucine zippers with a ranger of pIs and stabilities up to 10-15 M, *Protein Sci*., 10, 649, 2001.
13. Yin, J. et al., Labeling proteins with small molecules by site-specific posttranslational modification, *J. Am. Chem. Soc*., 126, 7754 , 2004.
14. Girish, A. et al., Site-specific immobilization of proteins in a microarray using intein-mediated protein splicing, *Bioorg. Med. Chem. Lett*., 15, 2447, 2005.
15. Tan, L.P. et al., Improving the intein-mediated, site-specific protein biotinylation strategies both *in vitro* and *in vivo*, *Bioorg. Med. Chem. Lett*., 14, 6067, 2004.
16. Tan, L.P., Chen, G.Y.J., and Yao, S.Q., Expanding the scope of site-specific protein biotinylation strategies using small molecules, *Bioorg. Med. Chem. Lett*., 14, 5735, 2004.
17. Lue, R.Y.P. et al., Versatile protein biotinylation strategies for potential high-throughput proteomics, *J. Am. Chem. Soc*., 126, 1055, 2004.
18. Lesaicherre, M.L, Intein-mediated biotinylation of proteins and its application in a protein microarray, *J. Am. Chem. Soc*., 124, 8768, 2002.
19. (a) Chong, S.R. et al., Single-column purification of free recombinant proteins using a self cleavable affinity tag derived from a protein splicing element, *Gene.*,192, 277, 1997; (b) Chong, S.R. et al., Utilizing the C-terminal cleavage activity of a protein splicing element to purify recombinant proteins in a single chromatographic step, *Nucleic. Acids. Res*., 26, 5109, 1998.
20. (a) Tolbert, T. and Wong, C.H., Intein mediated synthesis of proteins containing carbohydrates and other molecular probes, *J. Am. Chem. Soc*., 122, 5421, 2000; (b) Yeo, S.Y.D. et al., Cell-permeable small molecule probes for site-specific labeling of proteins, *Chem. Commun*., 23, 2870, 2003; (c) Srinivasan, R., Yao, S.Q., and Yeo, S.Y.D., Chemical approaches for live cell bioimaging, *Comb. Chem. High Throughput Screening*., 7, 597, 2004.

21. Cronan, J.E. and Reed, K.E., Biotinylation of proteins *in vivo*: A useful posttranslational modification for protein analysis, *Methods Enzymol*., 326, 440, 2000.
22. He, M.Y. and Taussig, M.J., Single step generation of protein arrays from DNA by cell-free expression and *in situ* immobilisation (PISA method), *Nucleic Acids Res*., 29, e73, 2001.
23. (a) Nemoto, N., Miyamoto-Sato, E., and Yanagawa, H., Fluorescence labeling of the C-terminus of proteins with a puromycin analogue in cell-free translation systems, *FEBS. Lett*., 462, 43, 1999; (b) Miyamoto-Sato, E. et al., Specific bonding of puromycin to full-length protein at the C-terminus, *Nucleic Acids Res*., 28, 1176, 2000; (c) Kawahashi, Y. et al., *In vitro* protein microarrays for detecting protein–protein interactions: Application of a new method for fluorescence labeling of proteins, *Proteomics*, 3, 1236, 2003.
24. Muir, T.W., Sondhi, D., and Cole, P.A., Expressed protein ligation: A general method for protein engineering, *Proc. Natl. Acad. Sci. USA*., 95, 6705, 1998.
25. Creighton, T.E., *Proteins: Structure and Molecular Properties*, 2nd ed., Freeman, New York, 1993.
26. Lesaicherre, M.L. et al., Developing site-specific immobilization strategies of peptides in a microarray, *Bioorg. Med. Chem. Lett*., 12, 2079, 2002.
27. Lesaicherre, M.L. et al., Antibody-based fluorescence detection of kinase activity on a peptide array, *Bioorg. Med. Chem. Lett*., 12, 2085, 2002.
28. Uttamchandani, M. et al., Combinatorial peptide microarrays for the rapid determination of kinase specificity, *Bioorg. Med. Chem. Lett*., 13, 2997, 2003.
29. Chen, G.Y.J. et al., Array-based technologies and their applications in proteomics, *Curr. Top. Med. Chem*., 3, 705, 2003.
30. Weng, S. et al., Generating addressable protein microarrays with PROfusion covalent mRNA-protein fusion technology, *Proteomics*, 2, 48, 2002.
31. Ramachandran, N. et al., Self-assembling protein microarrays, *Science*, 305, 86, 2004.
32. Zhu, H. and Snyder, M., Protein chip technology, *Curr. Opin. Chem. Biol*., 7, 55, 2003.

8 Protein *In Situ* Arrays through Cell-Free Protein Synthesis

Mingyue He, Farid Khan, Elizabeth Palmer, Mingwei Wang, and Michael J. Taussig

CONTENTS

INTRODUCTION

A major objective for proteomics studies is the assignment of protein function, in particular with regard to protein networks and interactions. Protein microarray technology is an appropriate tool for this purpose, as it allows the large-scale analysis of many hundreds of proteins in parallel even up to the level of entire proteomes.[1] The wide range of protein array applications is evident from other contributions to this volume. Major challenges for the technology include obtaining proteins for the arrays and maintaining their folding, function, and long-term stability on an array surface. Protein production usually involves recombinant protein expression in one of several *in vivo* expression systems followed by purification. However, using conventional bacterial expression systems is time-consuming and often problematic

due to insolubility, hydrophobicity, presence of disulphide bonds, etc., and many proteins, especially of human origin, are not expressed as functional molecules in heterologous hosts.[2,3] Arraying requires covalent or noncovalent attachment on an appropriate solid surface in such a way as to maintain long-term functionality (binding, enzymatic activity, etc), which can often decline due to degradation and inherent instability of proteins on the array surface.

To overcome these problems, we have developed two cell-free protein array technologies termed "Protein *in situ* Arrays" (PISA)[4] and more recently "DNA Arrays to Protein Arrays" (DAPA). The cell-free systems use DNA (usually PCR products) or mRNA as templates to direct protein synthesis by *in vitro* transcription (for DNA) and translation, enabling the rapid conversion of genetic information into functional proteins and the parallel synthesis of many proteins in a single reaction. In the PISA method as first described, the DNA was in solution or fixed to a bead, while in the new development of DAPA the DNA is surface-immobilized as an array, allowing conversion of DNA arrays directly into protein arrays. Both technologies combine cell-free protein synthesis from PCR DNA with simultaneous immobilization of the protein through a tag system, eliminating the need for independent cloning and protein purification. A particular utility is that proteins generated by PISA or DAPA can be analyzed immediately, avoiding the need for long-term storage and the risk of loss of function.

PROTEIN *IN SITU* ARRAYS (PISA)

Principle

PISA technology generates protein arrays by carrying out protein synthesis on a surface, which is precoated with the protein-capturing reagent, so that the newly-synthesized proteins are specifically captured and immobilized on the surface as soon as they are translated (Figure 8.1). Noncaptured proteins are removed by subsequent washing steps. Suitable array surfaces can include ELISA wells, magnetic beads or glass slides. Protein *in situ* immobilization can be achieved, in general, via a tag sequence, through which the protein is captured by a specific tag-binding reagent such as a chelator or antibody; the active region of the protein should then be available for functional analysis (Figure 8.1a). As an alternative, proteins can be captured by specific ligands coated on the surface, e.g., using antigens to localize expressed antibodies or other binders; in this case the purpose could be to identify binders from libraries (Figure 8.1b).

PCR DNA Construction

A PCR DNA template is required to direct protein synthesis in cell-free systems, such as rabbit reticulocyte lysate, wheat germ, or *E. coli* S30 extract, all of which are commercially available. The DNA construct contains essential elements for protein expression such as promoter (often T7), translation initiation signal, and transcription and translation termination signals. The translation initiation sites differ between eukaryotic and *E. coli* S30 systems, the former using a Kozak sequence

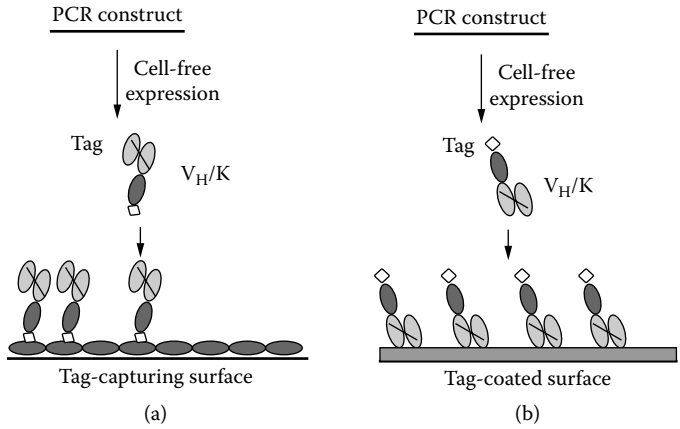


FIGURE 8.1 Principle of protein *in situ* arrays: (a) tag-capture method; (b) ligand-capture method. V_H/K is a three-domain single chain antibody fragment in which V_H is linked to K light chain.

while the latter requires a Shine-Dalgarno (S/D) sequence (Figure 8.2). Termination of transcription and translation has been shown to affect protein expression level, which is significantly decreased without those sequences. The presence of a poly(A) region following the stop codon also promotes protein expression by stabilizing mRNA level.[5]

To immobilize a newly synthesized protein, an affinity tag can be introduced at either the N- or C-terminus of the protein (Figure 8.2). It has been reported that a $(His)_6$ tag may not be accessible when located at the C-terminus in a number of proteins.[6] To remedy this and also reduce any possible interference of the tag sequence on protein folding, a flexible linker can be placed between it and the arrayed

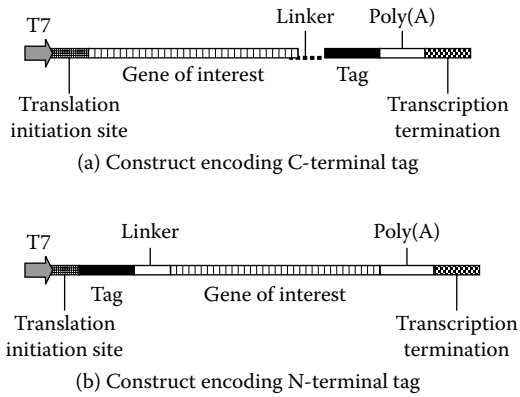


FIGURE 8.2 Structure of PCR constructs for cell-free expression in PISA systems. (a) C-terminal tagged; (b) N-terminal tagged.

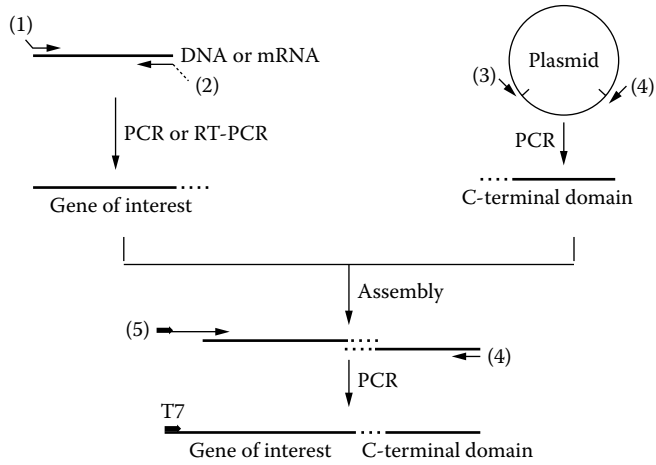


FIGURE 8.3 PCR assembly strategy for PISA expression constructs. The numbers indicate different primers.

protein. We have also designed a novel double-$(His)_6$ tag which binds particularly strongly to Ni-NTA (see below).[4,7]

The PCR DNA construct is generally produced by assembling the gene of interest with the elements for protein expression by overlapping PCR. To simplify the construction process, the sequence elements can be designed in a defined order and cloned into a plasmid, which is used as the template for PCR amplification when required (Figure 8.3). We have designed a plasmid encoding, in order, a flexible linker followed by the double $(His)_6$-tag sequence, two consecutive stop codons (TAATAA), a poly(A)28 region and a transcription termination region.[4] Similarly, the T7 promoter followed by a translation initiation sequence (eukaryotic or prokaryotic) can also be constructed into a plasmid. Quantities of these fragments can be easily produced by PCR and used for the generation of expression constructs for protein *in situ* arrays.

Cell-Free Expression Systems

Most cell-free systems make use of a crude cell lysate containing the protein synthesis machinery with an exogenous supply of essential amino acids, nucleotides, salts and energy-generating factors to direct protein synthesis from added DNA or mRNA template(s). Lysates have been produced from many organisms including human cells.[8,9] Commercially available systems include *E. coli* S30, rabbit reticulocyte, wheat germ, and insect cell lysates. They are designed either for 'coupled' synthesis, where DNA (PCR fragment or plasmid) is the template, or as 'uncoupled' systems, which require an mRNA template. A PURE system has also been developed in which purified individual protein components of the translation machinery are assembled.[10] PURE reconstitutes the coupled transcription/translation process by mixing 31 recombinant soluble protein factors with 46 tRNAs, their essential substrates

and corresponding enzymes, and has been used to produce a number of different proteins with yields of about 100 μg/ml.

The protein expression level of most of these systems has been improved continuously. Protein yields in the mgs/ml range have been reported using *E. coli* S30 and wheat germ extracts.[11,12] A wide-range of protein families, including proteins with molecular sizes up to 400 kDa, protein complexes, proteins with disulphide bridges and membrane proteins, which are often not produced in cell-based expression systems, have all been functionally expressed efficiently in cell-free systems.[13]

The use of cell-free systems has a number of other advantages. As well as giving a rapid synthesis of proteins from their corresponding PCR fragments, external components can be added, making them highly suitable for synthesis of folded or modified proteins under defined conditions.[14] Fluorescent or chemically-modified amino acids can be incorporated into proteins at specified positions during translation through tRNA methodologies, providing a powerful means of arraying labeled proteins for sensitive detection.[15]

PISA Procedure

Arraying by the PISA method is carried out by the following procedures.

Tag-Capture

The protein of interest is synthesized with an affinity tag fused either at the N- or C-terminus and immobilization achieved through interaction between the tag and a capturing reagent coated on the array surface. The tag can be introduced by fusing the open reading frame encoding the protein with a tag-encoding DNA sequence or labeling with small molecules during translation. A number of DNA-encoded tags have been employed,[16] including single and double-$(His)_6$,[4,7] GST[18] and AV1.[19] As cell-free systems allow site-specific incorporation of nonnatural or modified amino acids during protein synthesis using tRNA methodologies, newly synthesized proteins can be labeled with small molecules such as biotin, photo-reactive cross-linked groups or fluorescent moieties for immobilization and sensitive detection. Such labels can also be directed to a defined position in the primary sequence. For example, the N-terminus can be labeled using a fluorescently modified initiator methionine-tRNA (fmet-tRNA)[20] or suppressor tRNA[21] while C-terminal labeling can be achieved with puromycin analogues.[22] Labeling at internal sites has been achieved through the use of stop codon suppression.[15]

We have introduced a novel double-$(His)_6$ tag for binding to Ni-NTA-coated surfaces, which are available as microtiter plates, magnetic agarose beads, BIAcore chips and glass slides.[4,7] The double-$(His)_6$ tag sequence comprises two hexa-histidines separated by an 11-amino-acid spacer and has shown an order of magnitude or greater affinity for Ni-NTA modified surfaces than a conventional single-$(His)_6$ tag in ELISA and BIAcore studies[7] (Figure 8.4). Binding to Ni-NTA surfaces is sufficiently strong for the immobilized proteins to be reused after reagent stripping to remove the detection molecules.[4] We have also shown that the double-$(His)_6$ tag is detectable by anti-His antibodies even after binding to Ni-NTA.[7] Moreover, it significantly improves the functional properties of 'conventional' antibody arrays.[23]

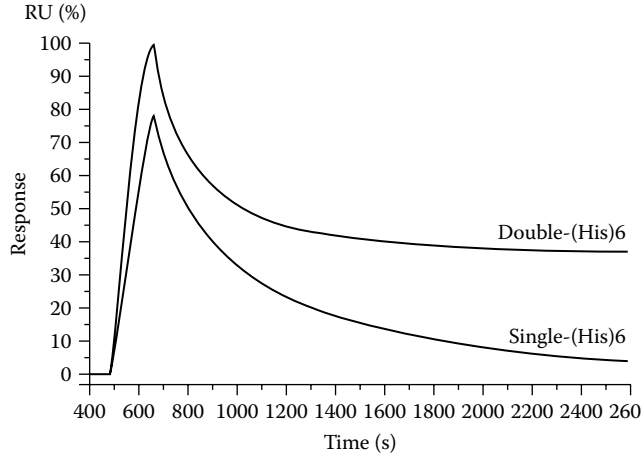


FIGURE 8.4 BIAcore data showing binding of a double-$(His)_6$ and single $(His)_6$-tagged green fluorescent protein (GFP) constructs to a Ni-NTA chip.

Ligand-Capture

The ligand-capturing method is carried out by cell-free synthesis of proteins on a ligand-coated surface. Proteins with binding activity are detected *in situ* on the surface for direct screening of functional binding activity and specificity. We have applied this procedure to screening of antibodies selected by ribosome display, leading to identification of high-affinity antibody fragments on antigen-coated slides (Figure 8.5).

In the original descriptions,[4,24] the PISA method was carried out in the wells of microtiter plates coated with the protein- or tag-capture reagent (e.g., Ni-NTA for the double-$[His]_6$ tag). Typically, 50 to 100 ng of DNA are mixed with 25 µl of cell free extract and incubated for 2 h at 30°C. The surface carrying the bound protein is then washed 3 times with PBS/0.05% Tween 20 to remove nontagged proteins. When using rabbit reticulocyte, there is some binding of free hemoglobin, which can be removed by washing with 20 m*M* imidazole. The procedure can also be

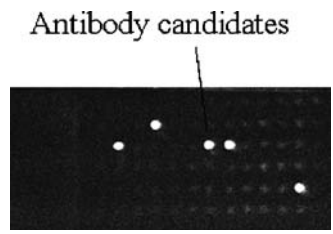


FIGURE 8.5 The use of a ligand-capture PISA for screening anti-progesterone antibodies on a progesterone-BSA coated slide; the identified binding antibodies are indicated.

miniaturized by spotting 100 nl of cell-free protein synthesis mixture together with the DNA onto a Ni-NTA coated slide (unpublished). We also showed that 5′-biotinylated DNA immobilized on streptavidin beads will template cell-free protein synthesis and that this could lead to new complex particles in which a protein and its encoding DNA were present on the same bead (a novel phenotype-genotype linkage).

MAKING PROTEIN ARRAYS FROM DNA ARRAYS

A procedure entitled Nucleic Acid Programmable Protein Arrays (NAPPA), recently reported by LaBaer and colleagues, produces a protein array from an immobilized DNA array template.[25] In this method, cloned plasmid DNA is immobilized on a glass slide, which is also precoated with a protein-capturing antibody (e.g., anti-GST where GST is the tag). A cell-free transcription/translation lysate is applied over the entire surface and the synthesized proteins are captured locally by the coated antibodies, with good spot morphology and minimal diffusion. This generates an *in situ* array in which the proteins are immobilized in the vicinity of their encoding DNA, so that each array location contains a mixture of plasmid DNA, antibody and captured protein.

NAPPA demonstrated that a DNA array template could be used directly to make a protein array as and when required. However, this method only permits a single conversion of the DNA array and the fact that each spot is a mixture of DNA, antibody and expressed protein may affect downstream applications. Recently, we have developed a novel system (DNA array to protein array, DAPA) to generate "pure" protein arrays from a PCR DNA array template. With DAPA technology, not only is the protein array generated on a separate surface in a single reaction, but also multiple copies of the same protein array can be produced through repeated use of the same DNA array template (Figure 8.6). In this method, cell-free protein synthesis is performed in a membrane, which is sandwiched between two solid surfaces (e.g., glass slides), one of which is arrayed with the DNA templates while the other is coated uniformly with a reagent to capture the translated proteins. Individual proteins synthesized in parallel from the arrayed DNA pass through the membrane to become immobilized on the opposite surface through interaction with the protein-capturing reagent, forming a protein array with the precise layout of the DNA array (He et al., in preparation).

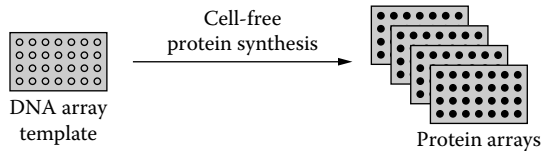


FIGURE 8.6 Principle of DAPA — converting a DNA array into multiple copies of the corresponding protein array.

APPLICATIONS OF CELL-FREE PROTEIN ARRAYS

The above technologies promise to be versatile and flexible tools, allowing rapid expression of cloned or un-cloned DNA sequences and arraying of full-length proteins and individual domains (including those which are difficult to express in bacterial systems) or peptides. Others and we have successfully used them to generate arrays from single-chain antibody fragments, binding-domains, fluorescent proteins, signal transduction proteins, DNA replication initiation proteins and enzymes. The following are a few specific applications:

1. Screening of antibody specificity: A small-scale antibody array was generated by the PISA tag-capture method on Ni-NTA-coated microtiter wells for analysis of binding specificity of human single-chain anti-progesterone fragments selected by ribosome display. The fragments were engineered with the C-terminal double-$(His)_6$ tag. The array was probed with both the antigen (progesterone-BSA) and control proteins and showed that binding could only be detected with the antigen itself. This both confirmed the specificity of the antibody combining site and demonstrated correct folding of the antibody after cell-free expression.[4,24] It was also possible to reprobe the same array after stripping of the detection reagents, due to the strength of protein immobilization by the double-$(His)_6$ tag.
2. Immobilization of enzymes: PISA was used to generate immobilized luciferase on magnetic beads from PCR DNA. Localized luciferase retained full enzymatic activity.[4]
3. Protein domain arrays for interaction mapping: Individual domains from the lymphocyte signal transduction protein Vav-1 were immobilized by PISA in order to map domain–domain interactions. In combination with ribosome display, interaction between the N-terminal SH3 domain of Vav-1 and Grb2 was identified, in agreement with biochemical methods and the yeast two hybrid method (He et al., in preparation).
4. Protein interactions and networks: A DNA array was generated using human genes encoding 29 proteins involved in DNA replication initiation and used as a template to create a protein array by the NAPPA method.[25] The protein array was then probed with each of the 29 proteins in turn, produced as free proteins in the same cell-free system. This study identified 110 interactions including many previously identified by genetic, two hybrid and biochemical methods.
5. Rapid isolation of cell-free synthesized proteins: The protein *in situ* immobilisation concept has been exploited for rapid purification of cell-free synthesized proteins.[26] Starting from DNA templates, $(His)_6$ tagged proteins were produced and isolated on Ni-agarose beads included in the lysate, followed by elution, in less than 1.5 hours. This method may have applications in proteomics where rapid expression and isolation of proteins is required.

CONCLUSION

Cell-free synthesis of proteins from DNA and their concurrent immobilization at a suitable surface can provide solutions to some major problems in array generation and application. Production of the proteins, which are often functional when made in cell-free systems, is multiplexed and easily scaled so that hundreds or thousands can be made in parallel. Once the DNA has been distributed, procedures are quick and avoid cloning and separate protein purification. They provide a means of generating protein arrays 'on demand,' as and when required, so that the likelihood of protein degradation or loss of function is minimized; in consequence, protein arrays should gain in reliability. *In situ* arrays are particularly suitable for functional analysis of protein domains or subregions, all that is required being the appropriate PCR fragments designed from a knowledge of the gene sequence. They complement conventional spotting of recombinant proteins, particularly where the latter are hard to express in cell based systems or un-cloned, and provide exciting new tools for high-throughput protein interaction analysis and functional proteomics.

APPENDIX: STEPS FOR PERFORMING PROTEIN *IN SITU* ARRAY

1. PCR construction
   (i) Generate target DNA construct by PCR (for a plasmid or cDNA template) or RT-PCR (for mRNA template) using designed primers and, where required, incorporating peptide tag.
   (ii) Assemble DNA fragments containing protein expression and termination elements at the 5' and 3' ends respectively of the target gene by overlapping PCR.
   (iii) Confirm construct identity by PCR mapping using primers at various positions along the desired sequence
2. Generation of protein *in situ* array
   (i) Set up a cell-free translation mixture according to the manufacturer's instruction.
   (ii) Apply the translation mixture to a precoated surface (wells, magnetic beads or glass slide). The volume can be varied between 0.1 and 25 µl.
   (iii) Incubate the mixture using the specific conditions for the particular cell-free system.
   (iv) Wash and analyze the arrayed proteins of interest.
3. Detection and analysis

Full methodological details can be found in He and Taussig.[4,24]

ACKNOWLEDGMENTS

We thank Hong Liu and Mike Bacon for technical assistance in the Technology Research Group. The development of the DAPA method is funded by the European

Commission Framework 6 Integrated Project 'MolTools' (www.moltools.org). Research at the Babraham Institute is supported by the BBSRC.

REFERENCES

1. Bertone, P. and Snyder, M., Advances in functional protein microarray technology, *FEBS J.,* 272, 5400, 2005.
2. Anderson, D.C. and Reilly, D.E., Production technologies for monoclonal antibodies and their fragments, *Curr. Opin. Biotechnol*., 15, 1, 2004.
3. Stevens, R.C., Design of high-throughput methods of protein production for structural biology, *Structure Fold. Des*., 8, R177, 2000
4. He, M. and Taussig, M.J., Single step generation of protein arrays from DNA by cell-free expression and *in situ* immobilization (PISA method), *Nucleic Acid. Res*., 29, e73, 2001
5. Michel, Y.M. et al., Eukaryotic initiation factor 4G-poly(A) binding protein interaction is required for poly(A) tail-mediated stimulation of picornavirus internal ribosome entry segment-driven translation but not for X-mediated stimulation of hepatitis C virus translation, *Mol. Cell. Biol*., 21, 4097, 2001.
6. Braun, P. et al., Proteome-scale purification of human proteins from bacteria, *Proc. Natl. Acad. Sci. USA,* 99, 2654, 2002.
7. Khan, F., He, M., and Taussig, M.J., A double-His tag with high affinity binding for protein immobilisation, purification, and detection on Ni-NTA surfaces, *Anal. Chem*., 78, 3072, 2006.
8. Keller, C. et al., Site-specific and temporally controlled initiation of DNA replication in a human cell-free system, *Nucleic Acids Res*., 30, 2114, 2001.
9. Landsverk, H.B. et al., Reprogrammed gene expression in a somatic cell-free extract, *EMBO Rep*., 3, 384, 2002.
10. Shimizu, Y. et al., Cell-free translation reconstituted with purified components, *Nat. Biotech*., 19, 751, 2001.
11. Kigawa, T. et al., Cell-free production and stable-isotope labeling of milligram quantities of proteins, *FEBS Lett*., 442, 15, 1999.
12. Madin, K. et al., A highly efficient and robust cell-free protein synthesis system prepared from wheat embryos: Plants apparently contain a suicide system directed at ribosomes, *Proc. Natl. Acad. Sci. USA*, 97, 559, 2000.
13. Jackson, A.M. et al., Cell-free protein synthesis for proteomics, *Brief. Funct. Genom. Proteom.*, 2, 308, 2004.
14. Ryabova, L.A. et al., Functional antibody production using cell-free translation: Effects of protein disulfide isomerase and chaperones, *Nat. Biotechnol*., 15, 79, 1997.
15. Noren, C.J. et al., A general method for site-specific incorporation of unnatural amino acids into proteins, *Science*, 244, 182, 1989.
16. Terpe, K., Overview of tag protein fusions: From molecular and biochemical fundamentals to commercial systems, *Appl. Microbiol. Biotechnol*., 60, 523, 2003.
17. Arnold, F.H., Metal-affinity separations: A new dimension in protein processing, *Biotechnology*, 9, 151, 1991.
18. Zhu, H. et al., Global analysis of protein activities using proteome chips, *Science*, 293, 2101, 2001.
19. Lue, R.Y.P. et al., Versatile protein biotinylation strategies for potential high-throughput proteomics, *J. Am. Chem. Soc*., 126, 1055, 2004.

20. Gite, S. et al., Ultrasensitive fluorescence-based detection of nascent proteins in gels, *Anal. Biochem*., 279, 218, 2000.
21. Rothschild, K. and Gite, S., tRNA-mediated protein engineering, *Curr. Opin. Biotechnol*., 10, 64, 1999.
22. Tabuchi, I., Next-generation protein-handling method: Puromycin analogue technology, *Biochem. Biophys. Res. Commun*., 305, 1, 2003.
23. Steinhauer C. et al., Improved affinity coupling for antibody microarrays: Engineering of double-$(His)_6$-tagged single framework recombinant antibody fragments, *Proteomics,* 6, 4227, 2006.
24. He, M. and Taussig, M.J., DiscernArray™ technology: A cell-free method for the generation of protein arrays from PCR DNA, *J. Immunol. Methods*, 274, 265, 2003.
25. Ranachandran, N. et al., Self-assembling protein microarrays, *Science*, 305, 86, 2004.
26. Kim T.W. et al., Cell-free synthesis and *in situ* isolation of recombinant proteins, *Protein Expr. Purif.*, 45, 249, 2006.

Section 3

Detection Methods for Protein Microarrays

9 Fluorescent Detection Methods for Protein Microarrays

Steven Roman and Scott Clarke

CONTENTS

INTRODUCTION

Fluorescence detection has a long history of use in biological assays, from fluorometric immunoassays to DNA sequencing, homogeneous FRET assays, microarrays, and a wide variety of cell-based assays. Because of the sensitivity and flexibility of fluorescent labels and advances in fluorescence detection methods, the pursuit of single molecule detection has mainly been based on fluorescence.[1] Fluorescence has been the dominant detection technology for DNA arrays[2,3] and, since protein arrays were initially developed from DNA arrays, much of the early work on protein arrays was based on DNA array methods and instrumentation. While a variety of detection methodologies have been applied to protein arrays,[4–6] the proven utility of fluorescence in protein-based biological assays and its dominance in DNA arrays have made it the leading detection method for protein arrays.[7]

There are several different types of protein arrays. Antibody arrays were among the earliest types of protein microarrays to be described in the literature.[8] These "capture" arrays are most typically used for multiplexed protein quantitation. Advances in high-throughput cloning and expression methods have led to "unbiased" functional protein arrays in which all or most open reading frames (ORFs) from an organism are cloned, expressed, purified, and arrayed. The first of such arrays was the yeast proteome array.[9] This volume focuses on non-antibody protein arrays, but given that such arrays — especially proteome arrays — are more or less in their infancy, our discussion of fluorescence detection methodologies will necessarily draw on methods from related fields, notably antibody and DNA arrays, that are likely to be applied to functional protein arrays.

The choice of a detection method is a crucial element of assay design and is based on the available reagents and instrumentation, the entity(ies) measured, and the goal of the measurement. Our discussion of fluorescence detection methods will be partly guided by this organizing principle: after brief overviews of the types of fluorescent labels commonly used, strategies for attaching fluorescent labels to biological molecules, and detection instrumentation, we will discuss options for designing fluorescence assays with functional protein microarrays according to the type of assay. Finally, we will consider future directions for fluorescence-based functional microarray experiments.

TYPES OF FLUORESCENT LABELS

Fluorescent compounds, whether natural or synthetic, have the property of undergoing electronic transitions induced by the absorption of energy (typically light energy) and resulting in the release of energy as light. Thus, the absorbed energy excites electrons in the fluorescent compound and light (usually of longer wavelength and lower energy relative to the excitation light; referred to as the Stokes shift) is emitted when the electrons relax back to ground state. While fluorescent compounds differ markedly in structure, they all have polycyclic aromatic ring systems favorable to absorbing light energy and undergoing the concomitant electron energy transitions.

Small Organic Dyes

A wide variety of small organic fluorescent dyes have been developed that differ in molecular and spectral characteristics, such as charge or hydrophobicity, photostability, extinction coefficient, quantum yield, fluorescence lifetime, and excitation and emission spectra. The small size of most organic fluorescent dyes can be an advantage over larger labels for protein labeling as it theoretically allows more dye to be conjugated per protein molecule, leading to increased specific brightness. Also, reactive functional groups on proteins that are inaccessible to larger labels may be more readily attacked by small dyes. A wider range of excitation/emission spectra is now available due, in part, to advances in solid state laser technology, providing access to an even greater variety of small organic dye properties.

When labeling proteins for microarray experiments, reactive fluorescent dyes should be evaluated on a number of key characteristics that are contributed by both the dye moiety and the reactive chemical group. Reactive dyes should form covalent bonds with the target proteins and have visible excitation and visible or near infrared emission to be compatible with existing detection instrumentation. They must also be reactive in aqueous solutions, have a high extinction coefficient (the efficiency with which the dye absorbs light energy), a relatively high quantum yield (the emission / absorption ratio, or how often an absorbed photon leads to a fluorescent event), and be as resistant as possible to photo-degradation. During protein labeling, the prevalence of unwanted side reactions must be kept to a minimum and the free unreacted dye should be easily deactivated and separated from the labeled proteins. Quenching through dye stacking or protein insolubility from labeling are characteristics that should be avoided. Furthermore, when labeling for microarray experiments, dyes must also have low non-specific binding to the array surface and blocking reagents. A summary of commonly used small organic fluorescent dyes organized by excitation spectra follows.

Ultraviolet (UV) and Violet Laser Excited Dyes (405 nm)

Very few microarray applications have used this region of the excitation spectrum due to limitations of commercial microarray scanners, although some commercial scanners now address this region. Also, most of these dyes emit at shorter wavelengths where auto-fluorescence of biological samples and some array substrates becomes a significant problem. Cascade Blue® and Alexa Fluor® 405 in

the pyrenes class and aminomethylcoumarin (AMCA) and Alexa Fluor 350 in the coumarin class of dyes are standards for this region of the excitation spectrum, with applications in flow cytometry and immunohistochemistry.[10] An example of using a UV/violet excited dye in a microarray assay is the work of Gosalia et al., in which coumarin-based fluorogenic peptides were arrayed for profiling protease specificity.[11]

Blue Laser Excited Dyes (488 nm)

Green emitting dyes excited by blue lasers were not used initially for microarrays, since blue lasers were not available on earlier scanners. These dyes have become more common with the advent of white light scanners and multi-laser scanners equipped with blue lasers. Fluorescein, Oregon Green®, and Alexa Fluor 488 are dyes appropriate for protein microarrays in this region of the excitation spectrum. Fluorescein has high quantum yield and good water solubility, but suffers from poor photo-stability, significant effects of pH on fluorescence, and dye quenching with increased degree of substitution (DOS), a phenomenon where, past a certain threshold, brightness decreases with increasing DOS. Oregon Green does not have the same DOS-related quenching or pH-dependent fluorescence as fluorescein and is therefore a preferred dye. Alexa Fluor 488 is brighter than fluorescein, matches the excitation and emission channels of fluorescein, is much less susceptible to photobleaching, and does not have DOS-related quenching or pH-dependent fluorescence.

Green Laser Excited Dyes (532 nm)

Tetramethylrhodamine (TMR) and carboxytetramethylrhodamine (TAMRA™) are chemically related species of rhodamine with similar spectral properties. They are photo-stable, but are prone to aggregation and suffer from DOS-related quenching. TMR is used in automated DNA sequencing, but has limited application on protein arrays. The rhodamine-like dyes Alexa Fluor 532 and Alexa Fluor 546, and the cyanine dyes Alexa Fluor 555; DyLight™ 547, and Cy™3 are standard dyes used on both DNA and protein microarrays. These dyes have high extinction coefficients and do not exhibit DOS-related quenching. Small but significant differences in structure confer differences in photo-stability and resistance to ozone degradation, with Alexa Fluor 555 being among the most stable.[12]

Orange Laser Excited Dyes (594 nm)

5-ROX™, Texas Red®, Lissamine™ rhodamine B, Alexa Fluor 594, and Cy3.5 have spectral properties intermediate between the green and red laser excited dyes. As such they have limitations for multiplexing because of the spectral overlap. However, dyes in this spectral range have been used successfully for array normalization in three or four color applications with DNA arrays.[13,14]

Red Laser Excited Dyes (633 nm)

Alexa Fluor 647, DyLight 647, and Cy5 are cyanine dyes with similar spectral properties and brightness characterized by very high extinction coefficients. As with the green laser excited cyanine dyes, small differences in structure confer significant differences in solubility, photo-stability, and ozone susceptibility, with Alexa Fluor 647 being the more stable. These dyes are among the most commonly used organic dyes for labeling proteins and DNA.

Far Red Laser Excited Dyes (650, 687 nm)

Alexa Fluor 660, Alexa Fluor 680, Alexa Fluor 700, Alexa Fluor 750, Cy5.5, and Cy7 represent the next generation of organic dyes with significant spectral separation from the red laser excited dyes. They are pH insensitive between pH 4 and pH 10, have varying degrees of photo-stability, but are susceptible to ozone degradation. The cyanine dyes in this group (Alexa Fluor 750 and Cy7) have high extinction coefficients making them particularly bright. Microarray instrumentation is not optimized for use with these near IR emitting dyes, as photomultiplier tubes (PMTs) in standard detectors are insensitive to wavelengths greater than 850 nm.

The BODIPY® dyes represent another class of dyes that offers a variety of substitutions resulting in emission maxima ranging from green to red. BODIPY dyes have high extinction coefficients and quantum yields, are non-charged, and are commercially available as amine reactive dyes. But, due to difficulty in obtaining good reaction efficiency in the aqueous labeling environments of protein extracts, they have found only limited application on protein microarrays. In one example, direct incorporation of BODIPY-FL-lysine into *in vitro* translation products was shown by dot blot detection,[15] demonstrating that ribosomes will utilize the fluorescent analog. Such an approach may be viable for protein array probe preparation. Membrane fraction experiments on microarrays may be an ideal application for BODIPY dyes, since these dyes are well-suited to hydrophobic environments and are currently used for lipid labeling.

Fluorescent Proteins

Naturally occurring fluorescent proteins and their engineered derivatives have found widespread use in biological applications.[16,17] The forerunner of the fluorescent proteins, green fluorescent protein (GFP), has a large Stokes shift with excitation in the blue region and emission in the standard fluorescein green. Engineered and mutant versions of natural fluorescent proteins have been developed that provide a wider choice in excitation and emission properties and differing extinction coefficients and quantum yields.[18] Fluorescent protein fusions have been used extensively in cellular studies, but the larger size of this class of fluorescent label makes it less suitable for some types of interaction studies (the family of GFP-related proteins has a molecular weight of ~30 kDa). Some fluorescent proteins, such as DsRed, only fluoresce as multimers and have unpredictable fluorescence responses as fusion proteins. Engineered monomeric versions of some of these proteins have become available that are more suited to protein expression studies. Use of recombinant green and red fluorescent protein fusions has been demonstrated on protein microarrays.[19]

R-phycoerythrin, is a 240 kDa fluorescent protein, derived from alga, with a broad excitation spectrum including several peaks (a minor blue laser excitation peak at 488 nm and a major green laser excitation peak at 532 nm) and emission compatible with the standard green channel filters of Alexa Fluor 555 and Cy3. R-phycoerythrin is very bright with an extinction coefficient of $1.96 \times 10^6\ M^{-1}\text{cm}^{-1}$ and a quantum yield of 0.68. The typical application for this fluorescent protein is

to conjugate it with streptavidin, anti-hapten, or anti-IgG antibody, and use it for secondary detection on DNA or protein arrays whose probes have been biotin- or hapten-labeled.

Lanthanide Time-Resolved Fluorescence

Lanthanide time-resolved fluorescence (TRF) has not often been used with protein arrays, due mostly to the lack of commercial TRF scanners.[20,21] However, adaptation of TRF microscopic imaging to microarrays could bring TRF to the forefront for array detection. TRF has the advantage of greatly reducing background fluorescence originating from non-lanthanide sources. Since lanthanide materials fluoresce over relatively long periods, signal is collected after flash excitation and a lag time, during which non-lanthanide sources of background fluorescence have decayed. Fluorescent lanthanide chelates exhibit a large Stokes shift, with excitation in the UV range and emission greater than 500 nm. The fluorescent emission peak profiles are also quite sharp, with half-widths being 10 nm to 20 nm. Commonly available lanthanides and their fluorescence characteristics appear in Table 9.1. Lanthanide labeling kits for proteins are commercially available that use standard succinimidyl or maleimide chemistry (see Labeling Methods section below) to covalently attach a lanthanide-metal chelate to proteins. The same issues of labeling with reactive fluorescent dyes (see Labeling Methods below) apply to this type of fluorescent tag.

There are several ways in which lanthanide labels are used in bioassays. The DELFIA® (Dissociation-Enhanced Lanthanide FlouroImmunoAssay, PerkinElmer) system induces fluorescence with an enhancement solution that dissociates lanthanides from lanthanide chelate-labeled proteins and re-chelates them in a highly fluorescent form. The LanthaScreen™ system (Invitrogen) uses Terbium chelates and fluorescein to produce time-resolved fluorescence resonance energy transfer (TR-FRET) where the lanthanide chelate acts as donor and fluorescein acts as acceptor. Terbium-labeled antibodies and fluorescein-labeled substrates are available for a wide variety of kinase assays. The HTRF® (Homogeneous Time-Resolved Fluorescence, Cisbio International) system uses Europium chelates and proprietary acceptors to enable a variety of homogeneous solution assays. The LANCE system (PerkinElmer) uses Europium chelate as donor and allophycocyanin as acceptor in TR-FRET assays. The TR-FRET and HTRF formats may be the most immediately applicable to protein arrays, since these systems use reagents that will remain fixed in addressable locations

TABLE 9.1
Commonly Available Lanthanide Chelates for Time-Resolved Fluorescence

Chelate	Emission (nm)	Excitation (nm)
Europium (Eu)	615	340
Samarium (Sm)	642	340
Terbium (Tb)	545	300

on an array surface. DELFIA may be adaptable to protein arrays, but is currently a solution phase system that releases fluorescent chelates into solution and is best suited to microplate assays and non-image-based readers.

Quantum Dots and Fluorescent Nanoparticles

Quantum dots (Qdots) are fluorescent nanometer-sized particles of semiconductor material that have broad excitation and narrow emission spectra. Emission maxima are governed by size and composition of the particles.[22] Thus, different Qdots can be excited by the same source and will fluoresce in different colors. Qdots typically consist of a cadmium selenide core surrounded by passivating and functionalizing layers. These outer layers provide relative biological inertness, photo-stability, and a substrate for conjugation to biologically active molecules. Qdots of various fluorescent emissions are commercially available as conjugates to primary or secondary detection proteins, or as carboxylate or amine species that can be activated and covalently linked to the protein(s) of interest. For the carboxylate form, the carboxyl group is activated with N-ethyl-N'-dimethylaminopropyl-carbodiimide (EDAC), followed by reaction with amines on the target protein. For the amine form, the amino group is activated with bis(sulfosuccinimidyl) suberate (BS3), a homobifunctional crosslinker, followed by reaction with amines on the target protein.

The application of Qdots to proteins on arrays provides a means of multiplexing using a single chemistry and functionality. Furthermore, a single UV or visible excitation can be used in multi-colored assays. However, the passivating layer results in Qdots having a much larger diameter than the semiconducting core. As a consequence of the large size, typical conjugations result in numerous proteins bound to a single Qdot. The resulting multi-valency may be advantageous or deleterious, depending on the application. In practice Qdots are usually used as secondary antibody or anti-hapten conjugates where multi-valency will enhance sensitivity with no effect on primary binding reactions.

Other types of fluorescent nanoparticles have been used for detection on microarrays. Dye-doped nanoparticles encased in a silica matrix have a large number of fluorophores per particle. As such they are bright and photo-stable because the silica matrix shields the dyes from degradation by ozone or oxygen.[23] Zhou and Zhou[24] used Cy3- and Cy5-doped nanoparticles on DNA arrays in a two color sandwich type assay. These particles were functionalized on their outer silica layer with thiols for conjugation to specific oligonucleotide probes. Lian et al. [23] used a variety of silica surface modifications for bioconjugation and tested their dye-doped nanoparticles in 96-well plate assays, immunohistochemistry, and immunocytochemisty, as well as DNA and protein arrays.

Amplification

Anti-hapten Dye Conjugates

Signal from hapten-labeled proteins can be amplified using anti-hapten antibody-dye conjugates. Fluorescein, biotin and its analogs, digoxigenin, and dinitrophenol are frequently used haptens that can be amplified using labeled anti-hapten antibodies

or, in the case of biotin, labeled streptavidin. Fluorescein has been used as a hapten in this context with fluorescent anti-fluorescein antibodies as the signal amplification step. Fluorescein-labeled protein signal can be effectively amplified with either Oregon Green- or Alexa Fluor 488-anti-fluorescein conjugates, as these dyes have the same excitation and emission spectra as fluorescein. The resulting complexes (Figure 9.1A) exhibit a signal increase of as much as 100-fold.[25] Amplified signal from biotin-, desthiobiotin-, and dinitrophenyl-biotin-labeled proteins can each be generated in a similar manner using either anti-biotin-dye or streptavidin-dye conjugates. Further amplification can be achieved by a subsequent round of biotin-anti-Ig (or anti-SA) followed by another application of labeled anti-biotin (Figure 9.1B).

Enzyme-Linked Fluorescence (ELF)

Proteins labeled with biotin or other haptens can be detected with alkaline phosphatase-anti-hapten antibody conjugates or alkaline phosphatase-streptavidin conjugates using alkaline phosphatase reagent and the fluorogenic ELF-97 phosphate substrate (2-(5′-chloro-2′-phosphoryloxyphenyl)-6-chloro-4-(3H)-quinazoline, Molecular Probes). Dephosphorylation converts ELF-97 phosphate to ELF-97 alcohol, a water insoluble fluorescent product that has UV excitation and orange emission. The insoluble product precipitates *in situ*, but, since it is not covalently linked to the local environment, signal resolution and quantitation may be compromised. The application of ELF-97 to microarrays is currently limited by a lack of UV lasers in commercial scanners.

Tyramide Signal Amplification (TSA)

TSA exploits the enzymatic conversion of tyramide to activated tyramide that forms covalent bonds with nearby protein tyrosine residues.[26] For signal amplification, tyramide-dye or tyramide-hapten conjugates are used. Localized, specific signal amplification is achieved by conjugating the enzyme that effects the tyramide conversion, horseradish peroxidase (HRP), to a secondary detection reagent, such as streptavidin (SA) or anti-IgG. Biotinylated or antibody-bound proteins of interest are reacted with HRP conjugate, and the protein-HRP complex is then exposed to tyramide and H_2O_2. When the tyramide is conjugated to a fluorophore, direct detection of deposited fluorescent tyramide is possible (Figure 9.1C). Tyramide-dye conjugates are available in a range of colors from blue to near IR emission. When the tyramide is conjugated to a hapten, an additional step with fluorescent anti-hapten is required for detection.

Rolling Circle Amplification (RCA)

Isothermal rolling circle amplification[27] produces a linear product consisting of hundreds of tandem copies generated from a circular DNA template. The tandem repeat product is then hybridized to a small labeled probe such that many copies of labeled probe bind to each RCA product, resulting in a very bright, amplified signal. If the product is tethered to a surface (such as an array), the signal remains localized. RCA has been adapted to protein arrays and has been used in one color microarray sandwich immunoassays[28–30] where detection antibodies are conjugated to a template-complementary oligo (Figure 9.1D). Two color applications (Figure 9.1E) have also been developed where crude protein samples are labeled

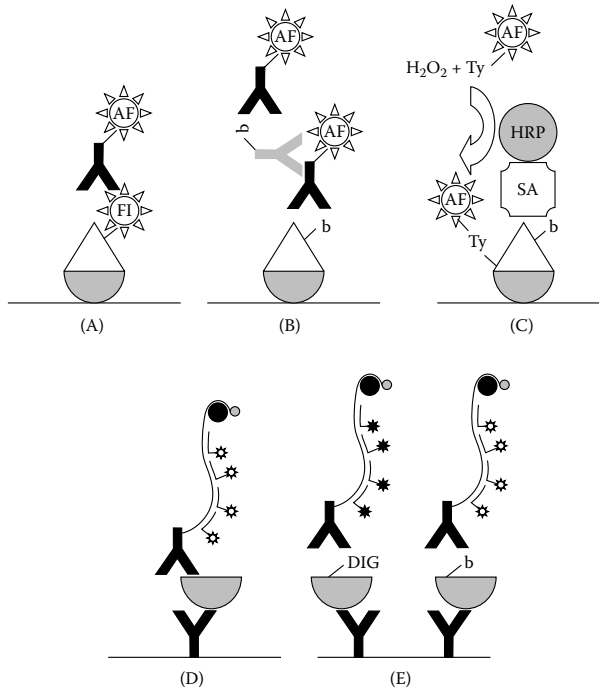


FIGURE 9.1 Selected Amplification Methods. (**A**) Signal from fluorescein (Fl)-labeled probe protein (triangle) bound to arrayed target protein (gray semicircle) is amplified with Alexa Fluor 488 (AF)-labeled anti-fluorescein antibody. Signals from Alexa Fluor 488 and fluorescein are additive since they share the same excitation and emission channels. For simplicity, only single labels are shown. In practice, proteins often bear multiple labels, leading to multiple binding events and greatly enhanced signals. (**B**) Biotinylated (b) probe protein (triangle) bound to arrayed target protein (gray semicircle) is detected with Alexa Fluor-labeled anti-biotin, then biotinylated anti-IgG, then a second application of Alexa Fluor-labeled anti-biotin. Again, for simplicity, only single labels are shown. (**C**) Tyramide signal amplification can be achieved by binding streptavidin (SA)-horseradish peroxidase (HRP) conjugate to target-bound biotinylated probe. Alexa Fluor-labeled tyramide (Ty-AF) and H_2O_2 are added, resulting in the HRP-catalyzed formation of activated Ty-AF, which reacts with nearby tyrosine residues. Amplified fluorescent signal is thus covalently attached to target and target-bound proteins. (**D**) Immuno-Rolling Circle Amplification (Immuno-RCA) is achieved by first forming a traditional antibody-antigen (grey semicircle)-antibody immune "sandwich." The secondary antibody is conjugated with an oligonucleotide complementary to a circular template (black circle). Isothermal replication of the circular template with an appropriate DNA polymerase (small grey circle) results in many tandem repeats of copied DNA ligated to the original oligonucleotide on the secondary antibody. Detection is achieved by addition of labeled short oligonucleotides complementary to the tandem repeat sequences. (**E**) Two color ImmunoRCA is achieved essentially as in (**C**), but using distinct template sequences on separate antibodies directed against different haptens, in this illustration, biotin (b) and digoxigenin (DIG). Thus, the same probe protein (semicircles) from different protein preparations, one labeled with digoxigenin and the other with biotin, can be differentially detected with different color fluorescent short oligonucleotide probes on the same array spot. As in other two color methods, the ratio of color1 signal to color2 signal indicates the expression ratio of probe protein in the two samples.

with digoxigenin (DIG) and biotin and bound to antibody arrays, followed by binding of template complementary anti-DIG and anti-biotin conjugated antibodies, and then by isothermal RCA and labeled probe hybridization.[31–33] Despite the many steps involved, RCA has proven to be a very sensitive amplification method.

LABELING METHODS

Modifying proteins with dyes or haptens carries with it the risk of affecting the characteristics of the protein of interest. Labeling can affect hydrophobicity, charge, or solubility, and may interfere with sites critical for protein functions or interactions. Another consideration with dye or hapten labeling of complex mixtures of proteins is that heterogeneous labeling of the proteins in the mixture is almost inevitable. Labeling bias, where some proteins are under-labeled while other proteins are over-labeled, leads to signal artifacts on arrays. Regional differences in the reactivity and accessibility of the functional groups on proteins are largely responsible for labeling artifacts.[34] Another potential problem is the loss of labeled protein due to precipitation during the dye removal step. Over-labeling can result in dye stacking and formation of hydrophobic pockets that reduce the solubility of proteins. In spite of these limitations, dye or hapten labeling is the predominant method for detecting proteins on functional protein microarrays *in situations* where high affinity secondary binding reagents are not available.

Amine Modification

Succinimidyl Esters

The amine reactive succinimidyl ester (SE) cyanine dyes are currently the most common choice for covalent protein labeling and are commonly available as the *N*-hydroxysuccinimidyl (NHS) ester form. These dyes are supplied as lyophilized powders and are stored dry at –20°C until use to prevent the unwanted hydrolysis side reaction. The labeling reaction proceeds rapidly in an aqueous environment at pH 8.6 with a half life of 10 minutes at 4°C.[34,35]

Tetrafluorophenyl Esters

Tetrafluorophenyl esters (TFPs) form the same covalent amide link as succinimidyl esters, but TFP reactive dyes have improved coupling efficiency due to a reduced susceptibility to hydrolysis.[25] TFP forms of some Alexa dyes (Molecular Probes) and biotin (Pierce Chemical Co.) are commercially available.

Sulfonyl Chlorides

Under alkaline conditions (typically pH 9 to 10), sulfonyl chlorides create a stable sulfonamide bond with the ε amine group of lysine in proteins. Sulfonyl chlorides are fairly sensitive to hydrolysis and must be stored desiccated to prevent breakdown. In aqueous environment at pH 8.3 Texas Red sulfonyl chloride was shown to be completely hydrolyzed in 2 to 3 minutes.[25] Sulfonyl chloride conjugates also react readily with DMSO, so DMSO should not be used for dissolving the reactive dye. The high pH of this reaction is not compatible with the stability of some proteins.

Isothiocyanates

Isothiocyanates react with ε and N-terminal amines to form relatively stable thiourea groups.[34] Although not as stable and often not as bright as the succinimidyl ester-dyes, they are still widely used in the form of fluorescein isothiocyanates (FITC) and tetramethylrhodamine isothiocyanates (TRITC) for modifying proteins. Since the active group is relatively unstable in aqueous environment, these reagents should be stored desiccated and frozen or refrigerated.

Thiol Modification

Maleimides

Covalent coupling through sulfhydryls is the second most important method of coupling small organic fluorophores to proteins. Because lysines for amine labeling are more abundant in proteins, they generally can't be used for precise site-specific dye attachment. Site-specific labeling through less abundant cysteines can be achieved with thiol reactive dyes. In principle the reaction with reduced cysteine thiols on the protein of interest proceeds very quickly and efficiently at neutral pH. Maleimides can undergo a ring-opening side reaction that destroys the sulfhydryl reactivity and this reaction may even happen after sulfhydryl coupling.[34] The ring-opening reaction typically becomes more prevalent at higher pH and different maleimides can differ significantly in their susceptibility.[34]

Antibodies are often labeled with maleimide chemistry, since the immunoglobulin G hinge cysteine can be conveniently reduced without subunit dissociation through the use of tris-(2-carboxyethyl) phosphine hydrochloride (TCEP) or dithiothreitol (DTT). The additional step of reducing agent removal is required prior to labeling, but beadimmobilized reducing agents, which can be easily and rapidly separated from the protein sample, can be used in place of soluble reducing agents.[34] In general, treatment of proteins with reducing agents may cause protein subunits to dissociate, thereby altering the interactions expected on a functional array. Use of TCEP over DTT can selectively reduce thiols in aqueous microenvironments while avoiding hydrophobic core thiols, thus minimizing the dissociation of proteins often seen with stronger reduction agents.[25] Reaction of maleimide with reduced antibodies is essentially complete within five minutes. Re-oxidation of the reduced thiol of the hinge region is negligible during this process, even though many protocols recommend nitrogen sparging of buffers prior to reaction.

Iodoacetamide

Iodoacetamides react readily with thiols to form thioethers. Iodoacetamides are handled the same as maleimides, with the additional complication that they are light sensitive, so reactions should be performed in the dark. Iodoacetamides are not susceptible to the ring opening hydrolysis reaction of maleimides, are less prone to water hydrolysis, and show the same degree of site-specific labeling as maleimides. Free thiols in solution will readily react with iodacetamide and quench the reaction with protein thiols. In a study on a fragment of Wiskott Aldrich Syndrome Protein with an introduced cysteine residue, labeling with an iodoacetamide derivative of Cy3 showed improved water solubility and reduced side reactions over the maleimide-Cy3.[36]

Carbonyl Modification

Carbonyl groups include carboxylic acids, aldehydes, amides, ketones, and esters. Some (carboxylic acids and amides) are naturally occurring in proteins, while others (aldehydes, ketones, and esters) can be formed by chemical modification. The carbon-oxygen double bond of carbonyl groups, in which the oxygen is more electronegative, presents a more electrophilic carbonyl carbon that is prone to attack by nucleophiles.

Carboxylate Modification

Carbodiimides

Carboxylates, present in proteins as free C-termini and on aspartic and glutamic acid side chains, can be activated using carbodiimide chemistry for subsequent coupling to amine-containing compounds. Thus, primary amine-modified fluorophores can be covalently bound to proteins following treatment of the proteins with carbodiimide 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDC).[34] Commercially available cadaverine-dye modified Cascade Blue, Texas Red, and Alexa dyes contain a primary amine side chain that can be coupled to proteins in this fashion. Carbodiimide catalyzed coupling provides an alternative choice of functional group for covalent protein modification. However, carbodiimide chemistry consumes the negative charge of carboxylate groups on the target protein, which, depending of the charge of the attached dye, could alter the net charge of the target protein.

Aldehyde and Ketone Modification

Aldehydes and ketones are reactive toward hydrazides, hydroxylamines, or amines. These reactive groups are not normally present on proteins and so must be introduced by chemical modification. For instance, aldehydes can be generated by periodate oxidation of carbohydrates on glycoproteins. Reaction of aldehydes and ketones with amines form relatively labile Schiff base compounds that must be subsequently reduced.

Hydrazide

Dye- or biotin-hydrazide conjugates can be used to covalently label aldehydes and ketones. Chen et al. demonstrated enzymatic ketone introduction and subsequent hydrazide labeling[37] of proteins. They labeled an acceptor peptide-tagged recombinant cell surface protein with a ketone derivative of biotin using biotin ligase (BirA), followed by hydrazide-hapten conjugation.

Alcohol Modification

Non-acylated N-terminal serine and threonine residues can be oxidized with periodate to form aldehydes that can subsequently be modified with dye-hydrazide, -hydroxylamine, or -amine derivatives. This method has been applied to peptide labeling.[38]

Carbohydrate Modification

Biotin-hydrazides and dye-conjugated hydrazides have been used to label carbohydrates of glycoproteins. Prior to hydrazide reaction, carbohydrates must be modified with periodate to generate aldehydes. Alternative methods of aldehyde formation using specific sugar oxidases can be used in place of periodate treatment.[34] Once introduced, aldehyde groups on proteins react directly with hydrazide as described above (Aldehyde and Ketone Modification).

Photo-reactive Chemical Reagents

Photo-reactive fluorescent cross-linking reagents currently have a minor role in labeling for protein microarrays, but could be developed into useful tools for protein-protein interactions on arrays. Sulfosuccinimidyl 2-(7-azido-4-methylcoumarin-3-acetamide) ethyl-1,3'-dithioporpionate (SAED) is a heterobifunctional fluorogenic cross-linker that has an active NHS ester group at one end for covalent attachment to protein amines and a UV light sensitive azide group at the other end that reacts with adjacent nucleophiles. These types of compounds have been used for mass spectroscopy analysis but not yet on protein arrays.

A strategy for demonstrating protein-protein interactions with photo-reactive cross-linkers makes use of benzophenone-4 maleimide to stabilize weaker interactions. This strategy could be applied to protein-protein interaction on protein arrays by first reacting benzophenone-4 maleimide with a fluorescent probe protein through the sulfhydryl reactive maleimide. After incubation of the benzophenone-modified probe on the array, UV light would be used to cross-link the probe to its target(s) via the photo-reactive benzophenone group. Unreacted species can be repeatedly photo-activated for subsequent coupling reactions.[34]

Indirect Labeling with Biotin and Other Haptens

The chemistries that are used to directly label with fluorescent dyes can be used for the covalent attachment of biotin (and its derivatives), digoxigenin, and dinitrophenol (DNP), as well as other haptens. Most hapten labeling reagents are available with several linker lengths designed to improve accessibility to binding by secondary detection reagents. Quantitation of the degree of substitution with haptenylation reagents can require specialized assays. Biotin quantitation has traditionally been done by a chromogenic assay using 2-hydroxyazobenzene-4′-carboxylic acid (HABA), but the assay is relatively insensitive and consumes large amounts of conjugate. A sensitive fluorescence assay for biotinylation that consumes very little conjugate has recently come on the market (Invitrogen/Molecular Probes). DNP labeling can be quantified by absorbance of the DNP chromophore. Labeling with other haptens is best determined through the use of fluorophore-hapten conjugates. Detection of haptenylated probe proteins on arrays is accomplished by using appropriate fluorescent anti-hapten reagents.

ULS™ (Cisplatin) Labeling

The DNA modifying property of cisplatins has made them useful as cancer therapies. Cisplatins bind to the N7 of guanine residues. Zhang et al. used a photo-cross-linking cisplatin-modified DNA to identify proteins involved in recognition and repair of DNA.[39] Cisplatins can also be used to modify proteins, where they react with the amino acids methionine, cysteine, and histidine. Cisplatin-based protein and DNA labeling reagents have been developed and commercialized by Kreatech Biotechnology (Amsterdam) as the Universal Labeling System (ULS™), but current protocols for protein labeling require 3- to 4-hour incubation at 37°C and the reagent is sensitive to competing thiols.

Staudinger Ligation

The Staudinger ligation is a reaction between a phosphine and an azide that, through an unstable intermediate, forms a primary amine and a phosphine oxide.[40] Variations on the Staudinger ligation have been developed to stabilize the intermediate so that the azide forms a stable amide bond with the phosphine, thus covalently linking the phosphine reagent to the azide-bearing entity. Azido groups can be chemically introduced on proteins with amine-reactive azido-NHS-esters. One potential advantage of this approach for functional protein array probe preparation is the small size of the azido group should be minimally disruptive to protein function while being easily modified with dye-phosphines to generate a fluorescent signal after incubation on an array. Another method of introducing Staudinger-modifiable azido groups on proteins is through the use of non-natural azido-amino acids for recombinant protein expression.[41] Metabolic incorporation of azido groups on glycoproteins has been accomplished using azido-modified glycosylation precursors,[40,42–44] enabling subsequent specific modification of the azido-glycoproteins by Staudinger ligation with phosphine probes. Fluorogenic phosphine reagents have also been developed for direct incorporation of a dye by the Staudinger ligation.[45]

Two-Color Applications

Two-color dye labeling for protein arrays requires that the dyes have spectral separation, similar reactivity towards proteins (minimal labeling bias), similar brightness, and equally low non-specific binding to the blocked array substrate. Two color labeling experiments with amine-reactive cyanine dyes on functional arrays show populations of proteins that have a shifted preference toward one dye or the other. Thus, in a homotypic labeling and probing study in which the same sample is labeled with two different dyes, there may be protein signals that will be weighted toward one color over the other (authors' unpublished observations). For two color applications the average DOS of each labeled solution should be matched. In practice, because of differences in labeling efficiencies between dyes, such label matching requires several labeling reactions at slightly different molar challenges. As with two color gene expression experiments on DNA arrays, normalization methods are generally employed to compensate for various types of bias (reviewed in [46]).

Recombinant Fusion Tags

Protein fusion tags have become a mainstay of recombinant protein technology, mainly for convenient purification and monitoring of expressed recombinant proteins.[47] Expression vectors for N- and C-terminal fusions are commercially available, making this an attractive method for studying protein characteristics. Such tags range from short stretches of amino acids to larger fusions with entire proteins. The growing number of commercially available tag and fluorescent protein fusion vectors and fluorescent anti-tag detection reagents provide many options for labeling a protein of interest for direct as well as indirect fluorescent detection on protein arrays.

The choice of fusion tag depends on the protein of interest, the expression system, and the purpose for the fusion. Fusions with green fluorescent protein (GFP) are

TABLE 9.2
Fusion Tags for Recombinant Protein Expression

Tag	Sequence/Size	References
Polyhistidine (6XHis)	HHHHHH	[119,120]
FLAG™ (T7 gene 10 leader sequence)	DYKDDDDKG	[121,122]
HA (from influenza hemagglutinin)	YPYDVPDYA	[123]
V5 (from simian virus 5 RNA pol α subunit)	GKPIPNPLLGLDST	[124]
c-*myc* (from murine c-*myc* oncogene)	EQKLISEEDL	[125]
BioEase™ (from *K. pneumoniae*, for *in vivo* biotinylation)	72 aa (7.1 kDa)	[126]
GST (Glutathione-S-tranferase)	223 aa (26 kDa)	[127]
FlAsH, etc. (biarsenical dyes that form covalent complexes with tetra cysteine motif)	CCPGCC	[16,17,128]
Fluorescent proteins (GFP, RFP, etc.)	GFP = 238 aa, 27 kDa	[17,129]

appropriate for *in situ* cellular studies, as recombinant products can be directly visualized in the cell. Smaller "epitope" tags, like V5, HA, FLAG™, and c-*myc*, are well suited to *in vitro* assays using available anti-epitope reagents. Epitope tags can often be combined such that one tag is used for purification while a second tag is used for an *in vitro* assay. Epitope tags used for purification can be engineered with protease cleavage sites so that the tag can be removed after purification. No single tag will fulfill all requirements for all proteins, and the selection of a tag or tags is empirical. The use of recombinant tags carries with it the risk that the fusion protein will be adversely affected by incorporation of non-native sequence elements.[48] On the other hand, a tag may confer increased solubility and/or expression of a recombinant protein, particularly in a heterologous expression system.[49] Table 9.2 lists some available fusion protein tags, but is not meant to be an exhaustive catalog (for more in-depth reviews, see [47,49]).

Nucleic Acid Labeling

Nucleic acids labeling is relevant to functional protein array experiments when DNA or RNA is used as the probing species, as in, for example, experiments to identify DNA binding proteins. There is a large body of literature on nucleic acid probe preparation and labeling for microarray experiments, a complete review of which is beyond the scope of this chapter. The reader is referred to various reviews, books, and websites with in-depth discussions of methods of nucleic acids probe preparation.[50–57]

In Vitro Protein Expression

In vitro translation is a convenient way to express proteins that may be otherwise difficult to obtain by *in vivo* methods. *In vitro* translation can be coupled to *in vitro* transcription and there are several commercial suppliers of *in vitro* transcription and translation (IVTT) kits, making this a flexible and convenient approach to protein production. Protein products from IVTT reactions have traditionally been labeled with ^{35}S-methionine, but

more recently, several fluorescent options have become available. Recombinant proteins expressed using *in vitro* translation systems can be tagged at their N- or C-termini with specific epitopes (reviewed above) or the tetra-cysteine moiety used for biarsenical dye labeling. These approaches require use of a fluorescent anti-epitope secondary antibody or the biarsenical fluorescent dye that binds to the tetra-cysteine tag. Methods of *in vitro* transcription and translation are covered in more detail in Chapter 3.

Several groups have used *in vitro* methods to synthesize and label proteins with either haptens or fluorescent dyes. Kawahashi et al.[58] prepared probe proteins for array experiments by cell-free translation using a fluorescent puromycin derivative that labeled proteins at their C-termini. In a similar approach, Tan et al.[59] used a biotinylated puromycin derivative to label cell-free synthesized proteins at their C-termini with biotin. Taki et al.[60] have developed a biotinylated tRNA(fmet) that can be used to label *in vitro* translation protein products at their N-termini. Coleman et al. used a BODIPY-FL-lysine-charged tRNA to label and monitor the expression of *in vitro* translation products.[15] These *in vitro* synthesis and labeling methods hold promise for more homogeneous labeling of proteins for array probing studies.

Non-covalent Methods

Use of stains for detection of protein binding events after the initial array probing step has potential advantages over covalent attachment of dyes in that protein-protein interactions are formed prior to modification by the detection reagent. However, some staining protocols require fixation of the proteins or alteration of the buffer conditions to maximize detection sensitivity, which are treatments that could introduce their own artifacts by interrupting some types of protein interactions. Examples of several stains specific for certain protein modifications or sequences follow.

Phosphoprotein Stain

Pro-Q Diamond (Molecular Probes), a phospho-protein specific dye, has been used successfully to stain phospho-peptides and phospho-proteins on array surfaces.[61,62] Pro-Q Diamond spectral properties fit the standard green laser line of 532 nm excitation and 580 nm emission. Use of this stain requires a mildly acidic buffer for fixation and dye binding.

Polyhistidine Stain

Staining for the polyhistidine epitope tag could provide a means for detecting the interaction of his-tagged recombinant protein probes on arrays without the potential interference of covalently modifying the probe with dyes. Pro-Q Sapphire (Molecular Probes) is a polyhistidine-specific dye with spectral properties that fit the standard blue laser line of excitation and 488 nm and emission at 515 nm.[63] Polyhistidine-tagged proteins currently being probed by surface plasmon resonance arrays could be validated on a fluorescence stain platform.

INSTRUMENTATION

Fluorescent microarray experiments require sophisticated instrumentation for data acquisition. The basic functions of the instrumentation are to excite the fluorophores on the array surface, to collect the emitted light from excited fluorophores, and to

convert the collected light signal into a digital image. Fluorescent microarray scanners currently available employ one of two basic methods to generate image data: laser excitation with a photomultiplier tube (PMT) detector, or white light excitation (usually filtered) with a cooled charge-coupled device (CCD) detector. Laser-PMT systems can be further divided into those with a confocal optical path and those with a non-confocal path. Confocal systems employ a second focusing lens that helps screen out noise arising from areas outside the focal plane. For multiple color applications, laser-PMT systems often have multiple lasers and PMTs. Systems capable of up to four colors are available. CCD-based systems, of course, do not require multiple excitation sources, but usually employ appropriate filter sets for excitation and emission. Each type of system has its strengths and weaknesses.[3,64] Software for analysis is usually bundled with instrumentation, although there are some stand-alone analysis packages. Table 9.3 lists some of the major microarray instrumentation suppliers.

Most commercial scanners excite fluorophores on the array surface by directing light (filtered white light or tuned laser) onto the surface of the array, usually through the use of a beam splitter that discriminates between excitation and emission wavelengths. The Zeptosens evanescent waveguide technology is unique in that it directs light into a specialized array substrate (e.g., a thin film of high refractive index material coated onto a glass microscope slide), creating a strong electromagnetic excitation field within the substrate that excites fluorophores close to the array surface. Fluorescent waveguide detection is reported to be more sensitive than epi-illumination or confocal systems due to suppression of background and efficient excitation of

TABLE 9.3
Commercial Suppliers of Microarray Scanners

Company	Scanner Type	Web Site
Affymetrix	Confocal laser scanner/PMT (proprietary array format)	www.affymetrix.com
Agilent	Laser scanner/PMT	www.home.agilent.com
AlphaInnotech	AlphaScan: Confocal laser scanner/PMT NovaRay: white light, CCD (slides and microplates)	www.alphainnotech.com
Applied Precision, LLC	White light, CCD	www.api.com
Axon Instruments	Laser scanners/PMT	www.moleculardevices.com
Biomedical Photometrics	Confocal laser scanners/PMT	www.confocal.com
Bio-Rad	Confocal laser scanner/PMT	www.biorad.com
Genetix	Confocal laser scanner/PMT	www.genetix.com
Illumina	Confocal laser scanners/PMT (proprietary array formats)	www.illumina.com
PerkinElmer	Confocal laser scanners/PMT	www.perkinelmer.com
Tecan	Confocal laser scanner/PMT	www.tecan.com
Zeptosens	Planar waveguide imager/CCD (proprietary array formats)	www.zeptosens.com

Note: All scanners take 1" × 3" microscope slides unless noted.

fluorophores. In addition, arrays can be read wet, allowing real-time data collection and making kinetic assays possible.[65]

Table 9.3 lists scanners for planar microarrays, but functional protein arrays can also be performed in solution on coded beads, such as those developed by Luminex (www.luminex.com). Bead array assays use a different instrumentation that, besides excitation sources (lasers) and emission collection (PMTs), also require fluidic components for handling beads so that they can be discretely decoded and analyzed in the instrument. Such assays are generally performed in 96-well microplates and beads are aspirated into the reader. Software for analysis is bundled with the reader.

FLUORESCENT DETECTION METHODS BY ASSAY TYPE

In the post-genomic era, focus is increasingly turning to the proteome and the immense variety of protein reactions and interactions that underlie cellular processes. Protein arrays are a significant tool for proteome research and in this section we will discuss fluorescent detection methods as they apply to specific assays using functional protein microarrays.

The types of assays envisioned for functional arrays can be divided into two main classes: binding assays and activity assays. Binding assays consist of protein-protein (including protein-Ab, a special case of protein-protein interaction), protein-nucleic acid, protein-small molecule (e.g., drug compound), protein-lipid, and protein-carbohydrate interactions. Fluorescent methods are available or are being developed for most of these applications. Activity assays on functional protein arrays will likely span the breadth of protein enzymatic activities cataloged in cells, whether the activity involves other proteins, nucleic acids, lipids, or other molecules of biological interest. Currently, activities such as phosphorylation, de-phosphorylation, proteolysis, and transcription are among those receiving the most attention on a proteomic scale and are amenable to investigation with functional arrays.

Binding Assays

Protein–Protein Interactions

Protein complex formation and disruption play major roles in cellular regulatory processes and elucidation of the "interactome" is of great interest for investigating signaling pathways and potential therapeutic targets.[66] One of the first tools for defining protein-protein interactions on a proteomic scale was the two-hybrid system first developed in yeast,[67–69] more recently applied to non-yeast interactomes,[70–72] and now available in other systems.[73]

Functional protein arrays are an important *in vitro* method of investigating protein-protein interactions and will help to validate interactions defined in other systems as well as define new interactions in their own right. Protein arrays also enable the characterization of antibody binding, both for specificity profiling and for immune response profiling, which is not readily achieved in *in vivo* systems such as two-hybrid.

Single Protein Probes

Experiments to investigate protein interactions on protein arrays can take any of a number of approaches. If one is interested in the interactions of a specific purified protein, there are several options for designing the experiment. If the protein of interest has been purified from a native source, it may be chemically modified with a small organic fluorescent dye (as described above) and used directly to probe the protein array (Figure 9.2A). After washing away non-specifically bound

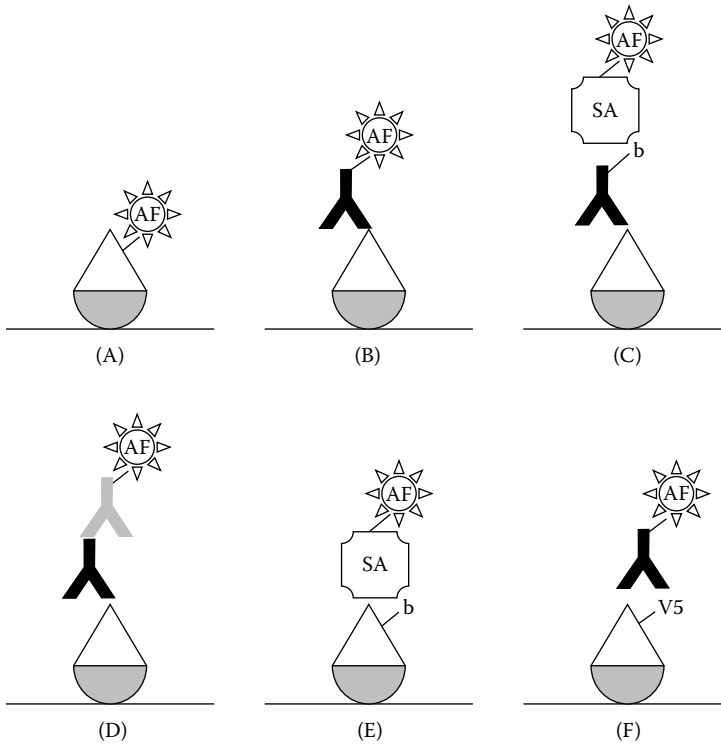


FIGURE 9.2 Detection Strategies — Single Protein Probes. Direct detection (**A**) of protein interaction between probe protein (triangle) and arrayed target protein (grey semi-circle) is achieved using Alexa Fluor (AF)-labeled probe protein. Indirect detection of probe protein binding (**B-F**) is achieved using a variety of strategies. (**B**) An Alexa Fluor-labeled antibody specific for the probe protein is added after binding of the probe to the arrayed target. (**C**) A biotinylated (b) antibody specific for the probe protein is added after binding of the probe to the arrayed target, followed by addition of Alexa Fluor-labeled streptavidin (SA). (**D**) An unlabeled antibody specific for the probe protein is added after binding of the probe to the arrayed target, followed by addition of Alexa Fluor-labeled anti-IgG. (**E**) A biotinylated probe protein is bound to the arrayed target, followed by addition of Alexa Fluor-labeled streptavidin. (**F**) A recombinant epitope-tagged probe protein, in this case bearing the V5 epitope, is bound to the arrayed target, followed by addition of Alexa Fluor-labeled anti-epitope antibody.

materials, interactions are detected with standard fluorescent microarray instrumentation. One caveat with this approach is that some fluorescent dyes can affect the function and/or properties of proteins, depending on the degree to which they introduce hydrophobic character or target residues crucial to protein function. If an antibody or other affinity reagent exists that is specific for the protein of interest, it can be used as a detection reagent, either fluorescently labeled (Figure 9.2B), haptenylated (Figure 9.2C), or unlabeled (Figure 9.2D). If unlabeled, an additional labeled detection reagent (e.g., fluorescent anti-IgG) must also be employed (Figure 9.2D). Alternatively, the protein of interest can be conjugated to a hapten, such as biotin, and used to probe the protein array. In this case, a fluorescent secondary detection reagent, such as streptavidin, is used to detect interactions (Figure 9.2E).

If the protein of interest has been cloned, it may be engineered for expression in a convenient system, such as *E. coli*, baculovirus, or mammalian cells. For recombinant protein probes, the options are numerous. One can employ any of the epitope or conjugation tags described earlier in this chapter (Figure 9.2F). Furthermore, with the use of epitope tags, it is not necessary to purify the protein of interest from the expression system, as long as it is the only protein in the system bearing the epitope. The disadvantage of crude probes is mainly in the quantitation of the amount of specific probe used in an array experiment, which can be more difficult and require a more specialized assay for a crude preparation. Epitope tagged recombinant proteins are used in conjunction with commercial anti-epitope reagents (usually antibodies) that quite often are available in fluorescently labeled forms or can be readily labeled with the fluorescent dye of choice. Epitope tagged probes are used in a two step experiment: first probing the array with the epitope tagged protein of interest (either crude or purified), followed by detection with the fluorescent anti-epitope reagent.

An example of protein interaction on functional arrays is the landmark study by Zhu et al. using yeast proteome arrays.[9] Representing the first whole proteome array in the literature, the arrays were produced by high throughput cloning, expression, and purification of the yeast proteome. The recombinant GST fusion proteins were purified and arrayed under mild conditions to preserve function. The arrays were probed with purified biotinylated calmodulin in the presence of Ca^{2+}, followed by Cy3-streptavidin detection. Many known and new interactions were detected and a new sequence motif was defined that was common to many of the calmodulin binding target proteins. Other studies also support the use of functional protein arrays for investigating protein-protein interactions.[58,74–80]

Complex Protein Probes

If one is interested in profiling the global interactions of a complex protein preparation, such as a cell or tissue lysate, physiological fluid, culture supernatant, or a subcellular protein fraction (e.g., nuclear or membrane extract), the protein preparation must be labeled prior to probing the array. The most expeditious method for complex protein samples is to label with an reactive fluorescent dye for a one-step probing procedure. One could also label with a hapten, then use a fluorescent anti-hapten reagent for a two-step procedure. An extension of this idea

is a gene expression profiling type of experiment adapted for proteins in which two crude protein preparations, representing two experimental conditions, are labeled with two different color fluorescent dyes (or two different haptens) and probed on a single array. The relative signal in each fluorescent channel for a particular array feature reflects the relative abundance of protein in the different samples that binds to the feature. This approach has been used successfully with antibody arrays, as numerous reports in the literature show.[31–33,81–87] Most of these studies use some type of data normalization to contend with bias introduced by labeling and other factors.[46]

The use of non-antibody functional protein arrays for protein profiling experiments has the potential to reveal additional biological information that antibody arrays may not. Protein profiling on antibody arrays presumably yields information on relative quantities of proteins in two samples, which may give insight into molecular differences between the samples. Protein profiling on functional arrays could possibly yield unique information on protein complex formation, thus potentially giving insight into mechanistic differences between the samples and affording insight into the "interactome." Several difficulties with profiling on functional arrays need to be overcome, however, to determine if this approach will bear fruit. First, it is likely that the affinities involved in protein-protein interactions of biologically relevant complexes may be lower than those of antigen-antibody interactions (where typical dissociation constants are in the μ*M* to high p*M* range). In addition, it is anticipated that biologically relevant binding partners may be expressed at moderate or low levels in the cell. Thus, sensitivity will be a challenge if one hopes to detect non-abundant proteins in such a system. This may be partly overcome by enriching for a particular class of probe protein and/or removing abundant proteins. Second, unlike antibody arrays where the identities of bound probe proteins are predictable based on antibody specificity, the binding of complex samples on functional protein arrays is unpredictable. This necessitates follow-on experiments, such as pull-down followed by mass spectrometry or protein sequence analysis, to verify and identify the interaction partners.

Antibody Specificity, Immune Response, and Autoimmune Profiling

Several groups have shown how protein arrays can be used to characterize antibody specificity,[88–90] profile immune response[91–94] and characterize autoimmune diseases.[95–97] All these types of experiments have in common the detection of antibodies bound to protein arrays. The requirement for functionality of arrayed proteins may be less stringent than for other applications, since antibodies are known to bind to linear (denatured) as well as conformational epitopes. Functional protein arrays can support the investigation of both conformational and denatured epitopes, since the arrays can be treated with denaturants before probing. Whole proteome arrays are a particularly powerful tool for characterizing antibodies as they present a wide variety of relevant epitopes in a single assay. However, whole proteome arrays are more likely to be used in latter stage characterization of antibody specificity, due to the expense and labor of constructing them. Smaller, more focused protein arrays could be quite valuable in earlier stage, higher throughput characterization of antibody panels during the screening and selection process. De Masi and colleagues[98] turned

the process around by coating slides with an antigen solution, then printing hybridoma supernatants onto the antigen slides as a high-throughput method of screening large numbers of hybridomas.

The most straightforward approach to detecting antibody bound to protein arrays is to probe the array with the antibody of interest and then utilize a fluorescent labeled secondary (anti-antibody) reagent, such as a species-specific anti-Ig. Such reagents are readily available from commercial sources in both labeled and unlabeled forms. Alternatively, in cases where purified antibody specificity is being examined, one can directly label the antibody of interest, either with fluorescent dye or a common hapten (e.g., biotin) for which a labeled secondary reagent is available. Direct dye labeling of the antibody avoids the need for a fluorescent secondary reagent, but may require some labeling optimization to avoid artifacts from the labeling itself. Labeling with a hapten allows the use of any of a number of commercially available fluorescent anti-hapten reagents, such as streptavidin for biotin binding.

Protein–DNA Interactions

Several groups have probed functional protein arrays with DNA as a means of investigating protein–DNA interactions.[78,79,99] These studies used DNA fragments as probes that were fluorescently labeled by standard nucleic acid probe preparation methodologies. In practice, assays with labeled nucleic acids are performed and analyzed similar to other array probing experiments: if the nucleic acids are labeled with a fluorophore, interactions are detected directly after probing and washing; if the nucleic acids are labeled with a hapten, then a fluorescent secondary detection reagent (anti-hapten) is used.

Protein–Lipid Interactions

Zhu et al. studied lipid binding on functional yeast proteome arrays by using liposomes containing five different phospholipids with a biotinylated tracer lipid for detection with Cy3-straptavidin.[9] A total of 150 protein targets were identified on the proteome array that bound one or another of the labeled liposomes, including a large class of membrane-associated proteins, a smaller class of lipid metabolizing proteins, and even some protein kinases. This study points out the utility of functional proteins arrays for protein–lipid studies, but also shows the need for specially designed detection reagents. Besides haptenylated lipids, fluorescent lipids could be used for direct detection of lipid binding to protein arrays. Fluorescent lipids are being developed and used in cellular lipid and membrane studies[100,101] and similar reagents should be applicable to protein arrays.

Protein–Small Molecule Interactions

Several studies investigating the binding of small molecules to functional protein arrays have appeared in the literature. MacBeath and Schreiber[74] probed functional protein arrays with fluorescent-BSA-coupled small molecules and showed specific interaction with arrayed target proteins. Fang et al.[102] printed functional lipid-bound

G-protein-coupled receptors (GPCRs) and probed them with fluorescent ligands and showed specific binding. In further studies with GPCR arrays, Hong et al.[21] probed the arrays with non-hydrolyzable europium-labeled GTP analogs to investigate GPCR activation by time-resolved fluorescence detection. Huang et al.[103] probed yeast functional proteome arrays with biotinylated small molecule inhibitors of rapamycin followed by Cy3-streptavidin detection, revealing several new components of the target of rapamycin (TOR) signaling pathway. These studies again highlight the need for specialized reagents compatible with small molecule assays and underscore the flexibility of fluorescence detection to meet the needs of these and similar applications. Small molecule probes can be directly conjugated with fluorescent dyes, with haptens, or with larger fluorescent labeled carrier proteins. Appropriate labeled secondary detection reagents must be employed if small molecule probes are haptenylated.

Activity Assays

Kinases and Phosphatases

Two early studies showed the feasibility of detecting protein phosphorylation events with protein arrays. MacBeath and Shreiber arrayed kinase substrates, then exposed them to cognate kinases and [γ-^{33}P]-ATP and observed on-chip phosphorylation by dipping arrays into photographic emulsion and imaging with a microscope.[74] Zhu et al.[104] cloned, expressed, and purified 119 GST-fusion kinases from yeast and arrayed them in microwells. After exposure to a panel of different kinase substrates and [γ-^{33}P]-ATP, on-chip phosphorylation was visualized by phosphorimaging. These studies point to the need for good fluorescent phosphorylation reagents to avoid radioactivity and achieve higher resolution images. More recent studies with p53 variants[78] showed on-chip phosphorylation using an anti-phosphoserine antibody and subsequent anti-IgG-HRP conjugate to generate chemiluminescent signal detected by film exposure. Ptacek et al. exposed functional yeast proteome arrays to purified kinases and [γ-^{33}P]-ATP and detected on-chip phosphorylation by autoradiography.[105]

Thus, it is possible to use arrayed proteins either as kinases or kinase substrates to perform on-chip phosphorylation studies. However, fluorescent reagents for phosphorylation activity have not made great inroads in microarray platforms. Molecular Probes (Eugene, OR) has developed a sensitive phosphoprotein stain, Pro-Q Diamond, capable of detecting phosphoproteins on planar arrays,[61] but these studies were not performed with on-chip phosphorylation. There are a variety of anti-phospho-amino acid antibodies commercially available that are amenable to fluorescent detection, either labeled directly or used in conjunction with a labeled secondary reagent. Many of the earlier anti-phospho-amino acid antibodies were developed for denaturing Western blot phosphoprotein assays and are not as well-suited to native phosphoprotein detection. Continued development of these kinds of reagents will lead to better native state assays for defining the phosphorylation state of proteins on arrays. Fluorescent reagents for solution phase high-throughput phosphorylation assays have been developed that utilize unique properties of fluorophores (such as fluorescence quenching, fluorescence

resonance energy transfer, and fluorescence polarization), but these kinds of reagents have not yet been applied to microarray formats. For such reagents to be useful on microarrays, they need to precipitate, attach covalently to targets, or be utilized in "multiple spotting" array formats[106] so that they remain in addressable locations. Chen et al. have been developing mechanism-based fluorescent molecules suitable for kinase and phosphatase assays on microarrays.[107] These fluorescent reagents are suicide inhibitors that bind covalently to target enzymes on arrays and are detectable with standard microarray scanners. Yee et al.[108] have developed wortmannin-based labels (biotin, BODIPY, and tetramethylrhodamine versions) for the selective covalent labeling of lipid and protein kinases. BODIPY-wortmannin is cell permeable and can specifically label proteins within cells. These probes appear to be activity based (wortmannin is a specific inhibitor of members of these classes of kinases) and should be useful reagents for probing kinase activities on functional arrays.

Proteases

Proteases are an important class of enzymes and are often targets of drug therapy. The ability to characterize protease inhibitor specificity is crucial to developing such therapies. Various groups have begun investigating proteolytic activity in microarray formats, but few studies have used functional protein arrays. Shao Yao's group in Singapore has been developing fluorogenic substrates and inhibitors for array-based assays.[107,109,110] They have been arraying the substrates or inhibitors, or mixtures of the enzymes plus substrates or inhibitors, but have also detected proteolytic activity when the enzyme itself is printed and the array is reacted with substrates or inhibitors.[107] Gosalia et al.[11] printed a library of fluorogenic protease substrates and then over-sprayed the array with aerosolized proteases and detected on-array proteolytic activity. Harris et al.[111] profiled the proteolytic activity in dust mite extracts by incubating extracts with peptide-nucleic acid- (PNA-) encoded cysteine protease probes and deconvoluting them on nucleic acid arrays. Winssinger et al.[112] used a similar approach with PNA-encoded fluorogenic substrates to profile proteolytic activity of purified and crude protease preparations. The development of these types of reagents, particularly the probes that remain bound to proteases, makes it possible to assay for proteolytic activity directly on functional protein arrays. Ideally, one would like to be able to probe functional arrays with specific fluorogenic substrates and inhibitors to profile classes of proteases on-chip and to characterize the binding of potential protease inhibitors to various classes of proteases. Another interesting type of profiling would be to expose a functional array — more specifically, a proteome array — to a given protease(s) for the purposes of determining which arrayed proteins are substrates for the protease(s). This type of experiment has not yet been done on functional arrays as there are no reagents available to generate signal from such reactions. However, proteolysis of arrayed proteins could possibly be examined by monitoring the loss of protein after protease treatment using a general protein stain or a reagent specific for N- or C-terminal tags on the arrayed proteins (assuming cleaved fragments become liberated from the array surface).

Other Enzymes

As proof of principle, several groups have performed various other enzymatic assays directly on planar arrays of proteins. For instance, Arenkov et al.[113] assayed for HRP, alkaline phosphatase (AP), and β-D-glucaronidase, the latter two enzymes with precipitating fluorogenic substrates (ELF-97 phosphate and ELF-97 β-D-glucaronide from Molecular Probes). Angenendt et al.[106] performed a similar study with HRP, AP, and β-galactosidase printed onto previously immobilized fluorogenic substrates.

In further demonstrations of on-chip enzymatic activity, several groups have investigated metabolic enzyme activity. Lee et al.[114] spotted cytochrome P450s in sol gel, then over-spotted various drugs, followed by contacting with a cellular monolayer. Conversion of drug to toxic metabolite was demonstrated by post-reaction staining of cells with a fluorogenic live / dead cell test reagent. Jung et al.[115] attached mRNA-protein fusions representing five enzymes in the trehalose biosynthetic pathway to microplate wells via hybridization with capture DNA in the wells. By addition of glucose to the system, they were able to demonstrate trehalose synthesis. Also, by modulating the levels of the five enzymes involved (by altering amounts of capture DNA in the plate wells), they were able to optimize the pathway for trehalose production. However, the readout for trehalose production was based on a colorimetric enzymatic assay.

While these examples illustrate the feasibility and utility of enzymatic assays in protein array formats, they also show that application of fluorescent detection to such systems often requires highly specialized reagents; in some cases these reagents are available and in other cases they are yet to be developed.

FUTURE DIRECTIONS

There are several overriding needs for the continued development of fluorescent detection methods for functional protein arrays and advances in each area will enhance the utility of functional arrays and drive their continued acceptance. First is new developments in fluorescent microarray instrumentation. Continued expansion of the range of excitation spectra available on commercial laser-based instruments will expand the range of usable fluorescent compounds. White light/CCD-based systems are intrinsically capable of a wider excitation spectrum, but these systems do not seem to be as popular as laser-based systems. UV excitation would open the door to lanthanide chelates and TRF, if combined with TRF imaging capabilities already developed for microscopy applications. Further exploration of the benefits of fluorescent waveguide excitation might lead to more sensitive systems that could measure assays at equilibrium rather than at end points subject to effects of ligand off-rates. All of these improvements would help provide an even wider range of fluorescent options for protein biochemical assays on arrays.

Second is new methods of protein labeling, particularly for labeling complex protein samples where current methods are plagued by labeling bias. Since complex protein samples present very heterogeneous targets for chemical modification, it is not

surprising that label would be heterogeneously dispersed in such samples, and that some proteins may not even label at all. It may be that uniform protein labeling of complex samples is only achievable by metabolic labeling methods, such as those discussed above for the introduction of azido groups for Staudinger chemistry. Other metabolic labeling methods developed for mass spec, such as stable isotope labeling by amino acids in cell culture (SILAC), may be applicable to array probe preparation, but adaptations of the methods for fluorescent detection are needed. Further exploration of fluorescent amino acid analogs like BODIPY-FL-lysine[15] might be useful both in cell culture applications as well as *in vitro* protein production. But, there remains the considerable problem of uniform labeling of complex protein preparations derived from protein extraction procedures. To achieve more homogeneous post-extraction or post-harvest labeling, new developments in protein chemical modification are needed.

Third is the continued development of specialized fluorescent reagents for enzymatic assays, proteolysis, phosphorylation, small molecules, etc. Given the variety of fluorescent labels and chemistries available, it is likely that such specialized fluorescent probes will continue to be developed on a case-by-case basis. Fluorescent reagents for FRET, TRF, and fluorescence polarization types of assays will likely continue to emerge for functional array applications.

Finally, the development of brighter and more stable fluorescent entities and new signal detection/amplification methods will continue to advance and will lead to increased sensitivity. For example, silver nanoclusters, bundles of silver atoms that exhibit fluorescence spectral properties in the visible and near IR range, may be useful for microarray applications. Currently this technology is still being developed and applications have not been demonstrated on functional protein arrays. The clusters are characterized by large Stokes shifts and narrow emission profiles whose wavelengths are dependent upon the number of atoms in the clusters. Passivation and functionalization of nanocrystals present the same kinds of challenges as those of Qdots. An example of a newer detection method is surface-enhanced Raman scattering (SERS). Although technically not a fluorescent method, SERS is a light-based signal detection method that uses various types of noble metal nanostructures.[116] SERS has been applied to immunoassay applications[117] and developed into an ELISA-like format where it has been used for detection of low levels of PSA in serum.[118] Advantages of this method include narrow spectral bandwidth, resistance to photo-bleaching and quenching, and long-wavelength excitation of multiple labels with a single excitation source.[118] SERS should be readily adaptable to protein array detection and may be a way to significantly boost fluorescent signals on arrays.

These developments, combined with the ongoing improvements in protein array surfaces, attachment chemistries, printing/manufacturing, and content will guarantee the widening use of protein arrays in discovery research and diagnostics.

TRADEMARKS

REFERENCES

1. Dittrich, P.S. and Manz, A., Single-molecule fluorescence detection in microfluidic channels — The holy grail in μTAS?, *Anal. Bioanal. Chem.*, 382, 1771, 2005.
2. Schena, M. and Davis, R.W., Genes, genomes, and chips, in *DNA Microarrays: A Practical Approach*, Schena, M., Ed., Oxford University Press, Inc., New York, 1999, p. 1.
3. Schermer, M.J., Confocal scanning microscopy in microarray detection, in *DNA Microarrays: A Practical Approach*, Schena, M., Ed., Oxford University Press, Inc., New York, 1999, p. 17.
4. Cutler, P., Protein arrays: The current state-of-the-art, *Proteomics*, 3, 3, 2003.
5. Haab, B.B., Methods and applications of antibody microarrays in cancer research, *Proteomics*, 3, 2116, 2003.
6. Seong, S.-Y. and Choi, C.-Y., Current status of protein chip development in terms of fabrication and application, *Proteomics*, 3, 2176, 2003.
7. Sun, C. et al., Advances in the study of luminescence probes for proteins, *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, 803, 173, 2004.
8. Ekins, R.P. and Chu, F.W., Multianalyte microspot immunoassay — microanalytical "compact disk" of the future, *Clin. Chem.*, 37, 1955, 1991.
9. Zhu, H. et al., Global analysis of protein activities using proteome chips, *Science*, 293, 2101, 2001.
10. Telford, W.G., Hawley, T.S., and Hawley, R.G., Analysis of violet-excited fluorochromes by flow cytometry using a violet laser diode, *Cytometry A*, 54, 48, 2003.
11. Gosalia, D.N. et al., Profiling serine protease substrate specificity with solution phase fluorogenic peptide microarrays, *Proteomics*, 5, 1292, 2005.
12. Berlier, J.E. et al., Quantitative comparison of long-wavelength Alexa Fluor dyes to Cy dyes: Fluorescence of the dyes and their bioconjugates, *J. Histochem. Cytochem.*, 51, 1699, 2003.
13. Forster, T. et al., Triple-target microarray experiments: A novel experimental strategy, *BMC Genomics*, 5, 13, 2004.
14. Staal, Y.C. et al., Application of four dyes in gene expression analyses by microarrays, *BMC Genomics*, 6, 101, 2005.
15. Coleman, M.A. et al., High-throughput, fluorescence-based screening for soluble protein expression, *J. Proteome Res.*, 3, 1024, 2004.
16. Zhang, J. et al., Creating new fluorescent probes for cell biology, *Nat. Rev. Mol. Cell Biol.*, 3, 906, 2002.

17. Miyawaki, A., Sawano, A., and Kogure, T., Lighting up cells: Labelling proteins with fluorophores, *Nat. Cell Biol., Suppl.*, S1, 2003.
18. Shaner, N.C., Steinbach, P.A., and Tsien, R.Y., A guide to choosing fluorescent proteins, *Nat. Methods*, 2, 905, 2005.
19. Kukar, T. et al., Protein microarrays to detect protein-protein interactions using red and green fluorescent proteins, *Anal. Biochem.*, 306, 50, 2002.
20. Luo, L.Y. and Diamandis, E.P., Preliminary examination of time-resolved fluorometry for protein array applications, *Luminescence*, 15, 409, 2000.
21. Hong, Y. et al., Functional GPCR microarrays, *J. Am. Chem. Soc.*, 127, 15350, 2005.
22. Bruchez, M., Jr. et al., Semiconductor nanocrystals as fluorescent biological labels, *Science*, 281, 2013, 1998.
23. Lian, W. et al., Ultrasensitive detection of biomolecules with fluorescent dye-doped nanoparticles, *Anal. Biochem.*, 334, 135, 2004.
24. Zhou, X. and Zhou, J., Improving the signal sensitivity and photostability of DNA hybridizations on microarrays by using dye-doped core-shell silica nanoparticles, *Anal. Chem.*, 76, 5302, 2004.
25. Haugland, R.P., *The Handbook: A Guide to Fluorescent Probes and Labeling Technologies*, Molecular Probes, Eugene, OR, 2005.
26. van Gijlswijk, R.P. et al., Fluorochrome-labeled tyramides: Use in immunocytochemistry and fluorescence *in situ* hybridization, *J. Histochem. Cytochem.*, 45, 375, 1997.
27. Lizardi, P.M. et al., Mutation detection and single-molecule counting using isothermal rolling-circle amplification, *Nat. Genet.*, 19, 225, 1998.
28. Schweitzer, B. et al., Multiplexed protein profiling on microarrays by rolling-circle amplification, *Nat. Biotechnol.*, 20, 359, 2002.
29. Schweitzer, B. et al., Immunoassays with rolling circle DNA amplification: A versatile platform for ultrasensitive antigen detection, *Proc. Natl. Acad. Sci. USA*, 97, 10113, 2000.
30. Kingsmore, S.F. and Patel, D.D., Multiplexed protein profiling on antibody-based microarrays by rolling circle amplification, *Curr. Opin. Biotechnol.*, 14, 74, 2003.
31. Zhou, H. et al., Two-color, rolling-circle amplification on antibody microarrays for sensitive, multiplexed serum-protein measurements, *Genome Biol.*, 5, R28, 2004.
32. Gao, W.M. et al., Distinctive serum protein profiles involving abundant proteins in lung cancer patients based upon antibody microarray analysis, *BMC Cancer*, 5, 110, 2005.
33. Orchekowski, R. et al., Antibody microarray profiling reveals individual and combined serum proteins associated with pancreatic cancer, *Cancer Res.*, 65, 11193, 2005.
34. Hermanson, G.T., *Bioconjugate Techniques*, Academic Press, San Diego, CA, 1996.
35. Cuatrecasas, P. and Parikh, I., Adsorbents for affinity chromatography. Use of N-hydroxysuccinimide esters of agarose, *Biochemistry*, 11, 2291, 1972.
36. Toutchkine, A., Nalbant, P., and Hahn, K.M., Facile synthesis of thiol-reactive Cy3 and Cy5 derivatives with enhanced water solubility, *Bioconjug. Chem.*, 13, 387, 2002.
37. Chen, I. et al., Site-specific labeling of cell surface proteins with biophysical probes using biotin ligase, *Nat. Methods*, 2, 99, 2005.
38. Geoghegan, K.F. and Stroh, J.G., Site-directed conjugation of nonpeptide groups to peptides and proteins via periodate oxidation of a 2-amino alcohol. Application to modification at N-terminal serine, *Bioconjug. Chem.*, 3, 138, 1992.

39. Zhang, C.X., Chang, P.V., and Lippard, S.J., Identification of nuclear proteins that interact with platinum-modified DNA by photoaffinity labeling, *J. Am. Chem. Soc.*, 126, 6536, 2004.
40. Saxon, E. and Bertozzi, C.R., Cell surface engineering by a modified Staudinger reaction, *Science*, 287, 2007, 2000.
41. Kiick, K.L. et al., Incorporation of azides into recombinant proteins for chemoselective modification by the Staudinger ligation, *Proc. Natl. Acad. Sci. USA*, 99, 19, 2002.
42. Hang, H.C. et al., A metabolic labeling approach toward proteomic analysis of mucin-type O-linked glycosylation, *Proc. Natl. Acad. Sci. USA*, 100, 14846, 2003.
43. Sprung, R. et al., Tagging-via-substrate strategy for probing O-GlcNAc-modified proteins, *J. Proteome Res.*, 4, 950, 2005.
44. Nandi, A. et al., Global identification of O-GlcNAc-modified proteins, *Anal. Chem.*, 78, 452, 2006.
45. Lemieux, G.A., De Graffenried, C.L., and Bertozzi, C.R., A fluorogenic dye activated by the Staudinger ligation, *J. Am. Chem. Soc.*, 125, 4708, 2003.
46. Eckel-Passow, J.E. et al., Experimental design and analysis of antibody microarrays: Applying methods from cDNA arrays, *Cancer Res.*, 65, 2985, 2005.
47. Terpe, K., Overview of tag protein fusions: From molecular and biochemical fundamentals to commercial systems, *Appl. Microbiol. Biotechnol.*, 60, 523, 2003.
48. Brothers, S.P., Janovick, J.A., and Conn, P.M., Unexpected effects of epitope and chimeric tags on gonadotropin-releasing hormone receptors: Implications for understanding the molecular etiology of hypogonadotropic hypogonadism, *J. Clin. Endocrinol. Metab.*, 88, 6107, 2003.
49. Waugh, D.S., Making the most of affinity tags, *Trends Biotechnol.*, 23, 316, 2005.
50. Murphy, D., Gene expression studies using microarrays: Principles, problems, and prospects, *Adv. Physiol. Educ.*, 26, 256, 2002.
51. Hegde, P. et al., A concise guide to cDNA microarray analysis, *Biotechniques*, 29, 548, 2000.
52. Kricka, L.J., Stains, labels and detection strategies for nucleic acids assays, *Ann. Clin. Biochem.*, 39, 114, 2002.
53. Holloway, A.J. et al., Options available — from start to finish — for obtaining data from DNA microarrays II, *Nat. Genet.*, 32 Suppl, 481, 2002.
54. Hilario, E., End labeling procedures: An overview, *Mol. Biotechnol.*, 28, 77, 2004.
55. Schena, M., Ed., *Microarray Biochip Technology*, Eaton Publishing Company, Natick, MA, 2000.
56. Schena, M., Ed., *DNA Microarrays: A Practical Approach*, Oxford University Press, New York, NY, 1999.
57. Shi, L., www.Gene-Chips.com, 2002.
58. Kawahashi, Y. et al., *In vitro* protein microarrays for detecting protein-protein interactions: Application of a new method for fluorescent labeling of proteins, *Proteomics*, 3, 1236, 2003.
59. Tan, L.P. and Yao, S.Q., Intein-mediated, *in vitro* and *in vivo* protein modifications with small molecules, *Protein Pept. Lett.*, 12, 769, 2005.
60. Taki, M., Sawata, S.Y., and Taira, K., Specific N-terminal biotinylation of a protein *in vitro* by a chemically modified tRNA(fmet) can support the native activity of the translated protein, *J. Biosci. Bioeng.*, 92, 149, 2001.
61. Martin, K. et al., Quantitative analysis of protein phosphorylation status and protein kinase activity on microarrays using a novel fluorescent sensor dye, *Proteomics*, 3, 1244, 2003.

62. Martin, K. et al., Strategies and solid-phase formats for the analysis of protein and peptide phosphorylation employing a novel fluorescent phosphorylation sensor dye, *Comb. Chem. High Throughput Screen*, 6, 331, 2003.
63. Hart, C. et al., Fluorescence detection and quantitation of recombinant proteins containing oligohistidine tag sequences directly in sodium dodecyl sulfate-polyacrylamide gels, *Electrophoresis*, 24, 599, 2003.
64. Mace, M.L., Jr. et al., Novel microarray printing and detection technologies, in *Microarray Biochip Technology*, Schena, M., Ed., Eaton Publishing, Natick, MA, 2000, p. 39.
65. Pawlak, M. et al., Zeptosens' protein microarrays: A novel high performance microarray platform for low abundance protein analysis, *Proteomics*, 2, 383, 2002.
66. Ghavidel, A., Cagney, G., and Emili, A., A skeleton of the human protein interactome, *Cell*, 122, 830, 2005.
67. Fields, S. and Song, O., A novel genetic system to detect protein-protein interactions, *Nature*, 340, 245, 1989.
68. Ito, T. et al., Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins, *Proc. Natl. Acad. Sci. USA*, 97, 1143, 2000.
69. Uetz, P. et al., A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*, *Nature*, 403, 632, 2000.
70. Rual, J.F. et al., Towards a proteome-scale map of the human protein-protein interaction network, *Nature*, 437, 1173, 2005.
71. Li, S. et al., A map of the interactome network of the metazoan *C. elegans*, *Science*, 303, 540, 2004.
72. Giot, L. et al., A protein interaction map of *Drosophila melanogaster*, *Science*, 302, 1727, 2003.
73. Lee, J.W. and Lee, S.K., Mammalian two-hybrid assay for detecting protein-protein interactions *in vivo*, *Methods Mol. Biol*., 261, 327, 2004.
74. MacBeath, G. and Schreiber, S., Printing proteins as microarrays for high-throughput function determination, *Science*, 289, 1760, 2000.
75. Espejo, A. et al., A protein-domain microarray identifies novel protein-protein interactions, *Biochem. J.*, 367, 697, 2002.
76. Lee, Y. et al., Proteochip: A highly sensitive protein microarray prepared by a novel method of protein immobilization for application of protein-protein interaction studies, *Proteomics*, 3, 2289, 2003.
77. Ramachandran, N. et al., Self-assembling protein microarrays, *Science*, 305, 86, 2004.
78. Boutell, J.M. et al., Functional protein microarrays for parallel characterisation of p53 mutants, *Proteomics*, 4, 1950, 2004.
79. Snapyan, M. et al., Dissecting DNA-protein and protein-protein interactions involved in bacterial transcriptional regulation by a sensitive protein array method combining a near-infrared fluorescence detection, *Proteomics*, 3, 647, 2003.
80. Coleman, M.A. et al., Identification of chromatin-related protein interactions using protein microarrays, *Proteomics*, 3, 2101, 2003.
81. Haab, B.B., Dunham, M.J., and Brown, P.O., Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions, *Genome Biol.*, 2, research0004.1, 2001.
82. Sreekumar, A. et al., Profiling of cancer cells using protein microarrays: Discovery of novel radiation-regulated proteins, *Cancer Res*., 61, 7585, 2001.

83. Miller, J.C. et al., Antibody microarray profiling of human prostate cancer sera: Antibody screening and identification of potential biomarkers, *Proteomics*, 3, 56, 2003.
84. Anderson, K. et al., Protein expression changes in spinal muscular atrophy revealed with a novel antibody array technology, *Brain*, 126, 2052, 2003.
85. Yeretssian, G. et al., Competition on nitrocellulose-immobilized antibody arrays: From bacterial protein binding assay to protein profiling in breast cancer cells, *Mol. Cell. Proteomics*, 4, 605, 2005.
86. Kopf, E., Shnitzer, D., and Zharhary, D., Panorama Ab microarray cell signaling kit: A unique tool for protein expression analysis, *Proteomics*, 5, 2412, 2005.
87. Hamelinck, D. et al., Optimized normalization for antibody microarrays and application to serum-protein profiling, *Mol Cell. Proteomics*, 4, 773, 2005.
88. Lueking, A. et al., Protein microarrays for gene expression and antibody screening, *Anal. Biochem.*, 270, 103, 1999.
89. Michaud, G.A. et al., Analyzing antibody specificity with whole proteome microarrays, *Nat. Biotechnol.*, 21, 1509, 2003.
90. Poetz, O. et al., Protein microarrays for antibody profiling: Specificity and affinity determination on a chip, *Proteomics*, 5, 2402, 2005.
91. Davies, D.H. et al., Profiling the humoral immune response to infection by using proteome microarrays: High-throughput vaccine and diagnostic antigen discovery, *Proc. Natl. Acad. Sci. USA*, 102, 547, 2005.
92. Nam, M.J. et al., Molecular profiling of the immune response in colon cancer using protein microarrays: Occurrence of autoantibodies to ubiquitin C-terminal hydrolase L3, *Proteomics*, 3, 2108, 2003.
93. Stone, J.D., Demkowicz, W.E., Jr., and Stern, L.J., HLA-restricted epitope identification and detection of functional T cell responses by using MHC-peptide and costimulatory microarrays, *Proc. Natl. Acad. Sci. USA*, 102, 3744, 2005.
94. Li, B. et al., Protein microarray for profiling antibody responses to *Yersinia pestis* live vaccine, *Infect. Immun.*, 73, 3734, 2005.
95. Robinson, W.H. et al., Autoantigen microarrays for multiplex characterization of autoantibody responses, *Nat. Med.*, 8, 295, 2002.
96. Joos, T.O. et al., A microarray enzyme-linked immunosorbent assay for autoimmune diagnostics, *Electrophoresis*, 21, 2641, 2000.
97. Hueber, W. et al., Antigen microarray profiling of autoantibodies in rheumatoid arthritis, *Arthritis Rheum*., 52, 2645, 2005.
98. De Masi, F. et al., High throughput production of mouse monoclonal antibodies using antigen microarrays, *Proteomics*, 5, 4070, 2005.
99. Hall, D.A. et al., Regulation of gene expression by a metabolic enzyme, *Science*, 306, 482, 2004.
100. Ishitsuka, R., Sato, S.B., and Kobayashi, T., Imaging lipid rafts, *J. Biochem.* (Tokyo). 137, 249, 2005.
101. Sanchez, S.A. and Gratton, E., Lipid-protein interactions revealed by two-photon microscopy and fluorescence correlation spectroscopy, *Acc. Chem. Res*., 38, 469, 2005.
102. Fang, Y., Frutos, A.G., and Lahiri, J., Membrane protein microarrays, *J. Am. Chem. Soc.*, 124, 2394, 2002.
103. Huang, J. et al., Finding new components of the target of rapamycin (TOR) signaling network through chemical genetics and proteome chips, *Proc. Natl. Acad. Sci. USA*, 101, 16594, 2004.

104. Zhu, H. et al., Analysis of yeast protein kinases using protein chips, *Nat. Genet.*, 26, 283, 2000.
105. Ptacek, J. et al., Global analysis of protein phosphorylation in yeast, *Nature*, 438, 679, 2005.
106. Angenendt, P. et al., Subnanoliter enzymatic assays on microarrays, *Proteomics*, 5, 420, 2005.
107. Chen, G.Y. et al., Developing a strategy for activity-based detection of enzymes in a protein microarray, *Chembiochem.*, 4, 336, 2003.
108. Yee, M.C. et al., A cell-permeable, activity-based probe for protein and lipid kinases, *J. Biol. Chem.*, 280, 29053, 2005.
109. Uttamchandani, M. et al., Nanodroplet profiling of enzymatic activities in a microarray, *Bioorg. Med. Chem. Lett.*, 15, 2135, 2005.
110. Srinivasan, R. et al., Activity-based fingerprinting of proteases, *Chembiochem.*, 7, 32, 2005.
111. Harris, J. et al., Activity profile of dust mite allergen extract using substrate libraries and functional proteomic microarrays, *Chem. Biol.*, 11, 1361, 2004.
112. Winssinger, N. et al., PNA-encoded protease substrate microarrays, *Chem. Biol.*, 11, 1351, 2004.
113. Arenkov, P. et al., Protein microchips: Use for immunoassay and enzymatic reactions, *Anal. Biochem.*, 278, 123, 2000.
114. Lee, M.Y. et al., Metabolizing enzyme toxicology assay chip (MetaChip) for high-throughput microscale toxicity analyses, *Proc. Natl. Acad. Sci. USA*, 102, 983, 2005.
115. Jung, G.Y. and Stephanopoulos, G., A functional protein chip for pathway optimization and *in vitro* metabolic engineering, *Science*, 304, 428, 2004.
116. Smith, W.E., Surface-enhanced resonance raman scattering, *Methods Enzymol.*, 226, 482, 1993.
117. Xu, S. et al., Surface-enhanced raman scattering studies on immunoassay, *J. Biomed. Opt.*, 10, 031112, 2005.
118. Grubisha, D.S. et al., Femtomolar detection of prostate-specific antigen: An immunoassay based on surface-enhanced raman scattering and immunogold labels, *Anal. Chem.*, 75, 5936, 2003.
119. Hochuli, E., Dobeli, H., and Schacher, A., New metal chelate adsorbent selective for proteins and peptides containing neighbouring histidine residues, *J. Chromatogr.*, 411, 177, 1987.
120. Hochuli, E. et al., Genetic approach to facilitate purification of recombinant proteins with a novel metal chelate absorbant, *Bio/Technology*, 6, 1321, 1988.
121. Hopp, T.P. et al., A short polypeptide marker sequence useful for recombinant protein identification and purification, *Bio/Technology*, 6, 1204, 1988.
122. Einhauer, A. and Jungbauer, A., The FLAG peptide, a versatile fusion tag for the purification of recombinant proteins, *J. Biochem. Biophys. Methods*, 49, 455, 2001.
123. Chen, Y.T., Holcomb, C., and Moore, H.P., Expression and localization of two low molecular weight GTP-binding proteins, Rab8 and Rab10, by epitope tag, *Proc. Natl. Acad. Sci. USA*, 90, 6508, 1993.
124. Southern, J.A. et al., Identification of an epitope on the P and V proteins of simian virus 5 that distinguishes between two isolates with different biological characteristics, *J. Gen. Virol.*, 72 (Pt 7), 1551, 1991.

125. Evan, G.I. et al., Isolation of monoclonal antibodies specific for human c-myc proto-oncogene product, *Mol. Cell. Biol.*, 5, 3610, 1985.
126. Schwarz, E. et al., The sodium ion translocating oxalacetate decarboxylase of *Klebsiella pneumoniae*. Sequence of the biotin-containing alpha-subunit and relationship to other biotin-containing enzymes, *J. Biol. Chem.*, 263, 9640, 1988.
127. Smith, D.B. and Johnson, K.S., Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase, *Gene*, 67, 31, 1988.
128. Griffin, B.A. et al., Fluorescent labeling of recombinant proteins in living cells with FlAsH, *Methods Enzymol.*, 327, 565, 2000.
129. Lippincott-Schwartz, J. and Patterson, G.H., Development and use of fluorescent protein markers in living cells, *Science*, 300, 87, 2003.

10 Functional Analysis of Protein Interactions Using Surface Plasmon Resonance-Based Microarrays

Alan McWhirter and Stefan Löfås

CONTENTS

INTRODUCTION

Surface plasmon resonance (SPR)-based biosensors are widely used for the characterization of protein interactions. Many thousands of papers from research laboratories and industry are testament to the versatility of the technology, from investigations into the mechanisms controlling fundamental cellular functions to hit selection and quality control in drug discovery processes.[1] SPR detection has also become an established analytical tool in the food industry, where precision and speed are invaluable in monitoring safety and quality.

Although SPR-based biosensors may be used to measure parameters such as the concentration of active protein in a solution or specificity for a particular binding partner, their main advantage over competing interaction technologies such as ELISA or affinity chromatography, is the provision of high resolution kinetics in real time over the entire course of an interaction. This provides a comprehensive and detailed profile of association and dissociation, imparting information about interactions that gives insights into protein function far beyond those that can be inferred from end point assays; the rates of association or dissociation of proteins, for example, enable a complex to be deconstructed in terms of recognition or stability and may form a basis for qualified proposals of interaction models.[2] Further, as the status of an interaction is followed according to changes in mass close to a sensor surface as a molecular complex forms and dissociates, there are no labeling requirements on any of the interacting partners, reducing the possibility of erroneous data arising from steric inhibition of the binding site.

Until now, SPR-based biosensors have been designed to deliver high quality data on a limited number of interactions. There are several reasons, however, to support the design and production of systems with greatly increased capacities for sample throughput. Perhaps the most pressing call for a commercially available protein interaction array is from the proteomics community; with a bewildering amount of novel proteins at hand since the completion of the human genome project, we will clearly be able to put to good use any technology that helps explain to us what they are all for. Additionally, many applications such as antibody screening, hit selection in drug development programs, peptide epitope mapping and even on-line quality control/safety testing during food production are all activities that could see practical and economic benefits from increased sample throughput on an array.

SPR-based arrays have recently appeared on the market, offering an "information-rich" technology rather than one with the onus purely on volume, delivering information on association, dissociation and strength of interaction. The concept of protein arrays is set to progress beyond simply high throughput and into the realm of high information content. Biacore AB (Uppsala, Sweden) offers two protein array products; Flexchip for simultaneous profiling of up to 400 protein interactions and Biacore® A100, which delivers high information content. Both systems are discussed in this chapter.

SURFACE PLASMON RESONANCE: THE TECHNOLOGY BEHIND THE DETECTION PRINCIPLE

SPR Detection

SPR-based biosensors monitor protein interactions in real-time using a label-free detection method. One of the interacting molecules is immobilized on a sensor surface, while the other is injected in solution and flows over the sensor surface. As molecules from the injected sample bind to the immobilized partners and then dissociate, an alteration in refractive index proportional to the change in mass close to the surface is recorded. These events are detected in real time and data are presented as the SPR response plotted against time (Figure 10.1). The plots display the formation and dissociation of complexes over the entire course of an interaction, with the kinetics revealed by the shape of the binding curve.

SPR occurs when polarized light, under conditions of total internal reflection, strikes an electrically conducting gold layer at the interface between media of different refractive index: the glass of a sensor surface (high refractive index) and a buffer (low refractive index). In "classical" Kretchmann configuration SPR detection, a wedge of polarized light, covering a range of incident angles, is directed toward the glass face of the sensor surface. An electric field intensity, known as an evanescent wave, is generated when the light strikes the glass. This evanescent wave interacts with, and is absorbed by free electron clouds in the gold layer, generating electron charge density waves called plasmons and causing a reduction in the intensity

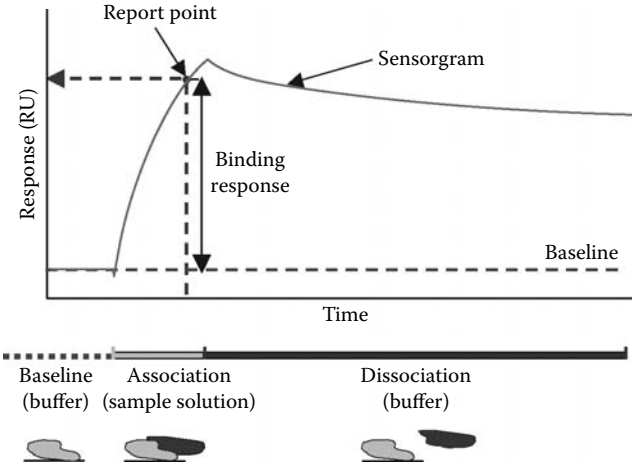


FIGURE 10.1 The plot of the interaction profile provides real-time information about the entire interaction. Binding responses at specific times during the interaction can also be selected as report points.

of the reflected light. The angle at which this intensity minimum occurs is a function of the refractive index of the solution close to the gold layer on the opposing face of the sensor surface.

The refractive index at the interface between the surface and a solution flowing over the surface changes as molecules bind or dissociate close to the sensor surface, altering the angle at which reduced-intensity polarized light is reflected from a supporting glass plane. The change in angle is proportional to the mass of bound material. When sample is passed over the sensor surface, the SPR response increases if the molecules interact. The response remains constant if the interaction reaches equilibrium. When sample is replaced by buffer, the response decreases as the interaction partners dissociate. Complete profiles of recognition, binding and dissociation are generated in real time.

The Development of SPR-Based Arrays

In contrast to most optical and acoustic transducer technologies, SPR is highly applicable to miniaturization and multichannel sensor design. Several optical configurations are conceivable and the first presented approach was based on SPR imaging.[3] In this setting, a metal-coated substrate was illuminated with a collimated beam of monochromatic light and a CCD camera was used to analyze the intensity differences in the reflected light across the surface. The minimum lateral resolution of detectable sample areas depends on the propagation length of the surface plasmon wave, which is in turn determined by the wavelength of the light source and metal types, but in practice can resolve below 50 μm.[4] Spatially differentiated binding reactions occurring on the surface can therefore be read individually and the technique has been used in applications such as lipid layer characterization,[5] the binding of streptavidin to biotinylated self-assembled monolayers (SAM)[6] and monitoring protein: DNA binding kinetics in which 120 different dsDNA strands were arrayed in a 10×12 matrix.[7]

SPR imaging systems are inherently limited in sensitivity, but improvements have been demonstrated by utilizing polarization contrast and patterned SPR structures.[8] Biacore's SPR sensor systems based on the Kretchmann configuration, however, where angle-resolved information is used, are even more sensitive. In the Kretchmann configuration, a wedge of monochromatic light beam is focused on a gold-coated surface and an arrangement of parallel flow channels matches a linear array of sensor areas over a distance of just a few millimeters. Binding reactions in each sensor area are individually interrogated using a CCD-based detector. The original Biacore instruments were designed to measure four sensor areas simultaneously.[9] As described in the following sections, significant design improvements in flow channel configurations have allowed the construction of eight[10] and, in Biacore A100, twenty sensing areas.

In addition, an array configuration based angular scanning of an SPR imaging system has been commercialized.[11] In this Flexchip instrument, based on diffraction grating coupled SPR (GC-SPR), up to 400 sensor areas can be individually and simultaneously interrogated. Real-time based readouts can thereby be acquired, delivering kinetic information from a multiplexed array (see Adapting SPR detection to Large-Scale Protein Arrays).

Flow Cell Systems in SPR-Based Protein Arrays

Interactions in Biacore's classical SPR-based biosensors take place within a flow cell formed by the superimposition of a plastic template engraved with microchannels onto a gold sensor surface. A series of pumps and valves control the flow of liquid through the channels and the steps of immobilization, interactant flow and regeneration are performed while the chip is docked in the apparatus. The flow cells were originally designed to address interactions involving one immobilized partner per flow cell, with the number of interactions in one cycle limited by the number of flow cells. The desire to bring protein arrays to the market has stimulated the design of new flow cells and has resulted in two formats; flow cells designed to accommodate large-scale protein arrays and those based on a novel format known as hydrodynamic addressing.

Adapting SPR Detection to Large-Scale Protein Arrays

To address an extensive protein array spotted with hundreds of immobilized proteins, GC-SPR is used in Flexchip and is a radical departure from the more familiar prism-based Kretchmann systems. This array is capable of delivering simultaneous interaction profiles on 400 spots on one sensor surface. The small coupling angle of the incident light is conducive to multiple imaging and is thus well suited to screening applications. Here, incident polarized light strikes the entire functional face of a finely grated sensor surface enabling simultaneous measurement of interactions on all spots and eliminating errors that could arise from sequential readings. As incident light strikes the sensor surface, interaction profiles are generated from the individual spots (Figure 10.2).

For large protein arrays, various approaches have been taken to immobilize interacting partners on the sensor surface. Techniques developed during the 1990s, principally for DNA arrays, have also been evaluated for SPR detection.[12] Sample deposition by contact or non-contact methods has become the most common alternative and several commercial spotters are available. A general review of protein arrays can be

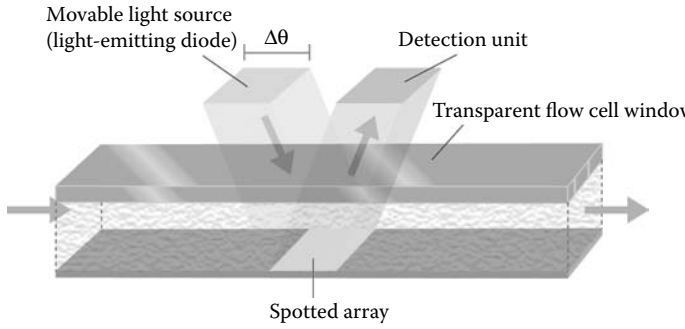


FIGURE 10.2 The Flexchip flow cell setup. A gasketed cell window with an inlet and an outlet valve hermetically seals the sensor surface (on which individual proteins have been spotted) to form a flow cell. The chip is then inserted into the instrument. Sample is injected through a single broad channel thereby interacting simultaneously with all spots on the array.

found in a review by Mann et al.[13] One possible limitation of these approaches is that relatively high protein concentrations are necessary, due to the need for a high surface density of active molecules. Ink-jet or piezoelectric printing devices originally employed for DNA applications can also be used for proteins, both in aqueous and organic solvents. However, careful optimization is needed when these technologies are used for proteins in buffered solutions, as deposits tend to accumulate and clog the ink-jet heads, particularly when solutions containing high concentrations of protein are used. Smearing and uneven intensities across spots have also been reported.

APPLICATION EXAMPLES

Large-Scale Protein Arrays

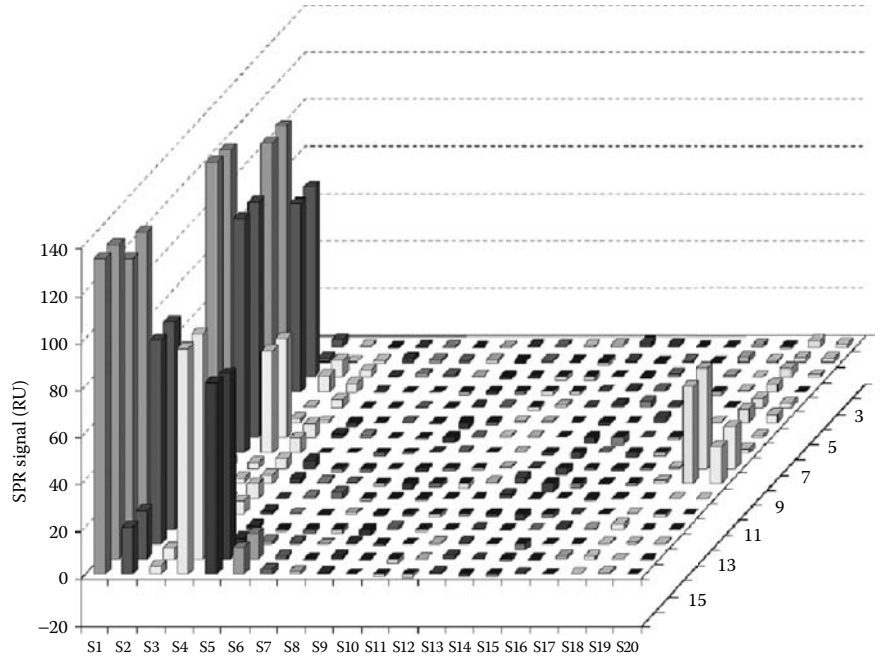
Interactions between proteins can often be probed using a peptide from one of the interacting partners rather than the whole protein. Such an array may be used, for example, by spotting overlapping peptides covering the entire sequence of one interacting partner in order to identify those peptides that have the highest binding activity. The array may then be further applied in an alanine scan to precisely identify the amino acid residues necessary for the interaction (Figure 10.3).

The array may also be used to precisely define how transcription factors bind to DNA by comparing interactions with wild type DNA oligomers to those containing mutations within the consensus sequence. Electrostatic interactions have been reported to influence the association of the transcription factor, NFκB to DNA and this may be mimicked *in vitro* by increasing the ionic strength of the running buffer, a condition that tends to favor specific over non-specific interactions. Interactions may thus be followed at different ionic strengths over a series of runs and the interaction profiles compared (Figure 10.4).

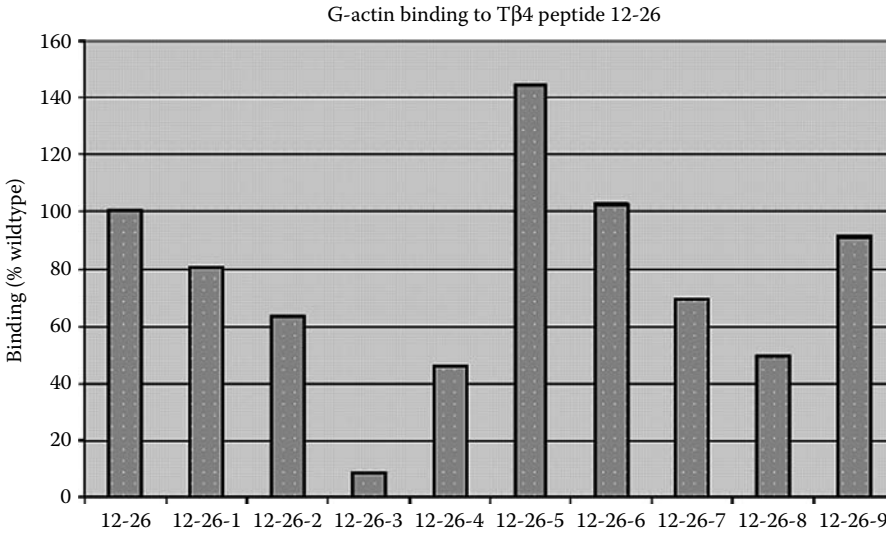
Epitope Mapping

Baggio et al. used Flexchip to probe the binding site for a monoclonal antibody (mAb) on a human class II MHC protein.[2] The location of the epitope was known from previous studies and Flexchip was used to assess the contribution of each residue within the epitope by alanine scanning, measuring the relative binding and kinetics of overlapping parent and mutant peptides from this region. Biotinylated peptides were spotted in triplicate at three different concentrations, all below saturation levels to minimize mAb rebinding and provide reliable kinetic data. Eighteen peptides were assessed, using a total of 162 spots. The data showed that changes in affinity were predominantly effects of increases or decreases in dissociation rates, suggesting that antigen may recognize the MHC protein in a conformation that is then fixed and stabilized on binding; an example of how kinetic data from an array may be interpreted in terms of a molecular interaction model.

Using mAbs that differentiate between open and closed conformations of HLA-DR1, Carven et al. investigated the nature of those structural changes and identified the regions involved.[14] They immobilized many overlapping 20-mer biotinylated peptides — either empty or peptide loaded — spanning the entire HLA-DR1 molecule and



(a)



(b)

FIGURE 10.3 (A) Select which peptides immobilized in an array best bind a protein in solution. (B) Spot peptides containing a single alanine substitution at one amino acid position. Pinpoint which amino acid(s) is (are) critical for the interaction. See color insert following page 236.

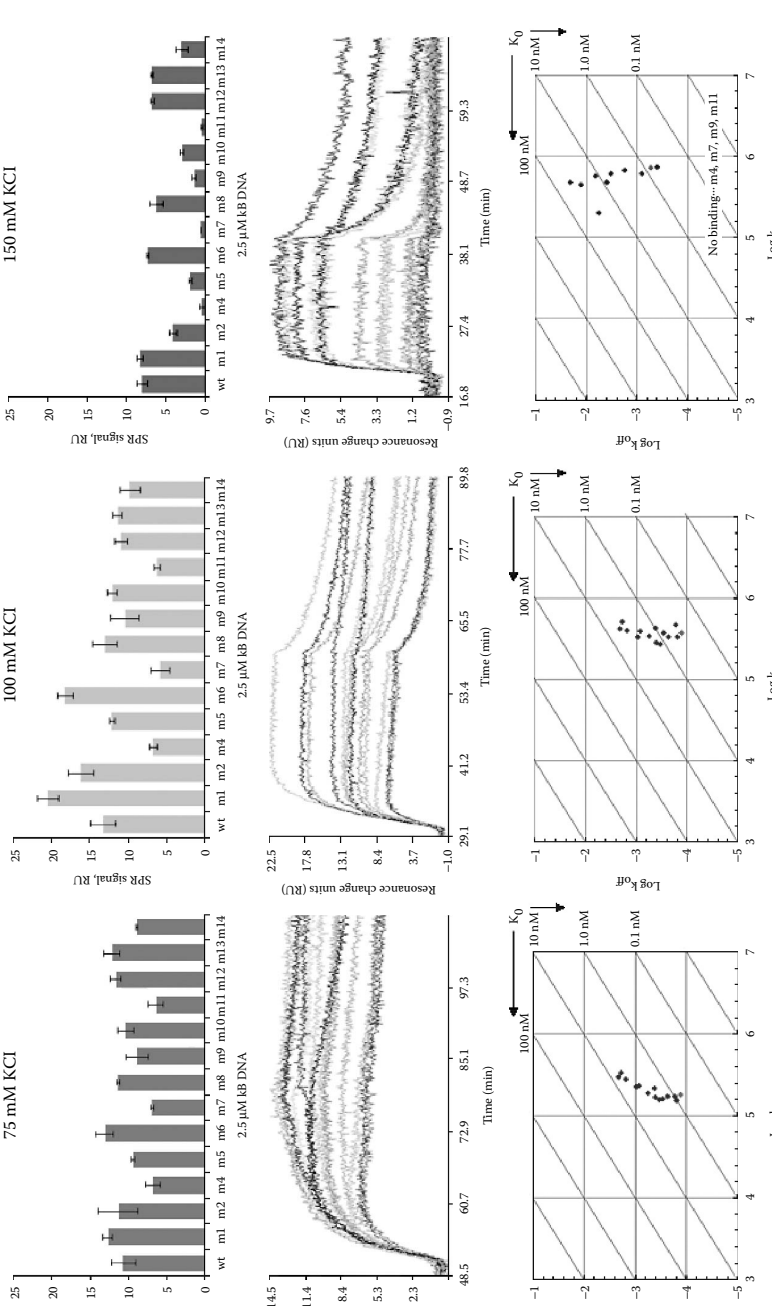


FIGURE 10.4 Interaction profiles involving wild-type and mutant sequences differ at higher salt conditions where alterations in binding specificity are due largely to changes in the dissociation rates. The results suggest that the three guanine bases at positions 1, 2, and 3 of the NFκB binding site contribute most to NFκB binding and stability.

injected mAbs over the surface in end point binding experiments. Using this strategy, it was possible to identify the epitopes to which antibodies bound, possibly inducing the empty peptides to change conformation to the loaded form.

Protein Expression Profiling

Usui-Aoki et al. developed antibody microarrays for protein expression profiling of a ubiquitous protein.[15] Crude adult mouse tissue preparations from brain, spleen, liver, thymus and testis were prepared and a protein A-purified antibody array was constructed to examine protein expression levels in different tissues and to identify tissue expression patterns of several related proteins by injecting crude samples from tissue homogenates over the array.

Kinetic Ranking

Hoet et al. constructed human Fab libraries with a combination of Ig sequences from non-immune donors and engineered regions of CDRs to produce high affinity antibodies without the need for lengthy *in vitro* maturation.[16] The selected target, human tissue kallikrein, was screened for dissociation rates on an array containing 355 recombinant synthetic Fabs using Flexchip. The screen made possible the selection of a number of high affinity antibodies.

Hydrodynamic Addressing Flow Cell Systems

Large-scale protein arrays such as Flexchip are designed to enable the collection of the maximum data set from interactions occurring on spots distributed over a single chip. Arrays from which it is desirable to derive the highest possible *information* content on each interaction, while maintaining high throughput, place different demands on the system design. Hydrodynamic addressing (HA) is a process by which multiple targets may be immobilized on several detection spots in a single flow cell, allowing simultaneous kinetic analysis of interactions (Figure 10.5). As there is no lag time between interactions, highly accurate reference subtraction allows the measurement of very rapid kinetics. Further, by immobilizing several targets in one flow cell, binding properties may be directly compared under optimal experimental conditions.

By adjusting the relative flow at the two inlets (one for the immobilized partner and the other for buffer), liquid can be directed to one or other of the addressable detection spots. The flow cell design allows rapid switching of flow between buffer and target and the transverse arrangement of the detection spots ensures that access of sample to all spots is simultaneous.

Biacore A100 has four parallel independent HA flow cells. Up to five different proteins can be immobilized in each. For assays requiring maximum sample throughput, identical immobilizations can be performed in all four flow cells, allowing four different samples to be analyzed in parallel during each analysis cycle. In assays where information output per sample is more important, up to twenty different interactants can be immobilized in the four flow cells and one sample per cycle is injected in parallel over all flow cells. Assays can be run in two configurations depending on the level on information required. The four flow cell, five spot per flow cell configuration enables up to 3800 interactions to be monitored in a 24 hour

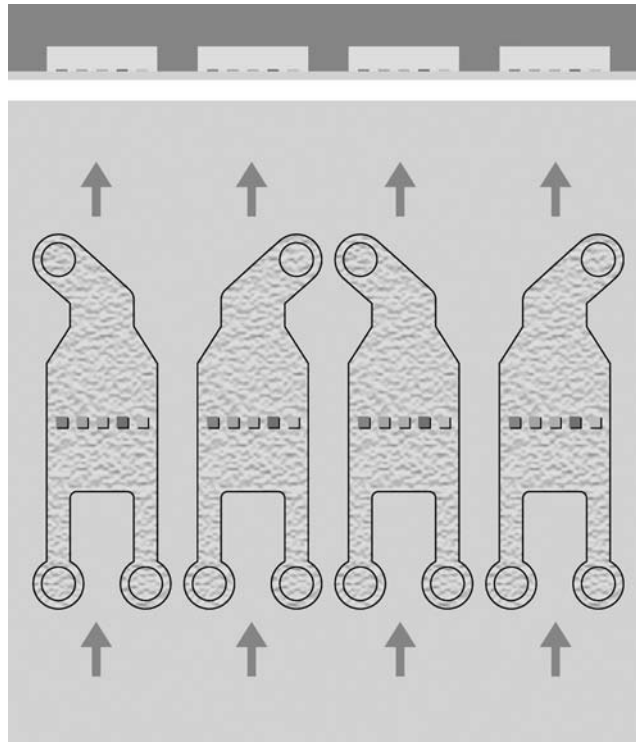
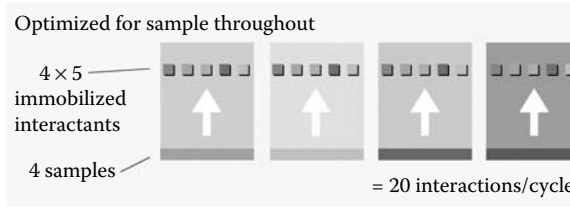


FIGURE 10.5 The flow cell system in Biacore A100, comprising four parallel HA flow cells. With five detection spots arranged across each flow cell, up to twenty interactions can be characterized during each analysis cycle.

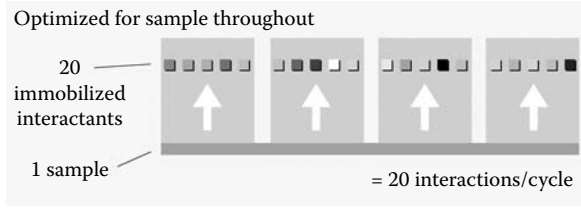
run, with a selectable configuration optimized either for maximum number of samples, or for maximum information per sample (Figure 10.6).

Productivity in biotherapeutic development may benefit from access to a high throughput array system that delivers kinetic data. For example, the development of mAbs is a complex and time-consuming process, involving generation, maintenance and screening of thousands of hybridoma clones. Early identification of those hybridomas that produce the best candidate antibodies is a critical step in successful, cost-efficient development. Rapid kinetic screening of many hundreds of hybridomas would efficiently enable selection of those candidates with the required kinetic profiles, discriminating between equal-affinity mAbs based on kinetic properties that are crucial for clinical success.

Even the most carefully designed and constructed biotherapeutics may be sensed as foreign proteins by the patient, causing an unwanted antibody response. The immunogenicity of newly developed drugs and vaccines is one area that could benefit from an array system in which serum antibody responses could be characterized. SPR-based protein arrays allow for the detection of potentially clinically relevant low/medium affinity antibodies, generating data on isotype, subclass specificity and kinetics from a single system using low quantities of sera.



(a)



(b)

FIGURE 10.6 Maximum sample throughput configuration with identical immobilizations in all four flow cells (A) and maximum information per sample (B), with up to twenty different interactants immobilized in the four flow cells.

TECHNOLOGICAL DEVELOPMENTS FROM THE LITERATURE

Immobilization of Interaction Partners

One challenge in constructing reliable protein arrays is the requirement to immobilize interacting partners at a defined density and in a uniform orientation. Although direct immobilization strategies such as amine coupling or aldehyde coupling are widely used in classical (i.e., non-array) SPR detection, they may not be suitable for arrays due to the difficulty in controlling for consistent orientation of all immobilized partners across the entire array. Capture of a protein by site-specific biotinylation on a surface prepared with streptavidin, so that the proteins adsorb to the surface in a single and predicted orientation is one alternative immobilization option.

Kwon et al. addressed the issue of spotting antibodies on an array in a uniform orientation using a method that exploits the activity of the enzyme, cutinase, a serine esterase that induces the formation of a covalent linkage between proteins and phosphonate groups on the sensor surface.[17] A fusion protein of antibodies and cutinase was captured on a SAM presenting a phosphonate capture molecule, allowing the immobilized antibodies to retain their affinity and selectivity for their targets. This method also made it possible to control the density of captured antibodies according to the density of the phosphonate groups on the sensor surface. As this capture layer is refractory to non-specific adsorption of proteins, it was possible to measure interactions with complex proteins in solution or crude biological matrices.

A further issue is to decide whether it is best to immobilize whole antibodies or to use engineered derivatives such as single domain antibodies, scFv fragments, Fab fragments or, alternatively, protein scaffolds such as fibronectin type III domains, which can be engineered for optimal protein binding. The advantages of engineered derivatives are that they are smaller, can be easily expressed in bacteria and can be optimized for stability, selectivity, and affinity.

As the activity of proteins is often compromised during the immobilization process due to denaturing or partial masking of the binding site, Peluso et al. compared random, direct panning of proteins with oriented capture via full size mAbs and Fab fragments.[18] Analyte binding was frequently improved using one of the tested capture methods, generating a more active surface. They tested randomly biotinylated IgG, IgG biotinylated on carbohydrate attachments, IgG biotinylated on the Fc domain, oriented Fabs and randomly biotinylated Fabs. Of the capture strategies tested, oriented Fab fragments were most frequently the optimal method in terms of retained activity of the captured molecule and because of the flexibility in selecting the density of the binding partner on the sensor surface — an important consideration in performing kinetic analyses — due to their comparatively small size.

The carboxymethylated surfaces of the most commonly used sensor chips provided by Biacore, however, largely overcome these problems of retention of activity. This is possibly because although these surfaces are ostensibly solid phase, the dextran linkers to which the carboxyl groups are attached, and which are open to amine, thiol or aldehyde coupling, are of sufficient length and flexibility to mimic a fluid environment and allow the immobilized partner a considerable degree of entropic freedom (Figure 10.7).

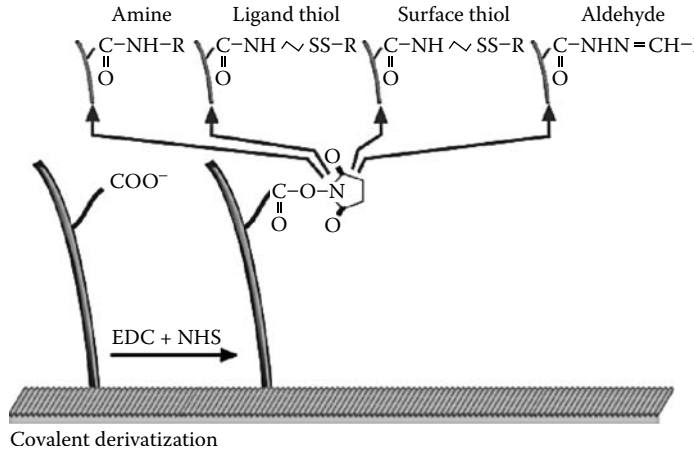


FIGURE 10.7 The carboxymethylated sensor surface. Dextran linkers provide a flexible structure to which carboxyl groups are attached. Molecules coupled to these groups, although securely tethered, maintain a large degree of rotational and lateral freedom, allowing the interaction to occur in an environment that mimics a fluid milieu more closely that provided by typical solid phase platforms.

Prospects for an SPR-MS Array

One of the most intriguing applications of SPR-based biosensors is the capture and characterization of unknown interacting partners from complex biological mixtures and the subsequent recovery of the bound molecules for identification using mass spectrometry (MS); there are many papers in the literature on this subject.[19–21] What are the possibilities for extending this technology to a multitude of captured molecules on a protein array? Nedelkov and Nelson have demonstrated the feasibility of the principle by immobilizing six different antibodies in a single flow cell of a Biacore instrument and probing with well-characterized and commercially available binding partners.[22] The work shows that readouts of parallel protein interactions in one flow cell are possible and that structural features of multiple interactants can be resolved by MS. Functional analysis and identification by MS of many multiplexed proteins from very limited amounts of starting material may one day be possible in an SPR-based protein array.

Carbohydrate Arrays

Although proteomics may be intuitively regarded as the linear successor to genomics, proteins are not alone in regulating or supporting biological processes. To emphasize this truism, it has become popular for scientists to attach the suffix "omics" to their own field of interest and thus we have, for example, "transcriptomics" to denote the study of the full complement of mRNA species expressed under specified conditions, or "lipidomics," the global analysis of how lipids interact with genes and proteins to determine cellular functions. Carbohydrate research, long the domain of biochemists with little opportunity for recourse to technologies like PCR or bacterial expression, may also soon be amenable to array technologies. Automated oligosaccharide synthesizers are finally opening the field to the possibility of producing carbohydrates in sufficient quantity and purity to make arrays feasible, a particularly attractive opportunity as miniaturized array technologies tend to be frugal in their consumption of reagents. The opportunity is certainly attractive as there are already numerous carbohydrate vaccines and drugs on the market. A "glycomics" approach mediated by the availability of a carbohydrate array may thus pave the way to many new promising therapeutics. Werz and Seeberger[23] and Ratner et al.[24] have written interesting articles on how far we can expect to go in terms of parallel thinking with protein chips when we consider the development of SPR-based carbohydrate arrays.

Peptide Nucleic Acid Arrays

There are few reports in the literature in which SPR-based biosensors have been used in a clinical setting to discriminate between homozygous and heterozygous individuals with genetically defined hereditary diseases. A more rapid assay than DNA sequencing that is sufficiently specific to detect disease-causing mutations defined by a single mismatch would be of considerable interest. Peptide nucleic acids (PNAs) have already been used in a Biacore assay to identify specific point mutations in PCR-generated targets derived from patients with cystic fibrosis, the

most common autosomal lethal disease in Caucasians.[25] PNAs are synthetic analogues of DNA in which the sugar-phosphate backbone is replaced by peptide linkages. They are of potential interest as reagents because hybridization of PNAs to complementary DNA is efficient, generating Watson-Crick double helices and are largely independent of the influence of structural features on the target DNA. PNAs are predicted to bind to single stranded PCR products more efficiently than oligonucleotides as the formed PNA: DNA duplexes have a higher melting temperature than DNA: DNA duplexes. Further, PNAs are resistant to nucleases and as they are not negatively charged, are not repelled by a DNA partner. Brandt and Hoheisel raise the possibility of constructing SPR-based PNA arrays[26] and the idea is particularly attractive for the type of screening applications in a clinical setting as proposed and developed by Roberto Gambari and Giordana Feriotto and colleagues at Ferrara University, Italy (see Feriotto et al.[27] and references therein).

Affibody Arrays

For truly global expression profiling, a great number and variety of antibodies would be needed. Alternative strategies include the use of Fab fragments generated from phage display libraries or affibodies, affinity proteins based on the 58 amino acid three helix bundle protein scaffold of the Z domain from staphylococcal protein-A and selected from a combinatorial protein library.[28] They have several advantages over antibodies in that they are small (always an advantage in an array), can be expressed in high yields and are easily synthesized and engineered. Renberg et al. showed that the specific activity of affibodies was improved compared with direct coupling when they were co-expressed with a biotin tag separated by a short spacer or when the affibodies were coupled to the sensor surface using thiol coupling.

SPR-Based Chemical Microarrays

Low-affinity screening using tethered drug-fragments on chemical microarrays is a highly promising approach for the rapid discovery and optimization of small molecule inhibitors. Currently, the Plasmon Imager® devices being pioneered by Graffinity AG are capable of routinely processing about 10,000 measuring points, delivering information on interaction affinities. Dickopf et al. used this technology in their search for new inhibitors of factor VIIIa, a serine protease involved in many pathological processes and consequently an attractive drug target.[29] This enzyme is partly characterized by the possession of a deep pocket, which is already the target of benzamidine-based inhibitors. These compounds, however, are not optimal for factor VIIIa as the S1 pocket of this protein contains a serine residue in place of alanine at the critical binding site. Candidates were thus selected firstly using a virtual screening process for small molecules to probe areas within the binding site, based on crystallographic data of factor VIIIa in complex with other known inhibitors. Filtering criteria such as molecular weight, number of rotatable bonds and lack of reactive groups were used to reduce 30,000 initial candidates to 1500. The candidates were conjugated to spacer molecules and immobilized on a microarray on top of a SAM using a variety of coupling chemistries. Hits were selected after injecting factor

VIIIa over the prepared surface and the signal to noise ratios were improved by subsequent injection of an anti-factor VIIIa antibody. The interactions were confirmed by functional studies of inhibition of enzymatic activity. The authors suggest that it may be ultimately be possible to fuse fragments to larger factor VIIIa inhibitors to make a highly effective drug. The fragments used in the study by Dickopf et al. had molecular weights of around 200 Da and demonstrated that SPR-based arrays are suitable for screening weak interactions involving low molecular weight molecules. Biacore A100 is fully capable of these performance standards in direct binding mode, with the protein target, instead of the low molecular weight binding partner, immobilized on the sensor surface.

Protein Microarrays within a Continuous Flow Cell

Wegner et al. have demonstrated that a protein microarray may be constructed using a continuous serpentine flow cell system and Kretchmann-based SPR detection.[30] The flow cell was constructed by etching a channel on a gold surface on top of an aluminum layer supported on a glass slide. The gold surface was modified with an amine-terminated SAM followed by the addition of a bifunctional linker, *N*-succinimidyl 3-(2-pyridyldithio) propionamido (SPDP), creating a disulphide-terminated surface. This modification makes it possible for cysteine-modified peptides to be covalently attached to the sensor surface via a thiol disulphide reaction on about 50 discrete areas throughout the length of the flow cell (see also Kanda et al.[31] for an illustration of how the targeted areas are formed in the flow cell). The system was used to study the relationship between a variety of S protein peptides and S protein, delivering kinetic and affinity data from several interactions.

DISCUSSION

Functional Protein Arrays: Fulfilling the Promise of Proteomics

The success of DNA arrays has perhaps led to misguided expectations that protein arrays will be on a similar scale, with potentially tens of thousands of proteins immobilized on a single chip. Although the advent of such a tool would be hailed in the proteomics field as a weapon akin to a battering ram, clearing the way for rapidly mapping protein interaction networks, the development of protein arrays is unfortunately a more complex undertaking than the DNA counterpart; proteins are more difficult to handle than DNA, as post-translational modifications vital for functionality are seldom preserved in the course of the amplification steps required to obtain reagents in sufficient quantity and purity. Consideration must also be given to variables such as immobilization conditions, orientation and the possibility that the immobilization process may impede or conceal the very binding site of interest. Further complications include the desirability of immobilizing proteins efficiently and at precise concentrations (important for obtaining meaningful association rates) across the entire array. Proteins are also less discriminatory in their choice of binding partners than DNA and so non-specific adsorption

to both the sensor surface and other proteins in a complex mixture such as clinical samples or hybridoma supernatants may complicate the interpretation of results from a multiplexed array.

Consider the type of information we wish to obtain from a protein array. Firstly, protein arrays should be regarded as supplementary technologies, rather than alternatives to existing methods such as 2-D gel electrophoresis (protein detection) or mass spectrometric follow-ups (protein identification). It should also be remembered that 2-D gels are limited by restrictions in throughput and are relatively insensitive; most of the interesting proteins expressed in response to physiological stimuli are present transiently, locally and at very low concentrations.[32] Besides high sensitivity, the wealth of information revealed by SPR-based analysis about an interaction, if used to its full potential, goes far beyond mere detection. While end point binding can be measured if required, a detection technology that monitors entire interactions and delivers data on recognition (association rate constant, k_a), stability (dissociation rate constant, k_d) and strength (affinity constant, K_D) means that we may also construct arrays designed to deliver information about protein function in which a restricted population of proteins in a cell, or a group of antibodies or peptides occupy defined spots on an array. SPR-based arrays, therefore, are likely to be smaller than DNA arrays. Although their development is partly governed by the fragility of proteins removed from their normal cellular environment, it is important that we do not obsessively cling to an intuition that says, "Because the DNA guys did it, so must we."

Arguably, the potential of proteomics will only be realized when researchers outside the confines of the proteomics community itself begin to routinely use the data to understand what all these thousands of proteins actually do! While the proteome itself is finite, the range of protein functions as defined by the manner in which they interact with each other (and with carbohydrates, lipids, nucleic acids and small organic molecules like vitamins or nucleotides) and how these complexes form and decay — the field of functional proteomics — is practically boundless. A rheumatologist, for example, will (or certainly should) be interested in identifying binding partners of proteins uniquely expressed in rheumatoid arthritis or learning whether the interaction profiles of key proteins in the rheumatoid arthritis proteome differ significantly from those in the normal proteome. However, he is most likely to develop faith in the benefits of pursuing a proteomics strategy if he has access to a technology such as a functional protein array that enables him to answer a far more explicit question, namely; What are the implications of these interaction patterns and how do they impact on severity, disease progression and therapeutic options?

In other words, the questions asked and addressed by recourse to a functional protein array are more focused and place more emphasis on information content from a limited number of selected interactions rather than a myriad yes/no answers from a crude screen. A functional protein array then, is likely to be attractive to those who wish to make use of the vast repository of data from proteomics initiatives to make proteomics itself a functional approach that solves real problems. It is the job of researchers within the proteomics community to alert those outside to the benefits of this global way of thinking and working.

REFERENCES

1. Rich, R.L. and Myszka, D.G., Survey of the year 2003 commercial optical biosensor literature, *J. Mol. Recognit.*, 18, 1, 2005.
2. Baggio, R. et al., Induced fit of an epitope peptide to a monoclonal antibody probed with a novel parallel surface plasmon resonance assay, *J. Biol. Chem.*, 280, 4188, 2005.
3. Rothenhäusler, B. and Knoll, W., Surface-plasmon microscopy, *Nature*, 332, 615, 1988.
4. Zizlsperger, M. and Knoll, W., Multispot parallel on-line monitoring of interfacial binding reactions by surface plasmon microscopy, *Progr. Colloid Polym. Sci.*, 109, 244, 1998.
5. Hickel, W. and Knoll, W., Surface plasmon microscopy of lipid layers, *Thin Solid Films*, 187, 349, 1990.
6. Piscevic, D., Knoll, W., and Tarlov, M.J., Surface plasmon microscopy of biotin-streptavidin binding reaction on UV-patterned alkanethiol self-assembled monolayers, *Supramol. Sci.*, 2, 99, 1995.
7. Shumaker-Parry, J.S., Aebersold, R., and Campbell, C.T., Parallel, quantitative measurement of protein binding to a 120-element double-stranded DNA array in real time using surface plasmon resonance microscopy, *Anal. Chem.*, 76, 2071, 2004.
8. Piliarik, M., Vaisocherová, H., and Homola, J., A new surface plasmon resonance sensor for high-throughput screening applications, *Biosens. Bioel.*, 20, 2104, 2005.
9. Löfås, S. et al., Bioanalysis with surface plasmon resonance, *Sens. Act. B: Chemical*, 5, 79, 1991.
10. Situ, C. et al., On-line detection of sulfamethazine and sulfadiazine in porcine bile using a multi-channel high-throughput SPR biosensor, *Anal. Chim. Acta*, 473, 143, 2002.
11. Brockman, J.M. and Fernandez, S.M., Grating-coupled surface plasmon resonance for rapid, label-free, array-based sensing, *Am. Lab.*, 33, 37, 2001.
12. Shumaker-Perry, J.S. et al., Microspotting streptavidin and double-stranded DNA arrays on gold for high-throughput studies of protein-DNA interactions by surface plasmon resonance microscopy, *Anal. Chem.*, 76, 918, 2004.
13. Mann, C.J., Stephens, S.K., and Burke, J.F., Production of protein microarrays, in *Protein Microarray Technology*, Kambhampati, D., Ed., Wiley-VCH, Weinheim, 2004.
14. Carven, G.J. et al., Monoclonal antibodies specific for the empty conformation of HLA-DR1 reveal aspects of the conformational change associated with peptide binding, *J. Biol. Chem.*, 279, 16561, 2004.
15. Usui-Aoki, K. et al., A novel approach to protein expression profiling using antibody microarrays combined with surface plasmon resonance technology, *Proteomics*, 5, 2396, 2005.
16. Hoet, R.M. et al., Generation of high-affinity human antibodies by combining donor-derived and synthetic complementarity-determining-region diversity, *Nature Biotechnol.*, 23, 344, 2005.
17. Kwon, Y. et al., Antibody arrays prepared by cutinase-mediated immobilization on self-assembled monolayers, *Anal. Chem.*, 76, 5713, 2004.
18. Peluso, P. et al., Optimizing antibody immobilization strategies for the construction of protein microarrays, *Anal. Biochem.*, 312, 113, 2003.
19. Borch, J. and Roepstorff, P., Screening for enzyme inhibitors by surface plasmon resonance combined with mass spectrometry, *Anal. Chem.*, 76, 5243, 2004.

20. Shang, C. et al., Mass spectrometric analysis of posttranslational modifications of a carrot extracellular glycoprotein, *Biochemistry*, 43, 6281, 2004.
21. Swietnicki, W. et al., Novel protein-protein interactions of the Yersinia pestis type III secretion system elucidated with a matrix analysis by surface plasmon resonance and mass spectrometry, *J. Biol. Chem*., 279, 38693, 2004.
22. Nedelkov, D. and Nelson, R.W., Design and use of multi-affinity surfaces in biomolecular interaction analysis-mass spectrometry (BIA/MS): A step toward the design of SPR/MS arrays, *J. Mol. Recognit.*, 16, 15, 2003.
23. Werz, D.B. and Seeberger, P.H., Carbohydrates as the next frontier in pharmaceutical research, *Chem. Eur. J.*, 11, 3194, 2005.
24. Ratner, D.M. et al., Tools for glycomics: Mapping interactions of carbohydrates in biological systems, *ChemBioChem*, 5, 1375, 2004.
25. Feriotto, G. et al., Peptide nucleic acids and biosensor technology for real-time detection of the cystic fibrosis w1282x mutation by surface plasmon resonance, *Lab. Invest.*, 81, 1415, 2001
26. Brandt, O. and Hoheisel, J.D., Peptide nucleic acids on microarrays and other biosensors, *Trends Biotechnol.*, 22, 617, 2004.
27. Feriotto, G. et al., Real-time multiplex analysis of four bb-thalassemia mutations employing surface plasmon resonance and biosensor technology, *Lab. Invest.*, 84, 796, 2004.
28. Renberg, B. et al., Affibody protein capture microarrays: Synthesis and evaluation of random and directed immobilization of affibody molecules, *Anal. Biochem.*, 341, 334, 2005.
29. Dickopf, S. et al., Custom chemical microarray production and affinity fingerprinting for the S1 pocket of factor VIIa, *Anal. Biochem.*, 335, 50, 2004.
30. Wegner, G.J. et al., Real-time surface plasmon resonance imaging measurements for the multiplexed determination of protein adsorption/desorption kinetics and surface enzymatic reactions on peptide microarrays, *Anal. Chem.*, 76, 5677, 2004.
31. Kanda, V. et al., Label-free reading of microarray-based immunoassays with surface plasmon resonance imaging, *Anal. Chem.*, 76, 7257, 2004.
32. Kodadek, T., Protein microarrays: prospects and problems, *Chem. Biol.*, 8, 105, 2001.

11 Leaving the Surface Behind: At the Intersection of Protein Microarrays and Mass Spectrometry

Darrell P. Chandler, Daniel S. Schabacker, Sergei Bavykin, and Igor M. Gavin

CONTENTS

INTRODUCTION

The discovery of protein biomarkers enables early detection and accurate prognosis of many diseases,[1] and helps identify the response of an organism and/or host to treatment. As clearly articulated elsewhere in this book, many new technologies are now converging within the field of proteomics and laying the groundwork for systems biology.[2] Several government programs and numerous corporate investments are also supporting research to provide the instrumentation and methods for analyzing entire proteomes through time and space, potentially at the level of a single cell.[3] Within the context of systems biology and the thesis of this book, we subscribe to the concept of functional proteomics as defined by MacBeath — namely, to understand and predict the function of every protein in a given organism

(as opposed to study the entire protein expression profile of a particular cell at any given time or under a specific set of environmental conditions). If the ultimate goal of functional proteomics and systems biology is taken seriously, then, we must acknowledge that to detect a protein is not the same as to identify or characterize a protein; that a protein complex is not necessarily a functional machine; that a cell is not an organ or community; and cell cultures are not necessarily representative of the host or an organism in its natural environment. Given this perspective, additional and/or alternative technology developments are still required to meet the functional proteomics challenge, whether for systems biology or drug discovery applications. The purpose of this chapter is therefore to describe recent developments at the intersection of protein arrays and mass spectrometry (MS), and how integrated protein array-MS technologies might be applied to challenging proteomics questions that may otherwise be intractable using historical or commercial, off-the-shelf systems.

THE CHALLENGE

Clearly, many analytical methods and instruments are necessary in order to meet the information demands of predictive biology and drug discovery. Currently, high-throughput versions for two-dimensional gel electrophoresis, yeast or bacterial two-hybrid screening, liquid chromatography, and matrix-assisted laser desorption/ionization (MALDI) or electrospray ionization (ESI) mass spectrometry (MS) methods dominate the technical landscape. In the near term, new methods for the "top-down" (intact proteins) or "bottom-up" (peptides) detection and characterization of proteins and protein function will continue to emerge, as indicated in several recent reviews.[2,4–8] In this context, both mass spectrometry and protein chip technologies appear to have important (and complementary) roles. The technical challenge facing postgenomic biology, however, is much more daunting than the pure scalability issue associated with the genome sequencing programs that inspired the current molecular revolution. As stated by Tyers and Mann,[9] proteomics (and hence, the underlying technologies) must deal with the unavoidable problems of limited and variable sample material, sample degradation, 10^6-fold dynamic range in protein abundance, a multitude (>200) of post-translational modifications that affect protein activity, and almost boundless environmental, developmental, and temporal specificities and perturbations.

Thus, "by all criteria, current instrumentation is far from optimal, in part because manufacturers have not yet had the necessary lead time to build machines and associated hardware that are perfectly tailored to protein analysis."[9] For example, two-dimensional gels are often criticized for their cumbersome nature, limited depth of coverage, detection limits and bias against membrane, acidic or basic proteins.[10] Mass spectrometry is generally ill-suited for the analysis of complex samples, and sample preparation is difficult to automate.[9] Liquid chromatography and capillary electrophoresis sample preparation methods suffer from peak capacity limitations and variable elution times/properties.[5] Coupled with ESI techniques, weak protein complexes can be disrupted before analysis.[11] The exquisite sensitivity and dynamic range of mass spectrometers place an analytical premium

on reproducible, high-throughput sample cleanup and preparation.[10] Global peptide mapping with accurate mass tags and high-resolution FTICR mass spectrometers provides a comprehensive view of protein expression,[12] yet a catalogue of protein "parts" does not by itself elucidate protein function or cellular networks (analogous to criticisms frequently leveled against expression profiling microarrays[7,13]).

Protein array technologies are rapidly developing and, in many respects, directly detect protein–protein interactions. For example, a whole-proteome yeast chip was developed and used to elucidate functional activity of 5800 different proteins.[14] However, planar chip surfaces are known to denature proteins, induce steric constraints on binding efficiency and protein function, and result in nonuniform spot morphologies that are difficult to quantitatively analyze across multiple chips.[8] Antibody, single-chain antibody and antibody-mimic affinity reagents for capturing or detecting proteins on microarray (or other) surfaces suffer from a lack of specificity and ability to detect posttranslational modifications. Only a fraction of antibodies "behave" or function properly on planar surfaces.[15] Labeling techniques to optically visualize protein–protein (and other protein-based) interactions may induce conformational changes in target proteins that destroy function or activity,[4,10,16] and sandwich detection schemes increase demands on high-throughput production of paired, high-quality antibodies for every target or complex.[8] Even mundane issues of spot homogeneity, protein distribution within immobilized spots, standards and reproducibility present significant data extraction and analysis problems for protein array methods.[17]

Thus, while recent technology developments and demonstrations are exciting, basic biological and technical challenges limit the extent or application of existing systems for functional proteomics. In particular, detecting macromolecules and characterizing their interactions, especially in complex mixtures, ultimately requires indicator-free methods[18] that offer the potential to simultaneously detect and identify previously uncharacterized molecules that interact with immobilized probes. As described in the remainder of this chapter, key developments in two disparate technical domains (three-dimensional microarrays and mass spectrometry) provide one technical solution and direction to address the fundamental technical challenges posed by functional proteomics questions and information requirements.

MALDI MASS SPECTROMETRY

A typical matrix assisted laser desorption ionization (MALDI) mass spectrometry experiment is performed by co-dispersing and crystallizing relatively simple analyte mixtures with a laser-absorbing matrix material on a target plate. The matrix (e.g., ferulic acid) absorbs pulsed laser energy, acts as a medium for energy transfer to the analytes, and thereby induces desorption and ionization of target molecules. The desorbed ions are separated in a mass analyzer (i.e., time-of-flight [TOF]) based on their mass-to-charge ratio (m/z) and sequentially detected and identified via their arrival time at the detector. A key advantage of MALDI over ESI is the static nature of the ionization technique, a property that makes MALDI more amenable to array

detection than the continuous flow of ESI. Second, singly charged ions are typically produced for intact proteins and most peptides. Therefore, coupling MALDI to linear TOF instruments (with a near limitless mass/charge range) allows for the observation of intact molecular ions in excess of 200,000 Daltons (e.g., intact PCR products or DNA fragments).

The most common application of MALDI-MS in proteomics is the identification of intact proteins after separation by one or two dimensional gel electrophoresis.[5,19,20] Multidimensional liquid separations (e.g., tandem liquid chromatography) have also proven effective for separating protein mixtures prior to tryptic digestion and MS analysis.[21] Many groups are developing on-chip protein separation systems based on retentate chromatography,[22] solid-phase microextraction,[23] capillary electrophoresis or other chip-scale liquid chromatography techniques,[24] usually in conjunction with ESI-MS. The direct analysis of intact proteins separated within polyacrylamide gels is also possible[25]; combined with "in gel" protein digestion, it is possible to reduce sample handling and loss typically associated with protein sample preparation.[26–28] Intact proteins or peptides can also be transferred to membrane surfaces for subsequent addition of matrix and MALDI-MS analysis as an alternative to liquid extraction and separations.[29]

Conceptually, the most direct and efficient MS separation technique for complex biological solutions is based on the affinity purification principle, with some recent developments demonstrating how on-chip affinity separations can be coupled with MALDI mass spectrometry. For example, Ciphergen commercially markets Surface Enhanced Laser Desorption/Ionization technology (SELDI; reviewed in Merchant and Weinberger[30]), which is predicated on the chromatographic separation of protein mixtures on low-density MALDI plates or to single, immobilized proteins as "bait." Metal ions, antibodies and lectins have also been attached to SELDI surfaces and used to selectively isolate proteins on a low-density (6-spot) chip prior to MALDI-MS analysis.[31–34] The basic SELDI surfaces can even be used to identify specific strains of bacteria.[35] On-plate digestion of captured proteins has also been demonstrated for peptide fingerprinting analysis without further sample handling steps.[36] However, the limitations discussed above for biomolecular interactions on planar (array) surfaces are also applicable here, including limited probe (protein, antibody, analyte) density and the potential for protein denaturation.

THREE-DIMENSIONAL PROTEIN ARRAYS

The last decade has seen considerable advances in microarray manufacturing technologies and applications, which now include tissue,[37] living cell,[38] peptide,[39] antibody/antigen,[40–42] protein,[14,43] carbohydrate[44] and small molecule arrays.[45] Three-dimensional gels were developed, in part, to overcome steric and probe density constraints imposed by two-dimensional surfaces while preserving the functional integrity of immobilized biomolecules.[46–48] Several leading bioscience companies (Perkin Elmer, Amersham, Schleicher & Scheull) presently manufacture and sell continuous gel- or membrane-layered glass slides for protein array applications. The three-dimensional gel element array platform was developed in

the early 1990s by Andrei Mirzabekov and colleagues, with the first gel element protein arrays reported in 1997.[49] A copolymerization technique for protein array fabrication has also been described.[50] By 2000, gel element arrays were evaluated in numerous functional protein assays,[51] including demonstrated activities and assays utilizing DNA-polymerase, DNA ligase and polynucleotide kinase.[52,53] Gel element arrays even support within-gel PCR amplification.[54] Thus, three-dimensional gel elements retain protein and enzymatic activity, and provide kinetic (binding or enzyme activity) data of relevance to systems biology or drug discovery. Given the commercial availability of continuous gel-layered substrates for microarray manufacture, the question then becomes, what value is added by using discreet gel elements rather than planar surfaces or a continuous gel layer for functional proteomic assays?

The fundamental difference between three-dimensional gel element arrays and other substrates is that individual polymeric gel elements literally create a high density array of three-dimensional "test tubes." Probes (or "bait") are covalently crosslinked to the polymer backbone instead of a solid substrate, and do not diffuse out of the gel matrix. Thus, each gel element retains a solution-phase test environment throughout manufacture and testing; immobilized molecules are randomly but uniformly oriented and available for interaction; and biomolecular interactions proceed according to well understood, liquid-phase thermodynamics and kinetics without uncharacterized or unknown surface effects (see, e.g., reviews in[8,13,14] and experimental data in[15,16]).

A common criticism of gel substrates is the (potentially) restricted pore size of the matrix.[8] Small pores can limit diffusion and preclude large complexes from either entering into or forming within the gel. The polyacrylamide gel formulation described by Arenkov[51] allowed a 300 kDa antibody-antigen-antibody complex to form within the matrix, and a 290 kDa enzyme (GUS) was immobilized and retained functional activity within the gel. To improve the range of applications for gel element microarrays and address the pore size constraints imposed by polyacrylamide, a number of alternative polymers are under development that allow one to tune the porosity of the gel element (from tens to hundreds of nanometers) by adjusting the concentration of polymer, cross-linking agents and solvents during photopolymerization. Several of Argonne's more promising gels, for example, have a nominal pore size of 300 nm, or 1/3 the diameter of an typical bacterium.

Initial functional tests with the alternative matrices gels are encouraging. As shown in Figure 11.1, the new gel element compositions have (at least) a five- to sevenfold greater protein binding capacity than commercially available substrates, including Hydrogel. The upper limit on protein binding capacity has not been determined, yet repeated application of concentrated protein solutions to the same gel element indicate that gel elements can continue to absorb and immobilize protein, whereas other substrates do not. The detection limit for the new gels in a standard cytokine sandwich immunoassay is 100 fg ml^{-1} of analyte (Figure 11.2), with a linear response over (at least) 4-logs of target concentration (not shown). Given the exceptionally high binding capacity of gel elements relative to other matrices and the demonstrated sensitivity for analyte detection,

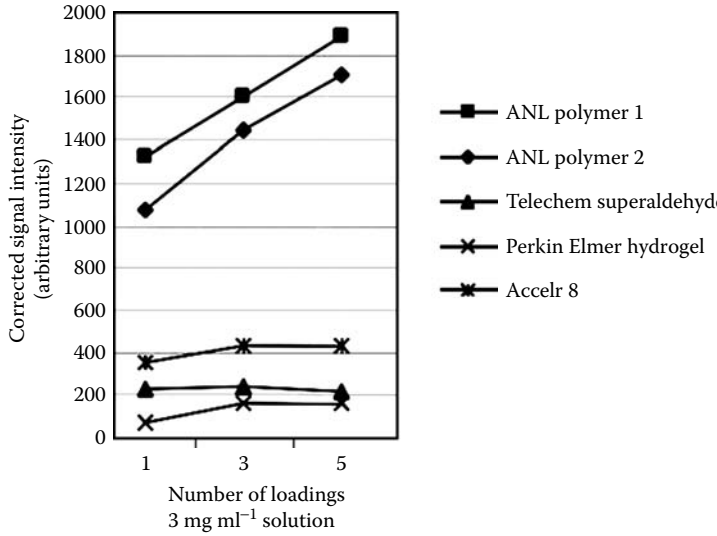


FIGURE 11.1 Repeated loading of biotinylated BSA solutions onto different biochip substrates. Each loading was 1 nanoliter of 3 mg ml^{-1} solution. After loading, biochips were reacted with 1 μg ml^{-1} streptavidin-Texas Red conjugate for 1.5 h and imaged on a Packard Biosciences Biochip Images scanner. The average fluorescence value for empty gel elements was subtracted from the average for elements loaded with biotinylated BSA. Results show the average of 5 replicate gel elements per loading and substrate.

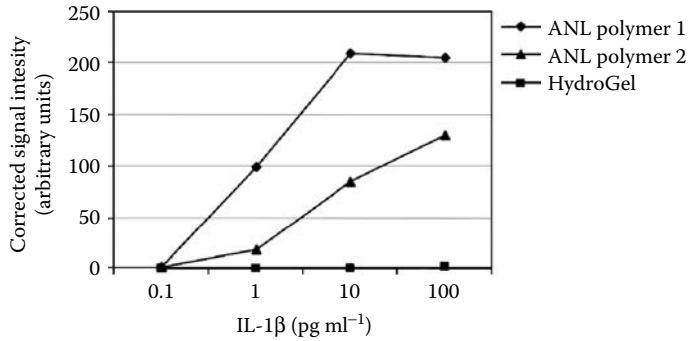


FIGURE 11.2 Three-dimensional array substrate performance in a standard sandwich immunoassay. Mouse Anti-Human IL-1β antibody or BSA was immobilized at 1 ng per spot or gel element (1 mg ml^{-1}) in 12 replicates. Incubation steps for the assay were (1) IL-1β at 0.1 to 100 pg ml^{-1}, 1.5 h; (2) biotinylated anti-IL-1β antibody at 1 μg ml^{-1}, 1.5 h; (3) horseradish peroxidase conjugated with streptavidin at 1 μg ml^{-1}, 1.5 h; (4) Tyramide Signal Amplification with biotin-tyramide, 30 min; (5) streptavidin-Texas red conjugate at 1 μg ml^{-1}, for 60 min. Fluorescent intensities were recorded on a stationary fluorescent microscope with a 1 sec acquisition time. The average signal intensity for BSA spots was subtracted from the average spot intensity for elements containing anti-IL-1β.

quantitative biochip responses may be possible over the 6-log dynamic range required for functional proteomics studies and at biologically relevant target concentrations.

LEAVING THE MALDI SURFACE BEHIND

In order to circumvent limitations of conventional LC or affinity separation techniques, provide a platform for global protein interaction analyses and specifically identify interacting partners on whole-proteome chips, we started using the new polymeric gel element microarrays as an affinity purification platform and interfacing them directly with a commercially available MALDI mass spectrometer as the detector (Figure 11.3). The concept builds off the premise of MALDI-MS from 2-D polyacrylamide gel pieces[25–28] and an earlier report of MALDI-MS detection of nucleic acid duplexes from gel element arrays.[56] To develop and demonstrate the 3-D MS protein chip technology, we initially used bovine trypsin and trypsin inhibitors to develop protein immobilization protocols, interaction assay conditions, and an analytical procedure for mass spectral detection of interacting proteins. As with prior (optical) protein array studies,[51] the MS-biochip system

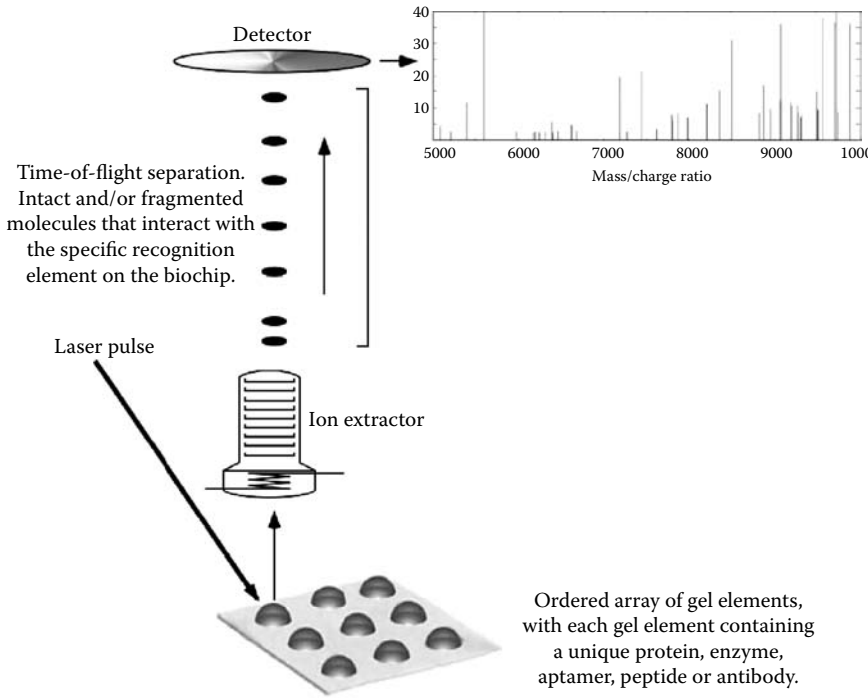


FIGURE 11.3 Conceptual drawing of the three-dimensional gel element array-MALDI MS detection system. A Bruker Biflex III mass spectrometer was used for all MS experiments described here.[31]

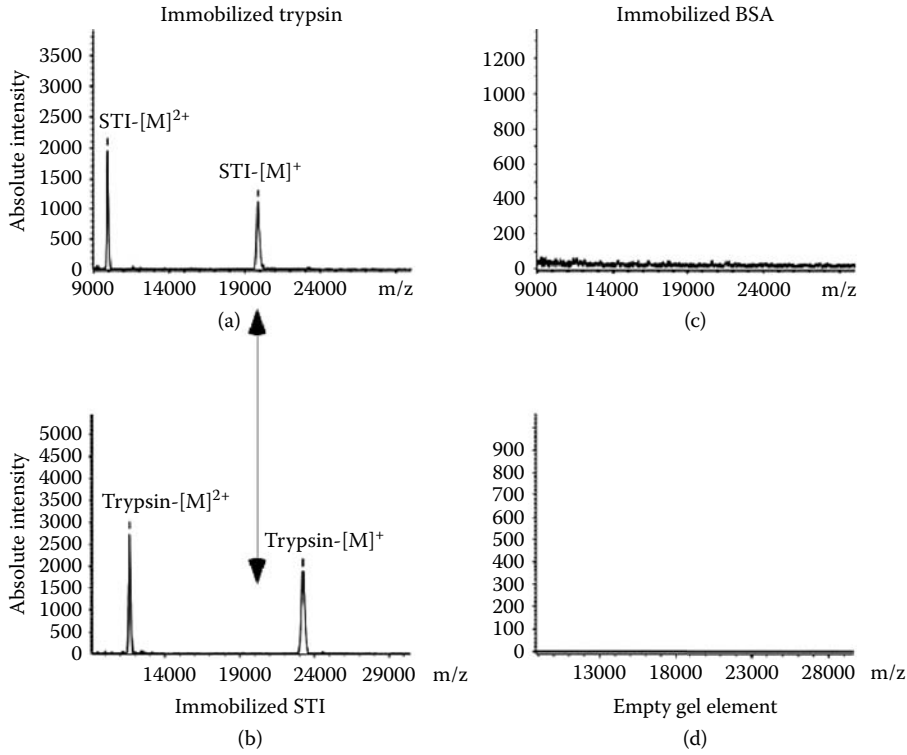


FIGURE 11.4 On-chip MALDI-MS analysis of trypsin interactions with trypsin inhibitor in three-dimensional gel element arrays. Bovine trypsin (A), soybean trypsin inhibitor (STI) (B), and BSA (C) were immobilized in the array elements using 10 mg ml^{-1} stock solutions. The mass spectra were obtained from the array elements hybridized to either 30 µl of 10 µg ml^{-1} STI (A, C) or 1 µg ml^{-1} trypsin (B). Panel (D) corresponds to the blank array elements that were incubated with the buffer during the immobilization procedure (i.e., negative control). The arrow between panels A and B indicates that immobilized STI is not ablated from the gel elements (B). Each mass spectrum is the representative signal from 200 laser shots per element from at least 5 elements in one experiment, and each experiment was repeated at least three times.

is able to detect specific interactions regardless of which protein is immobilized in the gel elements (Figure 11.4A and 11.4B), indicating that specific functional interactions can occur within the gel element array. Weakly interacting partners can also be separated and detected ($K_a = 10^6$ M^{-1}).[55] Control reactions between STI and immobilized BSA (Figure 11.4C) and application of MALDI matrix to a blank gel element array incubated with trypsin solution (Figure 11.4D) resulted in no detectable signal, suggesting that the gel element formulation is relatively immune to nonspecific protein binding. The absence of any signal corresponding to the immobilized protein (highlighted by the arrow between Figures 11.4A and 11.4B) demonstrates that gel-immobilized protein or capture probe is not released from the gel element during laser desorption and ionization. Competitive inhibition

was easily detected on-chip by adding 10 μg ml^{-1} bovine pancreatic trypsin inhibitor (BPTI) to a trypsin solution, and incubating the mixture with a gel element array containing immobilized sunflower trypsin inhibitor (SFTI); as expected, trypsin was not detected by MALDI-MS when BPTI was present in solution (I. Gavin, unpublished data).

Many protein interactions are based on a limited number of amino acid residues and contact points, some of which can be critical for the assembly of complexes that coordinate specific cellular functions.[57] Hence, understanding the location, nature and interaction of peptide recognition elements and motifs is important for drug discovery and as a means to verify computational predictions of protein structure-function.[58] To illustrate how the 3-D MALDI-MS gel element arrays can be used to address these questions, we synthesized an overlapping (tiled) set of trypsin peptides covering known interaction domains between trypsin and STI, and immobilized them at equimolar concentrations to discreet gel elements. The resulting arrays of trypsin peptides were hybridized to STI and the interacting peptides detected directly on-chip by MALDI mass spectrometry. As predicted and expected, STI interacted with peptides containing the strong binding domains A and B and the predicted hairpin structure required for STI interaction with trypsin (Figure 11.5). Other peptides in the tiled array or those corresponding to weak binding sites showed no detectable interaction with STI (I. Gavin, manuscript in preparation).

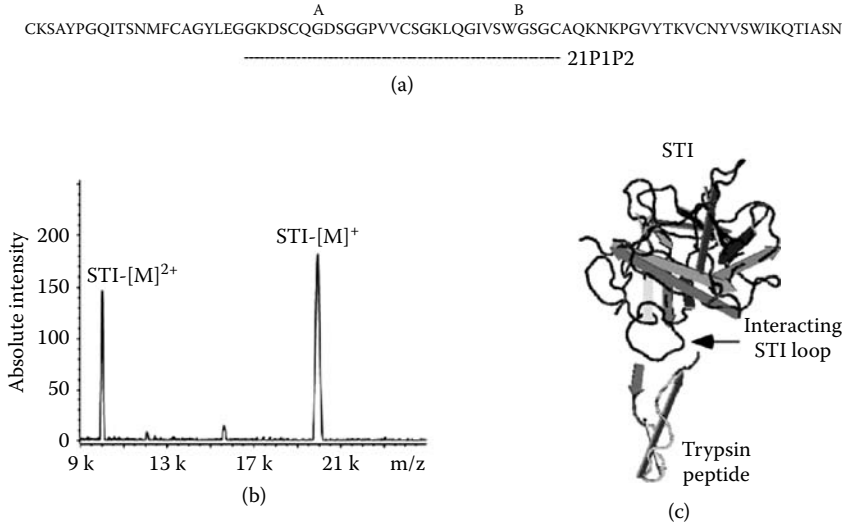


FIGURE 11.5 Mapping peptide interaction domains with 3-D MALDI-MS biochips. 15-mer trypsin peptides were immobilized with gel elements at 5 m*M* concentration in 50% acetonitrile. Arrays were incubated with 100 mg ml^{-1} STI and analyze by on-chip MALDI-TOF. (A) Trypsin primary sequence with two STI binding sites (A and B). Peptide 21P1P2 encompasses both binding sites. (B) On-chip mass spectrum for peptide 21P1P2. Weakly interacting peptides showed no STI binding. (C) Interacting peptides/structure map to known interaction sites in Trypsin-STI crystal structure.

To determine if immobilized trypsin also retains functional activity within the gel element, trypsin was immobilized in gel element arrays and incubated with 10 m*M* apomyoglobin overnight at 37°C. The spectrum of products of apomyoglobin proteolysis by immobilized trypsin was similar to the spectrum obtained from a test-tube control reaction, indicating that protein immobilization within the gel element does not destroy enzyme function and activity.[55] These data also indicate that on-chip, gel element tryptic digestion is possible and that sufficient quantity of peptide fragments remain for MS detection. This property of gel elements becomes very important as we endeavor (in the future) to specifically identify interacting partners within gel elements via tandem MS approaches.

MALDI MS-based detection techniques have also been used in combination with planar[34,59] and continuous Hydrogel[60] substrates. From the data shown in Figures 11.1 and 11.2, however, we believe that discrete gel elements provide higher probe immobilization capacity and, hence, improved detection limits and/or dynamic range than other substrates for on-chip MS analysis. To begin addressing this hypothesis, we determined[14] the lowest concentration of TNF-α cytokine in cell culture medium that can be detected with the MS biochip system on three three-dimensional substrates. Anti-TNF-α monoclonal capture antibodies were immobilized on the slides and hybridized to unpurified cell culture medium containing human recombinant TNF-α at various concentrations. A concentration-dependent peak corresponding to the secreted form of TNF-α was observed at concentrations higher than 1 ng ml^{-1} for gel element arrays but at > 10 ng ml^{-1} for Hydrogel arrays. Gel element biochips detected as little as 20 f*M* TNF-α in 10^7 excess of cell culture medium proteins (w/w) and generated the highest signal amongst the microarray substrates, even amidst nonspecific interactions between proteins present in the cell medium and the gel elements; in this case, there was no detectable interaction on planar chips (see Gavin et al.[55]). Comparable results were obtained using an STI model system and cell lysates, where STI was amended directly into cell lysate and applied directly to a gel element array containing immobilized trypsin (Figure 11.6). These (and other) data suggest that the planar surface is either denaturing protein, orienting proteins in a sterically unfavorable manner for interaction, providing limited probe binding capacity, or any combination thereof, as suggested by other protein chip studies. The improved signal of the gel element arrays relative to planar substrates or continuous Hydrogel suggests that the gel elements are providing a higher probe loading capacity (hence, improved capacity for interaction), or that the gel element formulation provides better binding conditions than Hydrogel. In either case, these experiments demonstrate the feasibility of using three-dimensional MS biochips for functional interaction assays in complex media, and the relative advantage of three-dimensional gel element arrays over commercial substrates for continued technology development and application.

MAKING THE MOST OF UNDEFINED PROTEIN CONTENT

Creating functional protein arrays typically requires a sequenced genome, well-characterized protein expression system and intensive expression and purification methods; regardless, the majority of proteins expressed *in vitro* lack post translational

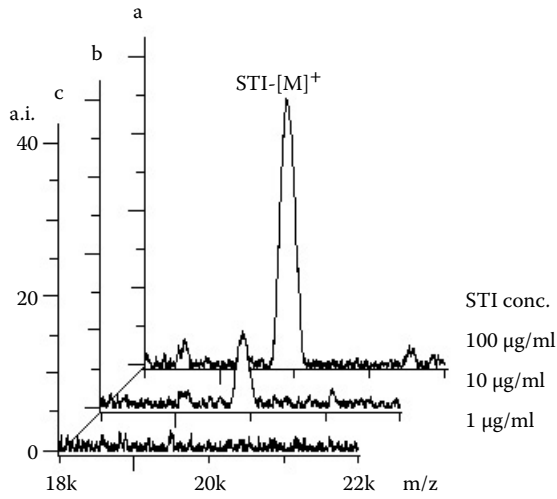


FIGURE 11.6 Functional protein interaction and detection in complex media. Human embryonic kidney cell line 293 was cultivated and lysed according to standard procedures. The crude cell lysate was amended with STI at 1 to 100 µg ml^{-1}, and incubated with gel element arrays containing immobilized trypsin at 10 ng per gel element that had been preblocked with Superblock for 1 hr. After incubation and a quick rinse in PBS, arrays were analyzed by MALDI-MS. STI was detectable when amended into the crude cell lysate at 10 µg ml^{-1}.

modifications that may be required for proper function or interaction.[61,62] Practical difficulties associated with protein expression and purification are therefore obvious impediments to protein array manufacture and use, resulting in a number of new methods for generating protein array content (as discussed elsewhere in this volume). In order to circumvent the protein content bottleneck, we have become particularly interested in recently described two-dimensional liquid phase separation technology (PF2D[63–69]) as a means of generating comprehensive functional protein arrays[70] for use with the MALDI-MS system described here. The allure of PF2D fractions for functional protein arrays and 3-D MALDI-MS biochips is that proteins are generated *in vivo* by the organism of interest; hence, a sequenced genome is not required in order to reproducibly generate protein content, and the resulting fractionated proteins retain all post translational modifications intact.

Preliminary methods for protein separation, protein array preparation, and protein array QA/QC analysis utilized *Yersinia pestis* KIMD27 as a model system. Bacterial cell pellets were lysed and 2 mg lysate separated at analytical scale (Eprogen; Darien, IL) according to pI and hydrophobicity, with the resulting fractions imaged and quantified as a two-dimensional map of protein content versus plate fraction (Figure 11.7A). In order to determine if the PF2D-array method could also separate and immobilize outer membrane proteins, we also performed a cell surface biotinylation procedure on intact Yersinia pestis cells

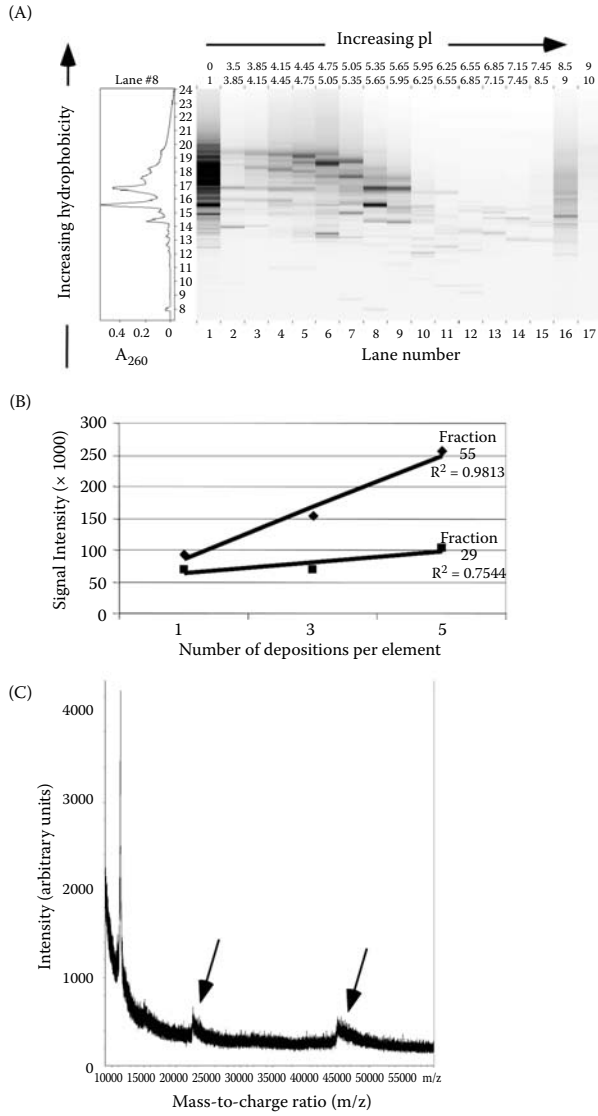


FIGURE 11.7 *Y. pestis* KIM-D27 was cultured under virulence-inducing conditions, washed, and subjected to cell surface biotinylation. After cell lysis, the total cell lysate was fractionated by ProteoSep 2-D liquid-phase fractionation (PF2D), resulting in an analytical-scale ProteoVue map (A). Fractions were automatically collected every 24 seconds, resulting in 864 total fractions within 9 × 96 deep-well plates. An A_{260} absorbance trace is shown for lane 8. Resulting fractions were applied to three-dimensional gel element arrays at 1, 3, or 5 depositions per gel element, and then developed with streptavidin-Texas Red to identify those fractions containing putative (biotinylated) membrane proteins (B). Fractions deposited within gel element arrays are themselves amenable to mass spectral analysis (C); Fraction 29 contained at least one prominent species that is putatively identified as an outer membrane protein due to its reactivity in (B), where arrows indicate the M^+ and M^{2+} ions.

prior to PF2D fractionation. Based on the PF2D ProteoVue profile and NanoDrop UV/vis quantitation, 88 protein fractions (from 864) containing the highest protein concentrations were deposited and immobilized onto 3-D gel element arrays at 1, 3, or 5 depositions (1 nl per deposition) per gel element. PF2D protein arrays were then reacted with streptavidin-Texas Red and analyzed with an optical detector. Two of the 88 immobilized PF2D protein fractions were putatively identified as (biotinylated) outer membrane proteins based on a linear correlation between microarray signal intensity and number of protein depositions onto the gel elements (Figure 11.7B). PF2D protein fractions were also analyzed by MALDI-MS directly on-chip to estimate the total number of proteins present per fraction and to understand the minimum MS-based detection limit for uncharacterized proteins (Figure 11.7C). These results suggest that both the PF2D fractionation method and 3-D gel element microarrays can isolate and immobilize intact proteins, including membrane proteins, for subsequent (functional) characterization and interaction assays.

While our preliminary studies show that PF2D fractions and intact proteins can be manipulated, immobilized and detected on-chip, detecting functional interactions between immobilized PF2D fractions and cell lysates requires successful immobilization of sufficient protein quantities to achieve successful detection and/or identification. The optical detection limit of 3-D gel element arrays is ~1 pg per gel element, whereas MS detection and identification requires a higher (and as of yet indeterminate) protein concentration per gel element. Preliminary functional interaction assays with analytical[17] scale *Yersinia* PF2D fractions indicate that there is insufficient fractionated protein in an analytical-scale PF2D separation (starting with 2 mg cell lysate) to achieve sufficient protein density within individual gel elements to detect functional interactions with the MALDI-MS system described here. However, there are several other methods by which the effective protein concentration in each PF2D fraction and gel element can be increased. First, pre-subcellular fractionation can be employed prior to PF2D fractionation to simplify the protein mixture applied to the PF2D system. Second, preparative-scale PF2D methods can be developed to increase total protein biomass. Third, pre-subcellular fractionation and preparative scale fractionation should simplify the protein mixture/complexity applied to each gel element, thereby increasing the effective concentration of individual proteins within each gel element. These modifications are the subject of ongoing research to extend the capabilities of the 3-D MALDI-MS arrays to the functional analysis of previously uncharacterized proteomes. In any case, we are cautiously optimistic that the three-dimensional gel element MALDI-MS chips described here will also find practical utility for analyzing and decoding functional protein networks starting from undefined or uncharacterized genome and proteome content.

SUMMARY

Mass spectrometry clearly offers the opportunity to detect proteins and identify their molecular masses without *a priori* knowledge of protein expression, or generating affinity reagents for detecting each target of interest. SELDI-mass spectrometry, in

particular,[18] is a well-known commercially available platform for chromatographic separations and protein detection on two-dimensional surfaces, but interaction assays on planar substrates are subject to the limitations of two-dimensional surfaces, including protein denaturation and limited probe immobilization capacity. The promise and potential of leaving the surface behind is that three-dimensional biochip substrates will overcome the steric and probe immobilization capacity constraints of planar biochips, while preserving biomolecular integrity, function and activity. As shown here and in Gavin et al.,[55] the substrate does matter for functional proteomics, with some clear performance advantages to discreet three-dimensional gel elements in both optical and MS-based detection modes.

The promise of MALDI-MS protein array technologies is to circumvent drawbacks of conventional antibody arrays, affinity reagents, and optical detectors. That is, a multiplexed antibody or ELISA array assumes that all interactions between target and cognate antibody are absolutely specific, an assumption that may or may not be valid. Generating an affinity tag for each and every protein of interest, for each and every proteome of interest, is a daunting task with significant technical and financial burdens. As illustrated in Figures 11.4 to 11.6, identifying molecular masses of spatially captured proteins from complex mixtures (e.g., cell lysates) will ultimately eliminate the need for specific detection antibodies and will allow functional protein interactions and analyses to be carried out with virtually an unlimited number of probes, irrespective of *a priori* knowledge or occurrence of interacting partners. Although presently less sensitive than conventional ELISA, the biochip MALDI-MS technique is able to "read through" complex backgrounds and any nonspecific binding of molecules, affording the opportunity to unambiguously detect (and eventually identify) interacting partners under natural (cellular, tissue or environmental) conditions. From this perspective, relatively nonspecific antibodies, capture probes and/or interacting partners can (in principle) be used as capture molecules in gel element arrays for affinity-based separations prior to MS detection, a capability that will facilitate the development and analysis of whole proteome chips, including those derived from previously un-sequenced genomes or uncharacterized isolates. Importantly, the preliminary data presented here (Figure 11.7) also indicate that the 3-D protein array manufacturing process may be effective on native membrane proteins, opening new opportunities for deciphering molecular networks and/or screening drug precursors and libraries. Hence, we are optimistic that further development of 3-D MALDI-MS gel element arrays will add value as a high-throughput, functional proteomics platform for functional proteomics, systems biology and drug discovery applications.

ACKNOWLEDGMENTS

This research was supported by the U.S. Department of Energy (DOE), Office of Science, under Argonne National Laboratory LDRD projects A03126 and A03316. Argonne National Laboratory is operated by the Uchicago LLC for the U.S. DOE under Contract No. DE-AC02-06CH 11357.

REFERENCES

1. Petricoin, E.F. and Liotta, L.A., Proteomic analysis at the bedside: Early detection of cancer, *Trends Biotechnol*. 20, S30, 2002.
2. Hood, L. et al., Systems biology and new technologies enable predictive and preventative medicine, *Science*, 306, 640, 2004.
3. Frazier, M.E. et al., Realizing the potential of the genome revolution: The Genomes to Life Program, *Science*, 300, 290, 2003.
4. MacBeath, G., Protein microarrays and proteomics, *Nature Genetics Suppl*., 32, 526, 2002.
5. Aebersold, R. and Mann, M., Mass spectrometry-based proteomics, *Nature*, 422, 198, 2003.
6. Harry, J.L. et al., Proteomics: Capacity versus utility, *Electrophoresis*, 21, 1071, 2000.
7. Quadroni, M. and James, P., Proteomics and automation, *Electrophoresis*, 20, 664, 1999.
8. Zhu, H. and Snyder, M., Protein chip technology, *Curr. Op. Chem. Biol*., 7, 55, 2003.
9. Tyers, M. and Mann, M., From genomics to proteomics, *Nature*, 422, 193, 2003.
10. Smith, R.D. et al., Rapid quantitative measurements of proteomes by Fourier transform ion cyclotron resonance mass spectrometry, *Electrophoresis*, 22, 1652, 2001.
11. Buchanan, M., Genomes to Life: Technology assessment for mass spectrometry, in DOE-OBER GTL Mass Spectrometry Workshop, Mariott Wardham Park Hotel, Washington, DC, 2001, p. 14.
12. Lipton, M.S. et al., Global analysis of the *Deinococcus radiodurans* proteome using accurate mass tags, *Proc. Natl. Acad. Sci. USA*, 99, 11049, 2002.
13. Schaeferling, M. et al., Application of self-assembly techniques in the design of biocompatible protein microarray surfaces, *Electrophoresis*, 23, 3097, 2002.
14. Zhu, H. et al., Global analysis of protein activities using proteome chips, *Science*, 293, 2101, 2001.
15. Haab, B.B., Dunham, M. J., and Brown, P., Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions, *Genome Biol*., 2, research0004.1, 2001.
16. Peter, J.-C., Briand, J.-P., and Hoebeke, J., How biotinylation can interfere with recognition: A surface plasmon resonance study of peptide-antibody interactions, *J. Immunol. Meth*., 274, 149, 2003.
17. Tseng, G.C. et al., Issues in cDNA microarray analysis: Quality filtering, channel normalization, models of variations and assessment of gene effects, *Nucl. Acids Res*., 29, 2549, 2001.
18. James, P., Chips for proteomics: A new tool or just hype?, *BioTechniques Suppl*., 4, 2002.
19. Kachman, M.T. et al., A 2-D liquid separations/mass mapping method for interlysate comparison of ovarian cancers, *Anal. Chem*., 74, 1779, 2002.
20. Zhang, W. and Chait, B.T., Profound: An expert system for protein identification using mass spectrometric peptide mapping information, *Anal. Chem*., 72, 2482, 2000.
21. Wang, T. et al., Reconstructed protein arrays from 3D HPLC/tandem mass spectrometry and 2D gels: Complementary approaches to *Porphyromonas gingivalis* protein expression, *Analyst*, 127, 1450, 2002.
22. Weinberger, S.R., Viner, R., and Ho, P., Tagless extraction-retentate chromatography: A new global protein digestion strategy for monitoring differential protein expression, *Electrophoresis*, 23, 3182, 2002.
23. Bergkvist, J. et al., Improved chip design for integrated solid-phase microextraction in on-line proteomic sample preparation, *Proteomics*, 2, 422, 2002.

24. Regnier, F.E. et al., Chromatography and electrophoresis on chips: Critical elements of future integrated, microfluidic analytical systems for life science, *Trends Biotechnol.*, 17, 101, 1999.
25. Ogorzalek-Loo, R. et al., Mass spectrometry of proteins directly from polyacrylamide gels, *Anal. Chem.*, 68, 1910, 1996.
26. Gross, J. and Strupat, K., Matrix assisted laser desorption/ionization-mass spectrometry applied to biological macromolecules, *Trends Anal. Chem*, 17, 470, 1998.
27. Katayama, H., Nagasu, T., and Oda, Y., Improvement of in-gel digestion protocol for peptide mass fingerprinting by matrix assisted laser desorption/ionization time-of-flight mass spectrometry, *Rapid Comm. Mass Spectrom.*, 15, 1416, 2001.
28. Shevchenko, A. et al., Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole-time-of-flight mass spectrometry and BLAST homology searching, *Anal. Chem.*, 73, 1917, 2001.
29. Bienvenut, W.V. et al., Toward a clinical molecular scanner for proteome research: Parallel protein chemical processing before and during western blot, *Anal. Chem.*, 71, 4800, 1999.
30. Merchant, M. and Weinberger, S.R., Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry, *Electrophoresis*, 21, 1164, 2000.
31. Alphonso, C. and Fenseleau, C., The use of bioactive glass slides for matrix assisted laser desorption/ionization analysis: Application to microorganisms, *Anal. Chem.*, 75, 694, 2003.
32. Bundy, J.L. and Fenselau, C., Lectin and carbohydrate affinity capture surfaces for mass spectrometric analysis of microorganisms, *Anal. Chem.*, 73, 751, 2001.
33. Fung, E.T. et al., Protein chips for differential profiling, *Curr. Op. Biotechnol.*, 12, 65, 2001.
34. Nedelkov, D. and Nelson, R. W., Analysis of native proteins from biological fluids by biomolecular interaction analysis/mass spectrometry (BIA/MS): Exploring the limit of detection, identification of nonspecific binding and detection of multiprotein complexes, *Biosensors & Bioelectronics*, 16, 1071, 2001.
35. Lundquist, M. et al., Descrimination of *Francisella tularensis* subspecies using surface enhanced laser desorption ionization mass spectrometry and multivariate data analysis, *FEMS Microbiol. Lett.*, 243, 303, 2005.
36. Nedelkov, D., Tubbs, K.A., and Nelson, R.W., Design of buffer exchange surfaces and sensor chips for biosensor chip mass spectrometry, *Proteomics*, 2, 441, 2002.
37. Kononen, J. et al., Tissue microarrays for high-throughput molecular profiling of tumor specimens, *Nature Med.*, 4, 844, 2998.
38. Endler, E.E. et al., Propagation of viruses on micropatterned host cells, *Biotechnol. Bioeng.*, 81, 719, 2003.
39. Emili, A.Q. and Cagney, G., Large-scale functional analysis using peptide or protein arrays, *Nat. Biotechnol.*, 18, 393, 2000.
40. Mendoza, L.G. et al., High-throughput microarray-based enzyme-linked immunosorbent assay (ELISA), *BioTechniques*, 27, 778, 1999.
41. Barry, R. and Soloviev, M., Quantitative protein profiling using antibody arrays, *Proteomics*, 4, 3717, 2004.
42. Tong, M. et al., A multiplexed and miniaturized serological tuberculosis assay identifies antigens that discriminate maximally between TB and non-TB sera, *J. Immunol. Meth.*, 301, 154, 2005.
43. Zhu, H. et al., Analysis of yeast protein kinases using protein chips, *Nature Genetics*, 26, 283, 2000.

44. Fukui, S. et al., Oligosaccharide microarrays for high-throughput detection and specificity assignments of carbohydrate-protein interactions, *Nature Biotechnol*., 20, 1011, 2002.
45. MacBeath, G., Koehler, A.N., and Schreiber, S.L., Printing small molecules as microarrays and detecting protein-ligand interactions en masse, *J. Am. Chem. Soc*., 121, 7967, 1999.
46. Mirzabekov, A. and Kolchinsky, A., Emerging array-based technologies in proteomics, *Curr. Opin. Chem. Biol*., 6, 70, 2002.
47. Angenendt, P. et al., Toward optimized antibody microarrays: A comparison of current microarray support materials, *Anal. Biochem*., 309, 253, 2002. 26
48. Wang, C.C. et al., Array-based multiplexed screening and quantitation of human cytokines and chemokines, *J. Proteome Res*., 1, 337, 2002.
49. Guschin, D. et al., Manual manufacturing of oligonucleotide, DNA and protein microchips, *Anal. Biochem*., 250, 203, 1997.
50. Vasiliskov, A.V. et al., Fabrication of microarray of gel-immobilized compounds on a chip by copolymerization, *BioTechniques*, 27, 592, 1999.
51. Arenkov, P. et al., Protein microchips: Use for immunoassay and enzymatic reactions, *Anal. Biochem*., 278, 123, 2000.
52. Dubiley, S. et al., Fractionation, phosphorylation and ligation on oligonucleotide microchips to enhance sequencing by hybridization, *Nucl. Acids Res*., 25, 2259, 1997.
53. Mikhailovich, V. et al., Identification of rifampin-resistant *Mycobacterium tuberculosis* strains by hybridization, PCR, and ligase detection reaction on oligonucleotide microchips, *J. Clin. Microbiol*., 39, 2531, 2001.
54. Pemov, A. et al., DNA analysis with multiplex microarray-enhanced PCR, *Nucl. Acids Res*., 33, e11, 2005.
55. Gavin, I.M. et al., Analysis of protein interaction and function with a 3-dimensional MALDI-MS protein array, *BioTechniques*, 39, 99, 2005.
56. Stomakhin, A.A. et al., DNA sequence analysis by hybridization with oligonucleotide microchips: MALDI mass spectrometry identification of 5-mers contiguously stacked to microchip oligonucleotides, *Nucl. Acids Res*., 28, 1193, 2000.
57. Pawson, T. and Nash, P., Assembly of cell regulatory systems through protein interaction domains, *Science*, 300, 445, 2003.
58. Tong, A.H.Y. et al., A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules, *Science*, 295, 321, 2002.
59. Nedelkov, D. and Nelson, R.W., Surface plasmon resonance mass spectrometry: Recent progress and outlooks, *Trends Biotechnol*., 21, 301, 2003.
60. Scrivener, E. et al., Peptidomics: A new approach to affinity protein microarrays, *Proteomics*, 3, 122, 2003.
61. Templin, M.F. et al., Protein microarray technology, *Trends Biotechnol*., 20, 160, 2002.
62. Predki, P.F., Functional protein microarrays: Ripe for discovery, *Curr. Opin. Chem. Biol*., 8, 8, 2004.
63. Lubman, D.M. et al., Two-dimensional liquid separations — mass mapping of proteins from human cancer cell lysates, *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci*., 782, 183, 2002.
64. Yan, T. et al., Molecular diversity and characterization of nitrite reductase gene fragments (nirK and nirS) from nitrate- and uranium-contaminated groundwater, *Environ. Microbiol*., 5, 13, 2003.
65. Yan, F. et al., A comparison of drug-treated and untreated HCT-116 human colon adenocarcinoma cells using a 2-D liquid separation mapping method based upon chromatofocusing PI fractionation, *Anal. Chem*., 75, 2299, 2003.

66. Zheng, S. et al., Two-dimensional liquid chromatography protein expression mapping for differential proteomic analysis of normal and O157:H7 Escherichia coli, *BioTechniques*, 35, 1202, 2003.
67. O'Neil, K.A. et al., Profiling the progression of cancer: Separation of microsomal proteins in MCF10 breast epithelial cell lines using nonporous chromatophoresis, *Proteomics*, 3, 1256, 2003.
68. Zhu, K. et al., Identification of low molecular weight proteins isolated by 2-D liquid separations, *J. Mass Spectrom*., 39, 770, 2004.
69. Van Le, T.S. et al., Functional characterization of the bladder cancer marker, BLCA-4, *Clin. Cancer Res*., 10, 1384, 2004.
70. Yan, F. et al., Protein microarrays using liquid phase fractionation of cell lysates, *Proteomics*, 3, 1228, 2003. 2930.

12 High-Resolution Label-Free Detection Applied to Protein Microarray Research

Lance G. Laing and Brian Cunningham

CONTENTS

INTRODUCTION

Microarrays have found great utility to date. However, when considering recent reviews and speaking with leaders in the field applying microarrays daily to scientific problems, one quickly learns that current microarray methods are limited in terms of analytical methods. Key issues reside in the nondestructive determination of the amount and activity of the material on the microarray slide and measurement problems created by the use of labels. This chapter will provide an introduction to one type of label-free biosensor technology (the BIND system), show data from demonstration applications with this system, and describe the uses and limitations of BIND as currently configured for protein microarrays. In the final section, a description is provided of the future of the technology and practical improvements planned for near term commercial availability.

BIND INTRODUCTION

Label-Free Technology

Photonic Crystal Biosensors

A new class of label-free biosensors based on the unique properties of optical device structures known as "photonic crystals" have been recently developed.[1,2] Photonic crystal structures have their historical roots in a phenomenon called "Wood's Anomaly." Wood's Anomalies are effects observed in the spectrum of light reflected by optical diffraction gratings.[3]

A photonic crystal is composed of a periodic arrangement (or gratings) of dielectric material in two or three dimensions.[4,5] If the periodicity and symmetry of the crystal and the dielectric constants of the materials used are chosen appropriately, the photonic crystal will selectively couple energy at particular wavelengths, while excluding others.[6] Device structures based on linear gratings and two-dimensional gratings (i.e., arrays of holes, posts, or veins arranged in checker-board or hexagonal close-packed grids along the sensor surface) have been demonstrated. Photonic crystal structure geometry can be designed to concentrate light into extremely small volumes and to obtain very high local electromagnetic field intensities. For example, sub-wavelength periodic structures have been developed to reflect only a very narrow band of wavelengths when illuminated with white light.[7]

For use as a biosensor, a photonic crystal may be optimized to provide an extremely narrow resonant mode whose wavelength is particularly sensitive to modulations induced by the deposition of biochemical material on its surface.[1] By attaching receptive molecules, biomolecules, or cells to the portion of the photonic crystal where the locally confined electromagnetic field intensity is greatest, the resonant coupling of light into the crystal is modified, so the reflected/transmitted output is tuned with changes directly proportional to the attached mass. The highly confined electromagnetic field within a photonic crystal structure provides high sensitivity and a high degree of spatial resolution consistent with their use in imaging applications, much like fluorescent imaging scanners.

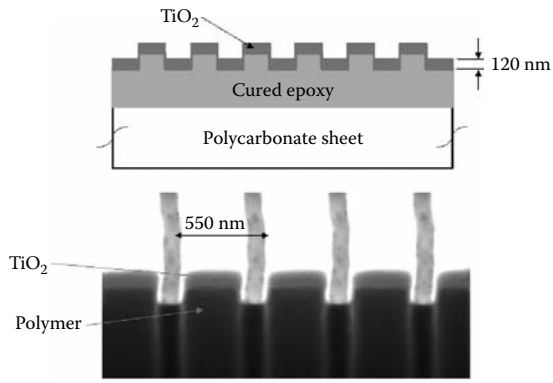
Biosensor Function

One of the first implementations of a photonic crystal biosensor has been recently demonstrated using one- and two-dimensional periodic structured surfaces produced on glass substrates and on continuous sheets of plastic film. For purposes of imaging microarrayed materials, operation of the sensor begins with illuminating the surface with white light, and collecting the reflected light with a noncontact imaging system, thereby independently measuring different locations on the sensor (please refer to Figure 12.1). The crystal reflected "peak wavelength value" (PWV) can be determined with 0.5-picometer resolution by illuminating with white light at normal incidence, gathering reflected spectra with a low-cost spectrometer, and applying advanced peak finding analysis. The magnitude of the PWV shift is quantitatively proportional to the amount of mass attached to the sensor. Previously published work shows a resulting optimized mass detection sensitivity of <1 pg/mm^2 (obtained *without* three-dimensional hydrogel surface chemistry), a result which has not been demonstrated by any other commercially available biosensor.[8,9] The sensor operates by measuring changes in the PWV of reflected light as biochemical binding events take place on the surface. For example, when a protein is immobilized on the sensor surface, an increase in the reflected wavelength is measured when a complementary binding protein is exposed to the sensor. Using low-cost components, the readout instrument is able to resolve protein mass changes on the surface with resolution less than 1 pg/mm^2. While this level of resolution is sufficient for measuring small-molecule interactions with immobilized proteins, the dynamic range of the sensor is large enough to also measure larger biochemical entities including live cells, cell membranes, bacteria, and viruses. A sensor measurement requires ~20 msec, so large numbers of interactions can be measured in parallel, and kinetic information can be gathered. The reflected wavelength of the sensor can be measured either in "single point mode" (such as for measuring a single point within a microplate), or an imaging system can be used to generate an image of a sensor surface with <9 micron resolution. The "imaging mode" can be used for many applications to increase the overall resolution and throughput of the system such as label-free microarrays, imaging plate reading, self-referencing microplates,[10,11] and multiplexed spots/well.[12]

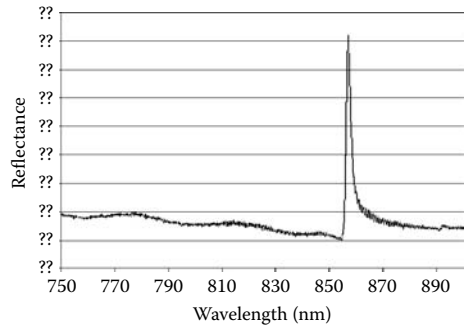
Biosensor Production

The BIND biosensor design enables a simple manufacturing process to produce sensor sheets in continuous rolls of plastic film that are hundreds of meters in length.[13] A fundamental advantage of this photonic crystal biosensor is the capability for inexpensive mass-manufacturing from plastic materials in continuous processes at a 1 to 2 ft/min rate. The mass manufacturing of a biosensor structure that is measurable in a noncontact mode over large areas enables the sensor to be incorporated into single-use disposable consumable items such as 96-, 384-, and 1536-well standard microplates as well as standard 1" × 3" microarray slides, thereby making the sensor compatible with standard fluid handling infrastructure employed in most laboratories.

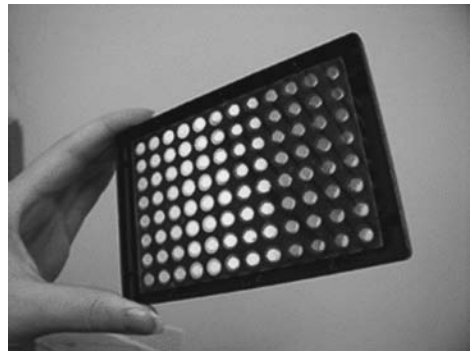
A sensor structure, shown in Figure 12.1b, consists of a low refractive index plastic material with a periodic surface structure, or grating, that is overcoated with



(a)



(b)



(c)

FIGURE 12.1 The BIND sensor system. (a) Photonic crystal biosensor structure utilizing a one-dimensional periodic surface structure of low dielectric constant polymer and high dielectric constant coating of TiO_2. Due to its simple structure, the sensor can be fabricated on continuous sheets of plastic film. When illuminated with white light, a narrowband reflectance spectrum is obtained (b) whose peak wavelength is tuned by the adsorption of biochemical material on the sensor surface. In (c), a 96-well microplate incorporating the photonic crystal biosensor into the entire bottom surface of each of the wells.

a thin layer of higher refractive dielectric material. The surface structure is replicated within a layer of cured epoxy from a silicon-wafer "master" mold (i.e., a negative of the desired replicated structure) using a continuous-film process on a polyester substrate. The manufacturing process[13] results in a >1000 meter long continuous plastic sheet of photonic crystal biosensors with TiO_2 dielectric as the active surface material. Appropriately sized sections are cut from the long sensor sheet, and attached with epoxy to the bottoms of bottomless microplates or microarray slides. Using this approach, photonic crystal sensors are mass-produced on a square-yardage basis at very low cost, with an extremely uniform process, yielding precise, high quality materials for analytical bioassays.

Imaging Detection Instrument

With the sensor structure illuminated at normal incidence by a collimated beam, only zeroth-order resonant coupling occurs. A photonic "band gap" for the photonic crystal structure is designed in the direction of periodicity (lateral to the surface) that cuts off the propagation of modes parallel to the surface. The zeroth-order coupling allows the sensor resonance to be detected as a mirror image of the surface, while the lack of lateral propagation ensures no optical cross talk between adjacent sensor regions. This capability enables high spatial resolution imaging detection of biomolecular binding density on the photonic crystal surface.

A single point illumination/single point spectrometer detection method described previously has been extended to incorporate an imaging spectrometer that is capable of generating high-resolution spatial maps of the PWV on the photonic crystal surface. This capability is possible due to the high degree of lateral optical confinement for photons resonantly coupled into the structure. This instrument allows the observation of patterns of biomolecule receptor attachment and hybridization interactions with high density. Because white light illumination is used, and because there is no optical contact required (such as a coupling prism) to the sensor, the imaging method can be performed on large sensor areas, such as entire microplates and microarray slides. As the same biosensor structure and peak-detecting method are used for single-point-based and imaging-based detection, the sensitivity (in terms of amount of PWV shift observed and resolution of PWV shift detection) of the approach is not compromised. The microtiter plate surface for protein microarray applications allows individually addressable portions of an array for rapidly testing different buffer or other binding interaction conditions.

A schematic diagram of the biosensor PWV imaging instrument is shown in Figure 12.2b. To generate a two-dimensional image of the sensor, a motorized stage translates the sensor in the direction that is perpendicular to the image line. The spatial separation of the image lines is determined by the step-size of the stage between each image-line acquisition. By this technique, a series of lines are assembled into an image. A large area can be scanned in a tiled fashion by translating the sensor in steps along the image-line direction.

Typically, a biosensor experiment involves measuring *shifts* in PWV, so the sensor surface is scanned twice, once before and once after biomolecular binding, and the images are aligned and subtracted to determine the difference in PWV as detected by the sensor. From the two scans, a quantitative determination of the mass attached at

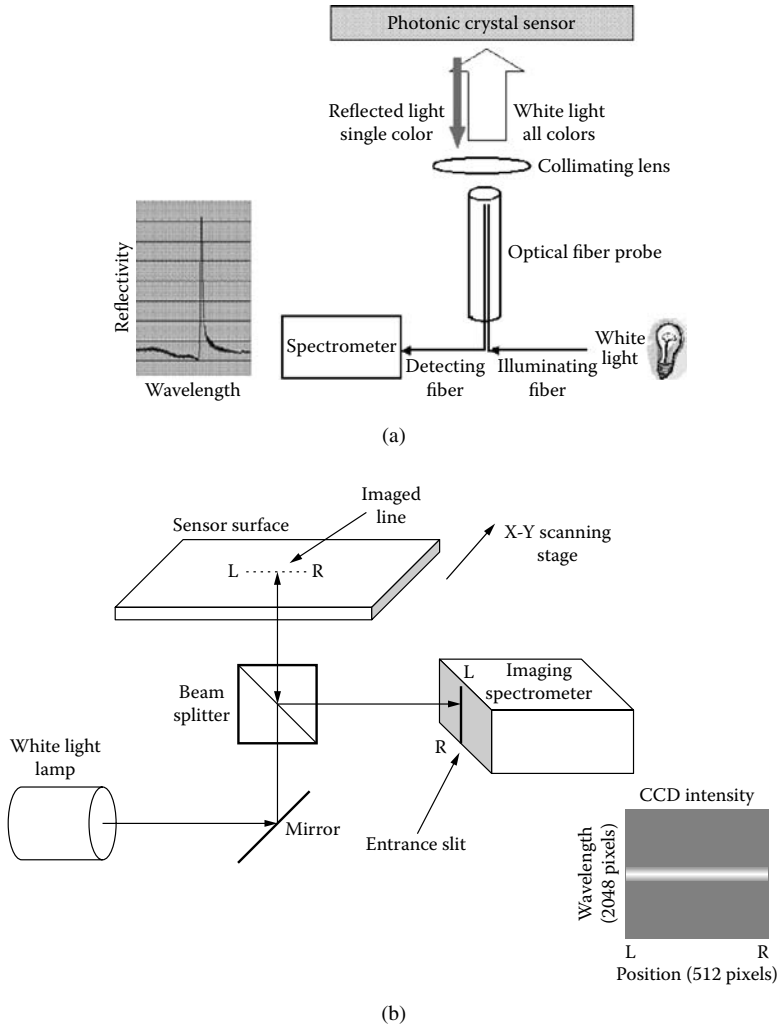
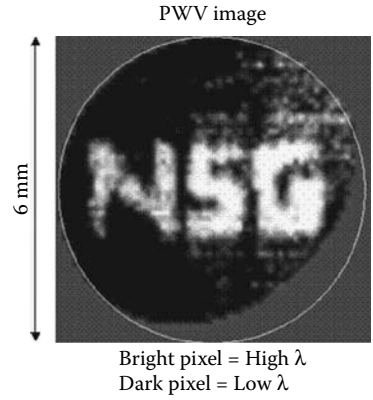
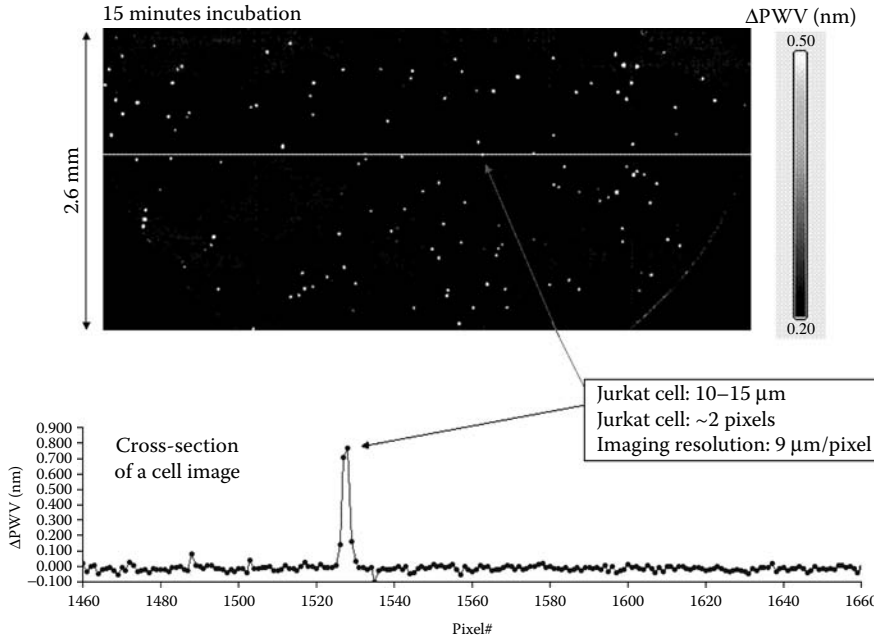


FIGURE 12.2 The BIND Reader and Scanner instruments. Excitation/detection instrumentation methods for photonic crystal biosensors. In (a), a light bulb illuminates the crystal surface at normal incidence through a fiber probe, and the reflected spectrum is gathered by a second fiber, connected to a spectrometer. High-resolution images of biochemical binding on the photonic crystal surface can be obtained using the instrument shown in (b), where an imaging spectrometer gathers hundreds of reflected spectra simultaneously from one line across the sensor surface.

each pixel may be determined from the PWV shift. This scanning method does not require the PWV of the imaged surface to be completely uniform, either across the surface or within a set of probe locations, or tuning of the sensor angle to a resonance condition as with SPR-imaging.[14] Figure 12.3 shows label-free biosensor images of PWV *shift* taken by pixel-by-pixel subtraction of a baseline PWV image from a PWV image captured after immobilization of protein (Figure 12.3A) or cells (Figure 12.3B).



(a)



(b)

FIGURE 12.3 PWV imaging. PWV shift images (bright regions represent regions of greater shift) of a photonic crystal sensor gathered using the instrument shown in Figure 12.2B. In (a), a 6-mm-diameter region of a biosensor is imaged at ~20 μm pixel resolution after writing the letters "NSG" with a Perkin-Elmer Piezoarray microarray spotting tool. In (b), the instrument is used to image a ~2.5 × 7.0 mm region of the biosensor surface at 9 μm pixel resolution to record the localized PWV shift caused by the attachment of individual cells. The cells themselves are typically 10 to 15 μm in diameter, so often they overlap two adjacent larger pixels, as shown in the cross section in the bottom right panel.

Application Issues

Advantages of Label-Free Methods

Many articles have described the problems and issues associated with current microarray experiments.[15–17] Technical studies are heavily focused on reducing experimental variation. A large proportion of the articles are devoted to issues dealing with the use of labels and validation of the resulting data.[18–21] The vast majority of assays currently used in pharmaceutical screening utilize some type of label to enable quantification of protein, DNA, small molecules, cells, or the interactions of these entities. The development of microarrays has relied upon labeling methods that have been available to date. Typical labeling methods include the use of fluorophores, radioligands, and secondary reporters. In contrast with the large variety of labeling methods, there are relatively few methods that allow detection of molecular and cellular interactions without labels. Label-free detection removes experimental uncertainty induced by the effect of the label on molecular conformation, blocking of active binding epitopes, steric hindrance, difficulties with getting the label into the labeling site, or the inability to find an appropriate label that functions equivalently for all molecules in an experiment.[1] Also, the labeled approach generally has the significant limitation of only reporting on the progress of an experiment when the labeled reagent is added to the reaction, and not on any other materials used. This can be a problem when trying to study the assembly of multi-subunit complexes as are often required in more interesting biological systems. This limitation also seriously disrupts quantification methods (concentration, activity, affinity, etc.) and data comparisons from experiment to experiment. Other problems arise when attempting to specifically identify molecules within complex pools of biological samples such as required for proteomics research. Label-free detection methods greatly simplify the time and effort required for assay development and provide quantitative analysis, while removing experimental artifacts from fluorescence labeling, quenching, shelf life, normalization, and background fluorescence.[22] Some chemiluminescent techniques introduce nonlinear signal enhancement, further complicating quantification challenges. Whether the microarray experiment is a survey array or scan array, both types of assays must satisfy scientific rigor, curiosity, and perhaps most importantly also be reproducible and interpretable beyond the single test itself. This is especially true if such data is to be used for drug approval or for the fulfillment of other federal diagnostic guidelines, as indicated by current trends in drug discovery.[23]

Quantification of Immobilized Material

In order for microarray results to become more analytical, quantification of reagent concentrations, binding, and correlation with experimental signal must be performed. The importance of this quantification can be seen in the simulated binding curves of Figure 12.4. The two curves represent two-body, 1:1 stoichiometry, noncovalent binding interactions of affinities differing by ten-fold. A horizontal dotted line drawn between the two curves at experimental signal equal to 0.5 represents a typical result (Ex. Fluorescence signal from immuno-based binding) where some reagent concentrations are not known or are largely uncertain. The experimental signal for the

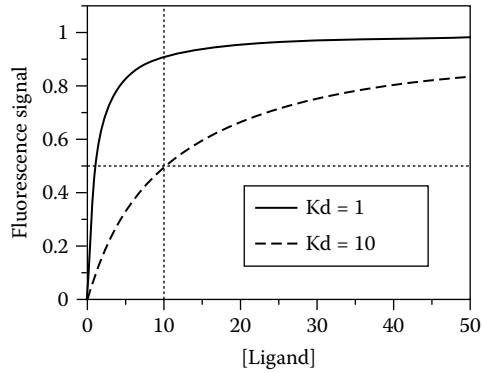


FIGURE 12.4 Simulated Binding Curves for two-body noncovalent interaction. Simulated binding curves (solid line and dashed line) for the noncovalent interaction of a receptor and ligand with single site binding on each demonstrate the importance of quantification of reagents. The curve with the solid line represents a tighter binding system (10×) than the curve with the dashed line. Dotted lines are selected cases for experimental conditions demonstrating the difficulty with work employing labels (horizontal line) and more favorable conditions (vertical line) when all concentrations are known beforehand.

weaker interaction is equal to the signal for the stronger interaction because the weaker ligand is at a ~5× higher concentration; no clear discernment of concentration or affinity can be made from this type of experiment. This scenario can be easily envisioned for many protein microarray experiments. A vertical dotted line through the two curves at [ligand] = 10 demonstrates a more optimal result and can only be obtained when the concentration of all of the reagents is well determined. In this case, the experimental signals are different and discernment of affinity can be made with systems of appropriate limits of detection.

Activity of Immobilized Material

In addition to being able to easily determine the amount of material immobilized on the array, quantifying the activity of the immobilized material is an important part of microarray validation. One important issue with microarray production and use currently is the ability to make the measurements in a nondestructive way such that the microarray can have a quality control test performed and still be similarly available for the actual test for which it was made. In the BIND Scanning system, the microarrayed proteins are quantified, assessed for activity, and the adjacent areas are all quantified similarly for ligand binding. The PWV signal that is recorded for the experiment is directly proportional to amount of mass that is bound. This approach easily allows for all the many different positive and negative references and controls that are required of an analytical system. With a label-free system capable of making binding or activity determinations in real time, not only can the quality control test be performed, but also the desirable characteristic of reversible

binding can be monitored during a kinetics off-rate experiment to further determine true functional characterization. As with most systems employing an immobilized component, attachment method plays an important role a strong preference for a uniformly oriented configuration.

BIND DATA

BIND label-free biosensors offer a resolution of binding sufficient for measuring small-molecule interactions with immobilized proteins. The dynamic range of the sensor is large enough to also allow measurement of larger biochemical entities including live cells, cell membranes, bacteria, and viruses. The following section provides data in support of the protein–protein and protein–small molecule, claims. The software for the BIND imaging system is able to translate the PWV shift into color-coded images for the pixels on the sensor as well as provide values for quantitative numerical analysis of the immobilization or binding interaction. The images can be grayscale or color enhanced to visualize the data with brighter images representing higher PWV shift values. Thus, protein spots would appear as bright areas against dark spots. Difference images for the addition of ligand can also be generated so that a visual representation of binding can made that also shows lighter areas against dark backgrounds. The data that follow use the translation tool to create images to help the reader see the variations in PWV, rather than report the data for hundreds of individual pixels.

Protein Systems

Protein microarray research involves the measurement of interactions of different sizes of proteins as well as small molecules. Data are provided below for antibody antigen interactions and several protein interactions with much smaller ligands.

Antibody

Protein A, IgG

A model system for the specific interaction of two proteins is offered in measuring of protein A capture of different species of immunoglobulin G (IgG). Figure 12.5 provides a color-enhanced PWV image of IgG from human, pig, rabbit, sheep, goat, and rat specifically binding to microarray spots of protein A that have been made on the sensor.

Experimental details — 300 nl of 100 μg/ml of protein A in water was spotted into each well in 96 well sensor plate. Following a typical incubation period, the sensor was washed 3 times with water and was blocked by applying Sea-block following suppliers recommendations. Different IgG in PBS were added across each row in the order as marked on the slide (Figure 12.5). Binding was carried out at room temperature for 30 min and then scanned. The results for the numerical analyses are shown using the slope method (as described previously) in Table 12.1. As expected, human, pig, and rabbit IgG bind stronger than sheep, goat, and rat

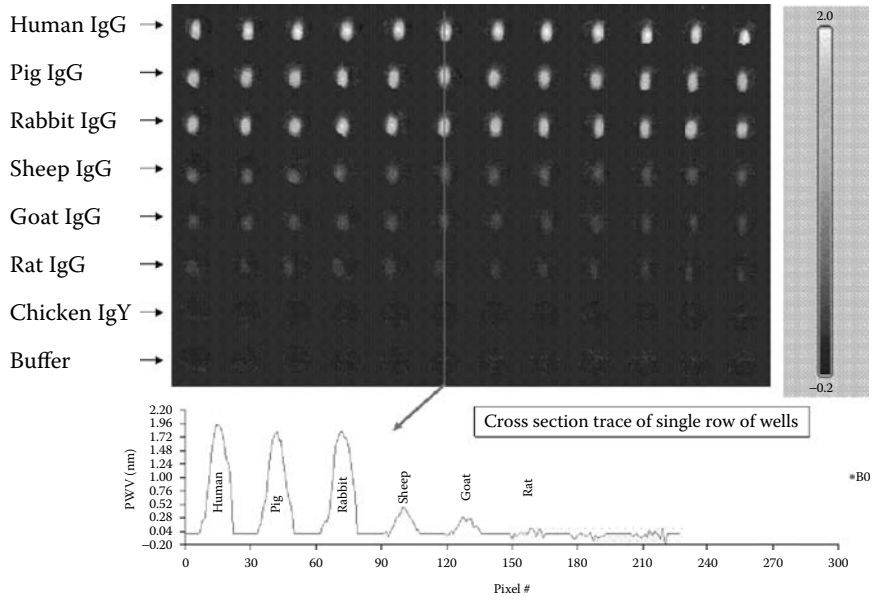
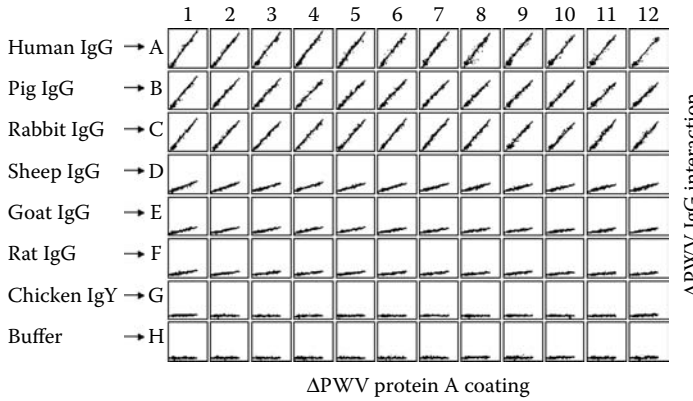


FIGURE 12.5 Spots of protein A binding to IgG of different species. Microarray spots of protein A were made onto a BIND sensor. IgG from different species were added to the isolated spots of protein A and the interactions were quantified. The lower portion of the figure shows a picture of a software tool provided for sampling cross-sections of the sensor image, plotting the pixel against the PWV shift for the pixel.

IgG, and there is not detectable signal for Chicken IgY (immunoglobulin Y) to protein A. Table 12.1 confirms this by showing steeper slopes for the tighter binding interactions. These results are in exact agreement with the supplier of the materials, Pierce.

Explanation of BIND slope analysis — Figure 12.6 shows how each pixel on the sensor provides a data point from the PWV shifts for the different read steps of the experiment on that single pixel. As an example, three different readings are made using the scanning instrument to measure a microarray. A first step records a baseline read prior to immobilization of a protein spot, a second step collects data for the amount of protein target that is immobilized, and a third step records data about how much ligand has bound to the protein spot. A pixel that receives no target is not expected to bind any ligand. This pixel would have no PWV differences for these steps and the pixel would be plotted near the origin on *x*-*y* axes (i.e., plotted *x* value equals target immobilization PWV shift while *y* value equals the ligand binding PWV shift). A best fit line is calculated for the plotted PWV shifts for all the individual pixels in that area. The magnitude of the slope indicates the binding of ligand on immobilized target protein. This approach takes advantage of ~300 independent PWV determinations within each of the protein A spots of Figure 12.5. This approach also normalizes for variances in target protein immobilizations within a spot.

TABLE 12.1
BIND Scanner Data Is Used to Quantify Binding Interaction via the Slope Method



Note: BIND Scanner data are used to quantify the binding interactions of protein A microarray spots with IgG from different species. A slope value is calculated from best fits to lines created by plotting each the change in PWV for each pixel in a given area. The *X* and *Y* values for a pixel are obtained from the PWV changes that are recorded for the immobilization of the target (giving target "concentration") vs. the PWV changes for binding of the ligand (giving "extent of binding"). The slopes recorded from this experiment are in exact agreement for values represented by PerBio, the supplier.

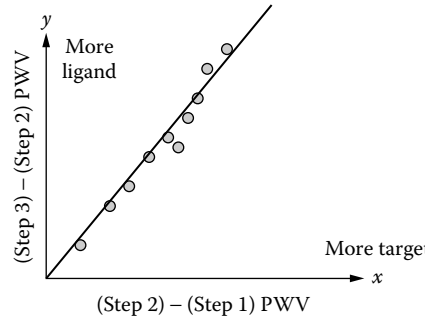


FIGURE 12.6 BIND Scanner data is used to quantify binding interaction. Difference data for steps of a BIND Scanner experiment are plotted on *x*-*y* axes pixel by pixel to quantify binding interactions. Pixels that have no target capture material are plotted near the *x* origin. These pixels should have no shift when the ligand is added at the next step and are thus expected to have *y* values near the origin. Pixels on the sensor receiving substantial target material and thus are likely to retain ligand are expected to be plotted at the other extremes of the *x* and *y* axes.

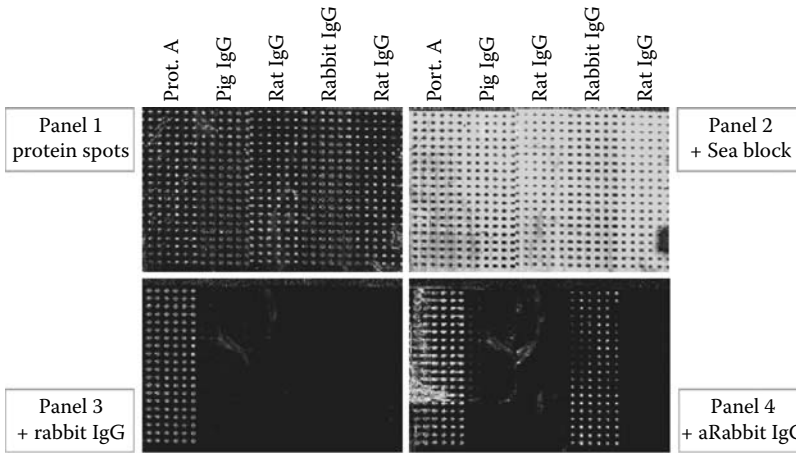


FIGURE 12.7 Multiple steps of protein–protein microarray interactions on a BIND sensor. The BIND Scanner results for a series of experimental steps are color enhanced to represent the PWV shifts occurring at the individual pixels on a protein microarray as multiprotein complexes are built up on the sensor. Brighter colors represent higher PWV shifts and greater protein attachment. Each panel has been processed to show only the difference image for each step as compared with the previous panel. Panels 1 to 4 step through spotting of target protein onto a BIND sensor, blocking of the sensor to prevent unwanted attachment of protein to the inter-spot spaces, addition and specific binding to the array of rabbit IgG, and in panel 4 addition and specific binding of anti-rabbit IgG. The far left columns of arrayed protein spots demonstrate the ability of the sensor to measure the building of multiple protein interactions. When the experiment reaches the point of panel 4, the sensor has the following proteins complexed specifically onto the spots: capture reagent, protein A, rabbit IgG, and anti-rabbit IgG.

Antibody, Antigen

Figure 12.7 demonstrates an important benefit of the BIND label-free system to quantify assembly of protein complexes. As a simple model system, protein A, blocker, rabbit IgG, and anti-rabbit IgG are reacted sequentially in panels 1 to 4 to demonstrate the ability to observe the formation of protein complexes in the static BIND system. When the experiment reaches the point of Panel 4, the sensor has the following proteins complexed specifically onto the spots: capture reagent, protein A, rabbit IgG, and anti-rabbit IgG. Protein assemblies are important in understanding human biological systems, especially their regulation.

Experimental details — Approximately six columns and 20 rows of 150 μm (<1 nl) diameter microarray spots of protein A on a 350 μm pitch were made on a BIND sensor with a contact printer (Figure 12.7). The BIND Scanner with user-adjustable image resolution with a setting at 20 μm pixel scanning resolution was used to image a microarray slide 5 times including the initial background scan. These four panels are setup to demonstrate multiplexed protein analysis on a single microarray slide. The same experiment could be performed with about 50 spots in a well of a 96-well plate. Panel 1 shows the color-enhanced PWV shift image for the spotting of various

proteins, some of which are antibodies. Panel 2 shows the PWV shift difference (from the previous step) image for the blocking step where SeaBlock (Pierce) was used. In this panel, only the areas of the sensor that do not already have protein spots (from Panel 1) are found to "light up" as the image is a difference image generated by subtracting the PWV shift of the previous step. Panel 3 shows the PWV shift difference image when the same slide is challenged with the addition of Rabbit IgG. The panel 3 image is generated as a difference image by subtracting the PWV shift of step 2. Only the spots of protein A bind the antibody as expected. Panel 4 shows the result for a fourth addition to the same slide of an anti-rabbit IgG. In this case, the first area, which had Rabbit IgG added in Panel 3, "lights up" and the area that had Rabbit IgG (see Panel 1) spotted originally also "lights up." Other areas of the slide with proteins spotted that did not interact with the Rabbit IgG did not have PWV shift signal.

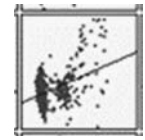
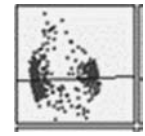
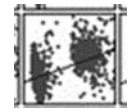
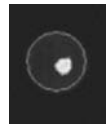
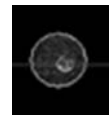
Enzyme System

The BIND system has high sensitivity and dynamic range in addition to quantitative ability. This allows the immobilization of individual proteins in microarray spots and quantifying the binding of small molecules, thus providing information that might be sought in a chemigenomics type experiment. The sections below demonstrate this capability for four model proteins and their small molecule ligands: carbonic anhydrase II, protein kinase A, streptavidin, and human serum albumin. Using the slope method described above, the PWV shifts for the protein immobilization and ligand addition steps are plotted and fitted to a linear analysis. When the molecular weights of the target protein and the ligand are known, a slope value for complete binding of the immobilized protein can be calculated as shown in the far right column of Table 12.2. The slope describes what fraction of target is bound by ligand. If concentrations of the ligand above the equilibrium binding constant and in excess of the target concentration are added to the microarrayed protein, a good assessment of the activity of the protein can be made. The "Notes" column of Table 12.2 gives values for the expected binding based upon calculations with molecular weights and stoichiometry of every target molecule by the appropriate number of ligand molecules.

Carbonic Anhydrase II

Carbonic anhydrase II (CA) is a 33,000-molecular-weight enzyme that has among its well characterized ligands a 201 molecular weight compound, 4-carboxybenzenesulfonamide (CBS). Microarray spots of the protein were made on a BIND sensor with 50 nl applications of 1 mg/ml or 0.5 mg/ml target concentrations, yielding ~2.5 nm PWV for the covalent immobilization of the protein. The protein spots were challenged with 10 μ*M* CBS in up to 5% DMSO. Row 1 of Table 12.2 shows a ring spot of the protein and the ring image of the specific area of the sensor the ligand interacts. The binding interaction has a literature value for equilibrium binding of 700 n*M* and a reported value for BIND of ~900 n*M*. A comparison of the empirical slope with the calculated slope suggests that the protein may only be about 64% active in this instance.

TABLE 12.2
Details of Microarrayed Proteins Binding Small Molecules on a BIND Sensor

Protein Name	Protein Image	Ligand Image	Slope Image	Slope Value	Notes
Carbonic anhydrase II				Avg 4 wells 0.015	Target 0.0233
Protein kinase A				0.0048	Target 0.011
Streptavidin				0.039	Target 3site 0.038 4site 0.051
Human serum albumin				0.0082	Target 1site 0.00513 2site 0.01026

Note: BIND Scanner images and data is reported for the PWV shifts of target protein immobilization and ligand binding to the microarray spot of protein. The empirical slope values reported are in good agreement with calculated values for expected slopes for significantly active protein.

Protein Kinase A

A 40,000-molecular-weight form of Protein kinase A was spotted onto a BIND sensor using 50 nl drops and 1.0 or 0.5 mg/ml protein. PWV shifts of ~5.2 nm were recorded for the immobilization of the kinase. A small molecule, staurosporin, at 466 molecular weight was applied to the kinase spots at 10 μ*M* concentration, well above the solution IC_{50} reported in the literature.[24] A comparison of the empirical and calculated slopes suggest that the protein may be about 44% active. As with the carbonic anhydrase data above, this reduced level of activity

may be a function of the random orientation of the protein on the surface of the sensor or of the capture method that may have disabled the binding site for the small molecule. Oriented capture method would greatly improve the activity of the protein. Nonetheless, the BIND technology is able to report on the activity of the immobilized protein and has been used now to characterize a number of different types of proteins.

Other Proteins

Streptavidin

Streptavidin is a well-characterized 55 kDa protein with high affinity four-site binding, providing a robust system for the study of protein small molecule interactions with low p*M* affinity constants. In a more sophisticated approach, this model protein can be used to monitor for distortions that may occur to the protein that inhibit full binding of biotin into the four available sites on each streptavidin molecule. Spots of 50 nl of 0.5 mg/ml streptavidin were made on a BIND sensor. The resulting microarrayed protein spots of 3 to 4 nm PWV shifts were recorded following incubation of ~4 h. The arrayed streptavidin spots were challenged with 500 n*M* biotin addition. Table 12.2, row 3, details the PWV shift images and the slopes for the binding interactions. Calculations for three-site binding on streptavidin give slope predictions of 0.038 and for four-site binding predicts a slope of 0.051. The empirical slope of 0.039 indicates between three- and four-site binding of the microarrayed protein.

Human Serum Albumin

Human serum albumin (HSA) is a ubiquitous protein of ~60,000 molecular weight that has among its well-characterized ligands a 308 molecular weight compound, warfarin. Warfarin has been variously reported to bind HSA with a first affinity of 3 to 7 μ*M* and a capacity for 2 to 3 more binding events per HSA molecule. Both the protein and the small molecule offer a challenge to a microarray system in that both have a nonpolar nature that makes them bind to many kinds of material often in undesirable ways. Table 12.2, row 4, details images and data for the immobilization of HSA spots on a BIND sensor. Spots of 500 nl, using concentrations up to 1 mg/ml protein were made and allowed covalent attachment of the protein. Between 7.5 and 8 nm PWV shifts for HSA immobilization were recorded prior to the addition of 50 μ*M* warfarin in 1% DMSO. Calculated values for single site binding show a slope of 0.00513 and two-site binding with a slope of 0.01026. The data show that between one and two sites are bound.

DESCRIPTION OF SYSTEM LIMITS

Figure 12.8 demonstrates one example of the current lower limit of detection for the concentration of added ligand is about 50 to 100 ng/ml for materials that bind in a specific manner to the sensor. This value is primarily a limitation of binding of low concentration materials to tips, tubes, and side walls of the sensor frame. With small molecule direct binding to immobilized proteins, detection of molecules with

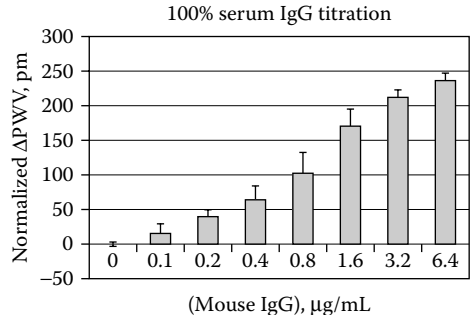


FIGURE 12.8 BIND sensor LLD in complex media. The data show the current BIND sensor has a typical lower limit of detection of biomaterials in 100% human serum of ~100 ng/ml. This limit is believed to be strongly influenced by retention of materials to tips, tubes, and walls of wells. The sensor can be tuned for sensitivity and dynamic range by control of the amount of capture reagent immobilized on the sensor surface.

molecular weights as low as 70 Da has been accomplished. The current typical technical limit of the sensor is about 1.33 pg/mm^2 as determined by correlation work with sensitive radioactivity titrations on to the sensor surface.[25,26] Resolution improves for higher density immobilized active protein concentrations, and depending on the protein, this can yield a 2–5 × resolution enhancement. Other reports have detailed the ability of the sensor to perform well despite the complex nature of the sample (see Figure 12.8). By using the BIND microarray scanner, spots of 50 µm diameter provide ~40 independent pixel readings using the 7 µm resolution setting of the scanning instrument. The time to scan the 4" × 6" area of an entire 384-well plate at ~150 µm resolution is under 2 minutes. Work with the BIND system has provided quantitative and sensitive detection of active protein with low variance for samples comprised of 100% human plasma, 100% human serum, or periplasmic extracts, all without the need for wash steps to measure the binding interaction. This has the great benefit of allowing measurement of weaker binding interactions or interactions with fast off-rates that normally cannot be viewed with label systems and their requisite wash steps.

FUTURE DIRECTIONS

Data and limits described herein are recorded using current commercially available sensors and instruments. Research and development of the next generations of sensors with better sensitivity and lower levels of detection are always ongoing in the areas of new surfaces, improved sensor design, multiple wavelength detections and enhancements, optimized attachment methods, various content added, value added formats (preformed for specific target capture situations), and flowed systems for kinetics measurements further characterizing the arrayed protein interaction. The main technical hurdles to widespread adoption of label-free detection have been

lacking of three critical elements, of sensitivity, of quantitative analysis, and of throughput. Additional factors affecting widespread adoption are a high cost/assay and instrument complexity. The first generation of products based upon photonic crystal label-free optical biosensors already demonstrate the ability to detect and quantify interactions of low molecular weight chemical compounds binding to high molecular weight proteins and the ability to image and detect attachment, proliferation, and apoptosis of individual cells in the same sensor and instrument platform. Similar to the way that silicon transistors have evolved from their early embodiments to today's high-performance integrated circuits, photonic crystal biosensors will also continue to develop as new design features, materials, and instrumentation approaches further push the limits of sensitivity and detection resolution. Already, devices with 5× higher sensitivity performance than the sensors used in the assays reported in this review have been routinely demonstrated in the laboratory.[27,28] Combinations of this label-free technology with other instrumental techniques such as mass spectroscopy and fluorescence spectroscopy have already been contemplated and tested. Advances in these areas will provide truly universal detection, quantification, and identification methods for any type of molecule in any type of media. The benefit to the microarray community will be that their work can move to a higher level of analytical characterization of protein–protein, protein–small molecule, and protein–cell interactions. This is necessary to fully understand the human proteome with less waiting and speculating on tools that are not fundamentally designed to, and therefore cannot, provide the detailed characterizations they need.

CONCLUSION AND SUMMARY

The SRU BIND free-space optics biosensor system has provided label-free data for protein–protein, protein–small molecule, and specific interactions with single cells. Results from work with the BIND label-free system demonstrate the ability to quantify protein attachment, activity, and specific binding interactions at high resolution of spotted biological materials on microtiter plates or microarray slides. This type of application should directly satisfy two of the top current needs of the protein microarray research community.

ACKNOWLEDGMENTS

As with all small technology-driven enterprises, the contributions from each SRU employee past and present has in some way contributed to the success of the current work. Their hard work, innovation and determination were essential for making the work shown here successful. The authors are especially grateful to John Gerstenmaier, Dr. Frank Wang and Dr. Christine Genick, for their efforts in performing experiments and providing the bioapplications data shown here, and to Dr. Peter Li for his efforts directing instrument and data systems development for BIND. Dr. Genick is acknowledged for her contributions in providing critical review and helpful suggestions with the manuscript.

REFERENCES

1. Brecht, A. and Gauglitz, G., Optical probes and transducers, *Biosens. Bioelectron.*, 10, 923, 1995.
2. Cunningham, A. J., *Introduction to Bioanalytical Sensors*, Wiley-Interscience Publications, John Wiley & Sons, New York, 1998, p. 418.
3. Scherer, A. et al., *J. Korean Phys. Soc.*, 42, 768, 2003.
4. Arakawa, T. and Kita, Y., Refractive index of proteins in organic solvents, *Analyt. Biochem.,* 271, 119, 1999.
5. Cunningham, B.T. et al., *Sens. Actuat.* B 81, 316, 2002.
6. Haes, A.J. and Duyne, R.P.V., A nanoscale optical biosensor: Sensitivity and selectivity of an approach based on the localized surface plasmon resonance spectroscopy of triangular silver nanoparticles, *J. Am. Chem. Soc.*, 124, 10596, 2002.
7. Magnusson, R. and Wang, S.S., *Appl. Opt.*, 34, 8106, 1995.
8. Magnusson, R. and Wang, S.S., *Appl. Phys. Lett.*, 61, 1022, 1992.
9. Cunningham, B.T. et al., *Sens. Actuat. B*, 85, 219, 2002.
10. Chan, L.L. et al., *Sens. Actuat. B*, accepted, February 2006.
11. Chan, L.L. et al., *IEEE Sens. J.*, accepted, February 2006.
12. Hessel, A. and Oliner, A.A., *Appl. Opt.*, 4, 1275, 1965.
13. Wood, R.W., *Philos. Mag*., 4, 396, 1902.
14. Li, P. et al., Label-free assays on the BIND system, *Sens. Actuat. B*, 99, 6, 2004.
15. Miklos, G.L.G. and Maleszka, R., Microarray reality checks in the context of a complex disease, *Nat. Biotechnol.*, 22, 615, 2004.
16. Tseng, G.C. et al., *Nucl. Acids Res.*, 29, 2549, 2001.
17. Stoll, D. et al., Protein microarrays: Applications and future challenges, *Curr. Opin. Drug Disc. Dev.*, 8, 239, 2005.
18. Kroll, T.C. and Wolfl, S., Protein microarrays: Applications and future challenges, *Nucl. Acids Res.*, 30, e50, 2002.
19. Wu, W. et al., Evaluation of normalization methods for cDNA microarray data by k-NN classification, *BMC Inform.*, 6, 191, 2005.
20. Fan, J. et al., Normalization and analysis of cDNA microarrays using within-array replications applied to neuroblastoma cell response to a cytokine, *Proc. Natl. Acad. Sci. USA*, 101, 1135, 2004.
21. Stoll, D. et al., Protein microarray technology, *Front. Biosci.*, 7, c13, 2002.
22. Lakowicz, J.F., *Principles of Fluorescence Spectroscopy*, Plenum Press, New York, 1983.
23. Espina, V. et al., Use of proteomic analysis to monitor responses to biological therapies, *Expert Opin. Biol. Ther.*, 4, 83, 2004.
24. Meyer, T. et al., A derivative of staurosporine (CGP 41 251) shows selectivity for protein kinase C inhibition and *in vitro* anti-proliferative as well as *in vivo* anti-tumor activity, *Int. J. Cancer*, 43, 851, 1989.
25. Cunningham, B.T. et al., Label-free assays on the BIND system, *Sens. Actuat. B*, 87, 365, 2002.
26. Lin, B. et al., *Biosens. Bioelectron.*, 17, 827, 2002.
27. Block, I.D., Chan, L.L., and Cunningham, B.T., *Microelectron. Mater.*, accepted, June 2006.
28. Block, I.D., Chan, L.L., and Cunningham, B.T., *Sens. Actuat. B*, accepted, January 2006.

Section 4

Applications of Functional Protein Microarrays

13 Studying Protein–Protein Interactions with Protein Microarrays: Rapid Identification of 14-3-3 Protein Binding Partners

Jun-ichi Satoh

CONTENTS

INTRODUCTION

The 14-3-3 Protein Acts as a Molecular Adaptor in Signaling Networks

The 14-3-3 protein family in mammalian cells consists of evolutionarily conserved, acidic 30-kDa proteins composed of seven isoforms named β, γ, ε, ζ, η, θ, and σ.[1,2] A homodimeric or heterodimeric complex composed of the same or distinct isoforms constitutes a large cup-like structure possessing an amphipathic groove with two ligand-binding capacity, and acts as a molecular adaptor by interacting with key signaling components of cell differentiation, proliferation, transformation, and apoptosis. The dimeric 14-3-3 protein regulates the function of target proteins by restricting their subcellular location, bridging them to modulate catalytic activity, and protecting them from dephosphorylation or proteolysis.[3,4] Although 14-3-3 is widely distributed in neural and nonneural tissues, it is expressed at the highest level in neurons in the central nervous system (CNS).[5,6] Aberrant expression and impaired function of 14-3-3 in the CNS are closely associated with pathogenetic mechanisms of various neurological disorders, such as Creutzfeldt-Jacob disease,[7–9] Alzheimer disease,[10] Pick disease,[11] Parkinson disease,[12,13] multiple system atrophy,[14,15] spinocerebellar ataxia,[16] amyotrophic lateral sclerosis,[17] Miller-Diecker syndrome,[18] multiple sclerosis,[19,20] and mitochondrial encephalopathy with lactic acidosis and stroke-like episodes (MELAS).[21,22]

In general, the 14-3-3 protein interacts with phosphoserine-containing motifs of its ligands, such as RSXpSXP (mode I), RXXXpSXP (mode II), and pS/pT($X_{1\text{-}2}$)COOH (mode III), in a sequence-specific manner.[23,24] Until present, more than 300 proteins have been identified as being 14-3-3-binding partners. They include Raf-1 kinase, Bcl-2 antagonist of cell death (BAD), protein kinase C (PKC), phosphatidylinositol 3-kinase (PI3K), and cdc25 phosphatase.[1,2,25] Binding of 14-3-3 to Raf-1 is indispensable for its kinase activity in the Ras-MAPK signaling pathway, while the interaction of 14-3-3 with BAD, when phosphorylated by a serine/threonine kinase Akt, inhibits apoptosis. Furthermore, recent studies indicate that the 14-3-3 protein may also interact with a set of target proteins in a phosphorylation-independent manner.[26–29] Increasing our knowledge of molecular interactions between 14-3-3 and target proteins would greatly help us to understand the biological function and pathological implication of the 14-3-3 protein networks.

The Advantages of Protein Microarray Analysis to Identify Protein–Protein Interactions

The yeast two-hybrid (Y2H) system is a powerful approach to identify novel protein–protein interactions in a high-throughput fashion.[30,31] However, Y2H screening requires a lot of time and effort, and is often criticized for detecting the interactions

unrelated to the physiological setting and obtaining high rates of false positive interactors caused by spontaneous activation of reporter genes and self-activating bait proteins.[32,33] Affinity purification coupled with mass spectrometry (APMS) is an alternative approach to identify the components of protein complexes on a large scale. This approach has been taken to identify a wide range of 14-3-3-interacting proteins involved in the dynamic control of cytoskeletons,[34] cell cycle regulation,[35] biosynthetic metabolism,[36] and oncogenic signaling events.[37] Although APMS screening detects binding partners of physiological significance, it is also time-consuming and expensive, requires a large amount of samples, and has a difficulty in detecting transmembrane proteins and loosely associated components that might be lost during purification.[38] Furthermore, the recognized interaction is not always direct, assisted by intermediary molecules.

Recently, protein microarray technology has been established for the rapid, systematic, and less expensive screening methods of thousands of protein–protein, protein–lipid, and protein–nucleic acid interactions in a high-throughput fashion.[39–43] It requires small sample volumes and affords the ability to control the experimental parameters, such as buffer pH, ion concentration, and reaction cofactors in a reproducible manner. This approach has diverse applications to discovery-based proteomics in the field not only of basic biological research but also of drug and biomarker discovery research, including identification of the substrates of protein kinases, the protein targets of small molecules, the consensus interaction of transcription factors, and autoantibody profiling.[44–51] Thus, this technology sounds pivotal for establishment of personalized medicine. The vast majority of protein–protein interactions occur between a domain located in one protein and a small motif spanning usually 8 to 15 amino acids in its ligand. They promote multimolecular protein complex formation that regulates diverse signaling networks. A recent study using the microarray containing 212 spots of protein domains, composed of two conserved tryptophans (WW), two conserved phenylalanines (FF), Src homology 2 (SH2), Src holmology 3 (SH3), pleckstrin homology (PH), forkhead-associated (FHA), PSD-95, DLG and ZO-1 proteins (PDZ), and 14-3-3-interacting modules, characterized the domain-specific binding profile of various signaling molecules in a single experiment.[52] More recently, the epidermal growth factor receptor (EGFR) signaling network was studied by using protein microarrays that contain virtually all SH2 and phosphotyrosine binding (PTB) domains encoded in the human genome, and probing with phosphotyrosine (pY)-containing peptides derived from EGFR, ErbB2, and ErbB3.[53]

Here, we have attempted to characterize a comprehensive human 14-3-3 interactome by analyzing a high-density protein microarray.

EXPERIMENTAL PROTOCOLS

Preparation of an Epitope-Tagged Probe for Microarray Analysis

Human embryonic kidney cells HEK293 whose genome was modified for the Flp-In system (Flp-In 293) were obtained from Invitrogen, Carlsbad, CA. Flp-In 293 cells contain a single Flp recombination target (FRT) site targeted for the site-specific

recombination, integrated in a transcriptionally active locus of the genome, where it stably expresses the *lacZ*-Zeocin fusion gene driven from the pFRT/*lac*Zeo plasmid under the control of SV40 early promoter. Flp-In 293 cells were maintained in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 100 U/ml of penicillin, and 100 μg/ml of streptomycin (feeding medium) with inclusion of 100 μg/ml of Zeocin (Invitrogen) according to the methods described previously.[54]

To prepare the probe for protein microarray analysis, the open reading frame (ORF) of the human 14-3-3ε gene (YWHAE, GenBank accession No. NM_006761, amino acid residues 2 to 255) was amplified from cDNA of NTera 2-N cells, a model of differentiated human neurons in culture,[55] by PCR using PfuTurbo DNA polymerase (Stratagene, La Jolla, CA, USA) and the sense (5'gatgatcgagaggatctggtgtac3') and antisense (5'ctgattttcgtcttccacgtcctg3') primers. The PCR product was then cloned into a mammalian expression vector pSecTag/FRT/V5-His TOPO (Invitrogen) to produce a fusion protein with a C-terminal V5 (GKPIPNPLLGLDST) tag, a C-terminal polyhistidine (6xHis) tag, and an N-terminal Ig κ-chain secretion signal. This vector, together with the Flp recombinase expression vector pOG44 (Invitrogen), was transfected in Flp-In 293 cells by Lipofectamine 2000 reagent (Invitrogen) (Figure 13.1). A stable cell line was established after incubating the cells for approximately one month in the feeding medium with inclusion of 100 μg/ml of Hygromycin B (Invitrogen). The stable cell line was named 293eV5.[56] In this system, the recombinant protein was secreted into the culture medium after the Ig κ-chain secretion signal sequence was processed by an endogenous signal peptidase-mediated cleavage. Therefore, it has an advantage of easily purifying the recombinant protein, compared with the system where the recombinant protein is expressed in the cytoplasm, mixed with various unnecessary proteins.

To purify the recombinant 14-3-3ε protein, the culture supernatant of 293eV5 incubated in the serum-free DMEM/F-12 medium for 48 hours was harvested and concentrated at an 1/40 volume by centrifugation on an Amicon Ultra-15 filter (Millipore, Bedford, MA). It was then purified by the HIS-select spin column (Sigma, St. Louis, MO), and concentrated at a 1/10 volume by centrifugation on a Centricon-10 filter (Millipore). The purity and specificity of the probe were verified by Western blot analysis using mouse monoclonal anti-V5 antibody (Invitrogen) and rabbit polyclonal antibody specific for the 14-3-3ε isoform (IBL, Gumma, Japan) (Figure 13.1).

Protein Microarray Analysis

ProtoArray human protein microarray (v1.0; Invitrogen) we utilized contains 1752 human proteins of various functional classes spotted in duplicate on a nitrocellulose-coated glass slide. (After a quality control procedure, the number of total arrayed proteins is reduced from 1900 originally listed in the array.) Nitrocellulose-coated surface provides a nearly quantitative retention of the spotted proteins and significantly higher detection sensitivity than the other surfaces.[40] All the proteins immobilized on the array were expressed as an N-terminal glutathione-S transferase (GST)-6xHis fusion protein derived from the genes selected from the human

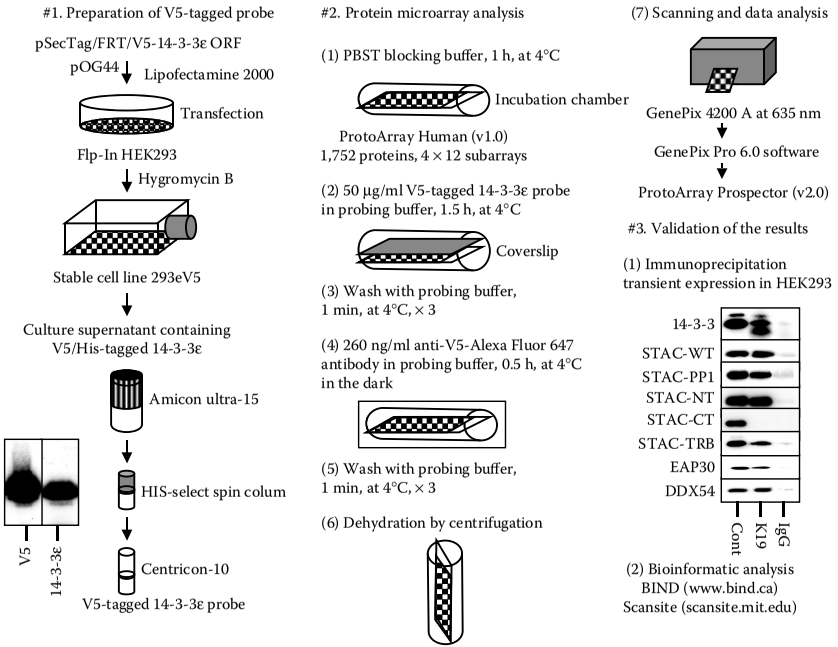


FIGURE 13.1 Protein microarray analysis of 14-3-3-binding proteins. The experimental protocol is comprised of the following three steps. (1) Preparation of V5-tagged probe. The recombinant human 14-3-3ε protein tagged with V5 was purified from the concentrated culture supernatant of a stable cell line 293eV5. The purity and specificity of the probe were verified by Western blot analysis using anti-V5 antibody or anti-14-3-3ε antibody. (2) Protein microarray analysis. After blocking nonspecific binding, the microarray containing 1752 human proteins was incubated with the probe, followed by incubation with anti-V5 antibody labeled with Alexa Fluor 647, and scanned on the microarray scanner. The significant binding was identified by analyzing the data with the ProtoArray Prospector software. (3) Validation of the results. The specific binding to 14-3-3 of STAC (WT, wild type; PP1, WT with inclusion of protein phosphatase-1 during protein extraction; NT, the N-terminal half; CT, the C-terminal half; TRB, the truncated form lacking both the RYYSSP motif and the cysteine-rich domain), EAP30 and DDX54 was validated by immunoprecipitation (IP) analysis of the corresponding recombinant proteins expressed in HEK293 cells. Lanes represent the input control (Cont), and IP with K-19 or normal rabbit IgG. The 14-3-3 interactors identified by protein microarray analysis were further evaluated by bioinformatic analysis of protein–protein interaction networks via BIND and Scansite database searches.

ultimate ORF clone collection (Invitrogen). They represent either the full-length or the partial fragment of recombinant proteins. They were expressed in Sf9 insect cells by using the Bac-to-Bac Baculovirus expression system (Invitrogen), purified under non-denaturating conditions by glutathione affinity chromatography in the presence of protease inhibitors, and processed for spotting on the slides (Invitrogen application note).

The proteins are spotted in an arrangement composed of 4 × 12 subarrays equally spaced in vertical and horizontal directions. Each subarray includes 16 × 16 spots, composed of 48 control spots (C), 80 human proteins (P), and 128 blanks (B) (Figure 13.2a). The controls include 14 positive control spots; four spots of an Alexa Fluor 647-labeled antibody (rows 1, 8; columns 1, 2), six spots of a concentration gradient of a biotinylated anti-mouse antibody with a capacity to bind to mouse monoclonal anti-V5 antibody conjugated with Alexa Fluor 647 (row 8; columns 3 to 8), and four spots of a concentration gradient of V5 protein (row 8; columns 13 to 16). They also include 34 negative control spots; six spots of a concentration gradient of bovine serum albumin (BSA) (row 1; columns 3 to 8), four spots of a concentration gradient of a rabbit anti-GST antibody (row 1; columns 9 to 12), four spots of a concentration gradient of calmodulin (row 1; columns 13 to 16), 16 spots of a concentration gradient of GST (row 2; columns 1 to 16), two spots of buffer only (row 8; columns 9,10), and two spots of an anti-biotin antibody (row 8; columns 11, 12).

Nonspecific binding was blocked by incubating the microarray for 60 min at 4°C in the PBST blocking buffer composed of 1% BSA and 0.1% Tween 20 in phosphate-buffered saline (PBS), as described previously (Figure 13.1).[56] Then, it was incubated for 90 min at 4°C with the probe described above at a concentration of 50 μg/ml in the probing buffer composed of 1% BSA, 5 m*M* $MgCl_2$, 0.5 m*M* dithiothreitol (DTT), 0.05% Triton X-100, and 5% glycerol in PBS. The array was washed three times with the probing buffer, followed by incubation for 30 min at 4°C with mouse monoclonal anti-V5 antibody conjugated with Alexa Fluor 647 (Invitrogen) at a concentration of 260 ng/ml in the probing buffer. The array was washed three times with the probing buffer, dehydrated by brief centrifugation, and then scanned by the GenePix 4200A scanner (Axon Instruments, Union City, CA) at a wavelength of 635 nm. The data in a format specified by the GenePix Pro 6.0 microarray data acquisition software (Axon Instruments) were analyzed by using the ProtoArray Prospector software v2.0 (Invitrogen) following acquisition of the microarray lot-specific information online (www.invitrogen.com/protoarray). The spots showing the background-subtracted signal intensity value greater than the median plus three standard deviations of intensities of all protein features were considered as having a significant binding.

The Z-Score was calculated by the following formula: $Z_k = (X_k - \mu_s)/\sigma_s$, where X_k represents the signal intensity value of the k^{th} protein feature, μ_s is the mean signal intensity of all protein features, and σ_s expresses the standard deviation of intensities of all protein features. The Z-Score reflects a binding specificity determined by the definition how far and in what direction a signal from a specific protein feature deviates from the mean signal intensity of all the protein features.

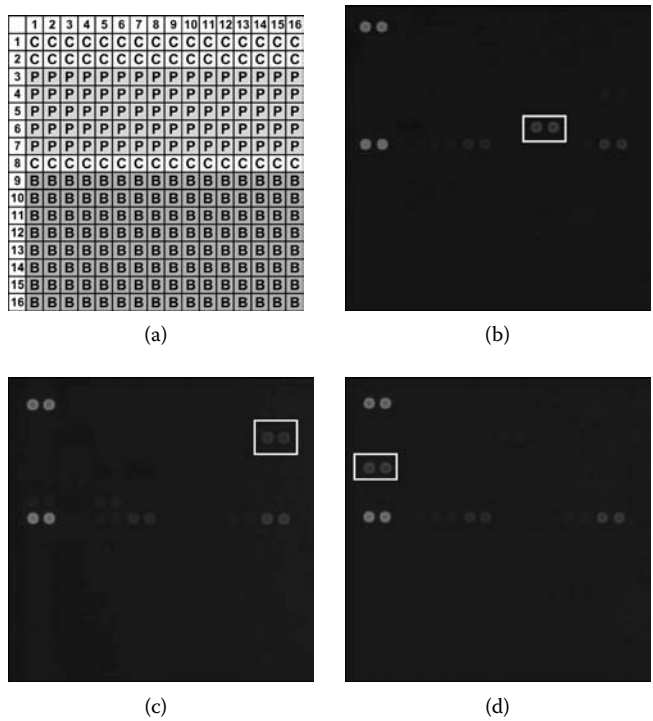


FIGURE 13.2 Detection of 14-3-3-binding proteins on protein microarray. The microarray we utilized contains 1752 distinct human proteins of various functional classes spotted in duplicate on a nitrocellulose-coated glass slide. They are printed in an arrangement of 4×12 subarrays equally spaced in vertical and horizontal directions. (a) Layout of the subarray. Each subarray includes 16×16 spots composed of 48 control spots (C), 80 human proteins (P), and 128 blanks (B). (b) EAP30 on the subarray 1. The spots of (row 7; column 1) and (row 7; column 12) indicated by a square represent EAP30. (c) DDX54 on the subarray 27. The spots of (row 3; column 15) and (row 3; column 16) indicated by a square represent DDX54. (d) STAC on the subarray 39. The spots of (row 5; column 1) and (row 5; column 2) indicated by a square represent STAC. In these subarrays (b–d), the positive control spots represent an Alexa Fluor 647-labeled antibody (rows 1, 8; columns 1, 2) that provides the strong signals, a concentration gradient of a biotinylated anti-mouse antibody with a capacity to bind to mouse monoclonal anti-V5 antibody labeled with Alexa Fluor 647 (row 8; columns 3 to 8), and a concentration gradient of V5 protein (row 8; columns 13 to 16). The signals are only visible at the higher concentration in the latter two.

VALIDATION AND EVALUATION OF THE RESULTS OF PROTEIN MICROARRAY ANALYSIS

Transient Expression of Recombinant Proteins in HEK293 Cells

To verify the results of protein microarray analysis, the ORF of the genes encoding EAP30 subunit of ELL complex (EAP30, NM_007241, amino acid residues 2 to 258), dead box polypeptide 54 (DDX54, NM_024072, amino acid residues 2 to 881), and

src homology three (SH3) and cysteine rich domain (STAC, NM_003149, amino acid residues 2 to 402, full-length) were amplified by PCR using PfuTurbo DNA polymerase and the specific primer sets (5'caccgccgcggggtgggagctggc3' and 5'tcaggggagggcttctctg-gcctc 3' for EAP30; 5'gcggccgacaagggcccggcggct3' and 5'tcacatcctcttccgcatcttgcc3' for DDX54; and 5'atccctccgagcagcccccgcgag3' and 5'tcagatgttttctagtacatcaag3' for STAC). The N-terminal half of STAC (amino acid residues 2 to 333, NTF), the C-terminal half of STAC (amino acid residues 234 to 402, CTF), and two distinct truncated forms of STAC (amino acid residues 2 to 164 named TRA and amino acid residues 2 to 105 named TRB) were amplified using the corresponding primer sets (5'atccctccgagcagcccccgcgag3' and 5'tcaaagatctgaagtagaggttct3' for NTF; 5'gtggaggt-tcctgaggaagccaat 3' and 5'tcagccacctggatgcagaccagc3' for CTF; 5'atccctccgagcagc-ccccgcgag 3' and 5'tcatggcagcttgcccatgcaccg3' for TRA; and 5'atccctccgagcagcccccc-gcgag 3' and 5'tcagccacctggatgcagaccagc3' for TRB).

They were then cloned into a mammalian expression vector pcDNA4/HisMax-TOPO (Invitrogen) to produce a fusion protein with an N-terminal Xpress tag. To express the STAC mutant with a single amino acid substitution S172A (the single mutant; SMT) or with double amino acid substitutions S172A and S173A (the double mutant; DMT), the pcDNA4/HisMax-TOPO vector containing the full-length wild-type (WT) STAC gene was modified by consecutive site-directed mutagenesis using QuikChange II site-directed mutagenesis kit (Stratagene) and the primer sets (5'gtttcggcgttactacgcctccccttgctcattc3' and 5'gaatgagcaaggggaggcgtagtaacgccgaaac 3' for SMT and 5'cggcgttactacgccgcccccttgctcattcat3' and 5'atgaatgagcaaggggcggc-gtagtaacgccg3' for DMT). All these vectors were transfected in HEK293 cells by Lipofectamine 2000 reagent.

Coimmunoprecipitation Analysis

For coimmunoprecipitation analysis, total protein extract was prepared by homogenizing the cells in M-PER lysis buffer (Pierce, Rockford, IL) supplemented with a cocktail of protease inhibitors (Sigma), either with inclusion of phosphatase inhibitors (Sigma) to maintain the protein phosphorylation status or with inclusion of recombinant protein phosphatase-1 (PP1) catalytic subunit α-isoform (5 U/ml; Sigma) instead of phosphatase inhibitors to induce the protein dephosphorylation reaction.[57] The homogenate was centrifuged at 12,000 rpm for 20 min at 4°C. After preclearance, the supernatant was incubated for 3 hours at 4°C with 30 μg/ml rabbit polyclonal anti-14-3-3 protein antibody (K19)-conjugated agarose (Santa Cruz Biotechnology, Santa Cruz, CA) or the same amount of normal rabbit IgG-conjugated agarose (Santa Cruz Biotechnology). After several washes, the immunoprecipitates were processed for Western blot analysis using mouse monoclonal anti-14-3-3 protein antibody (H-8, Santa Cruz Biotechnology) and mouse monoclonal anti-Xpress antibody (Invitrogen). K-19 and H-8 antibodies recognize all 14-3-3 isoforms. The specific reaction was visualized by using a chemiluminescent substrate (Pierce).

Bioinformatic Analysis

In addition to validation of the specific interactions by wet experiments, we evaluated them by bioinformatic analysis. The information on known 14-3-3 interactors, molecular

function, molecular weight, and subcellular localization was obtained from Biomolecular Interaction Network Database (BIND; www.bind.ca), Human Protein Reference Database (HPRD; www.hprd.org), Prediction of Protein Sorting Signals and Localization Sites in Amino Acid Sequence Database (PSORT II; psort.ims.u-tokyo.ac.jp), and PubMed Database (www.pubmed.gov). The 14-3-3-binding consensus motif mode I (RSXpSXP) located in target proteins was surveyed by the Scansite 2.0 Motif Scanner (scansite.mi.edu),[58] which assesses the probability of a site matching the candidate motif under high, medium or low stringent conditions (Figure 13.3).

RESULTS

Protein Microarray Analysis Identified 20 Distinct 14-3-3-Interactors

Western blot analysis verified the purity and specificity of the recombinant 14-3-3ε protein tagged with V5 (Figure 13.1). Among 1752 proteins on the microarray, 20 were identified as the proteins showing significant binding to the probe, all of which were previously unreported 14-3-3-binding partners by the BIND search.[56] Seven were categorized into hypothetical clones of uncharacterized function, derived from either the Mammalian Genome Collection (MGC) or the Full-Length Long Japan (FLJ). They include FLJ10415 (GenBank accession number NM_018089), LOC57228 (NM_020467), MGC17403 (NM_152634), LOC137781 (BC032347), LOC92345 (NM_138386), FLJ10156 (NM_019013), and FLJ25758 (NM_001011541). Thirteen proteins with annotation are as follows:

1. *EAP30 subunit of ELL complex (EAP30; NM_007241)* (Figure 13.2b). This is a 30-kDa component of the ELL complex (estimated MW is 28,866 suggested by HPRD; putative subcellular location is cytoplasmic suggested by PSORT II), which confers derepression of transcription by RNA polymerase II.[59] EAP30 is also named VPS22, a component of the ESCRT-II endosomal sorting complex that plays a key role in the multivesicular body (MVB) pathway.[60] The 14-3-3-binding consensus motif mode I is not identified by the Scansite Motif Scanner, although the Z-Score of two corresponding spots on the array shows the highest values, 22.9 and 24.6 respectively. The similarity in the scores between distinct spots supports the reproducibility of the results of protein microarray analysis.
2. *Lymphocyte cytosolic protein 2 (LCP2; NM_005565).* This is a 72-kDa protein (MW 60,191; nuclear), alternatively named SH2 domain-containing leukocyte protein of 76kD (SLP76), which associates with the Grb2 adaptor protein and provides a substrate of the ZAP-70 protein tyrosine kinase.[61] LCP2 plays a key role in promoting T cell development and activation. It contains three mode I motifs with low stringency; pS297 (TTERHERSSPLPGKK), pS376 (SSFPQSASLPPYFSQ), and pT456 (DSSKKTTTNPYVLMV).

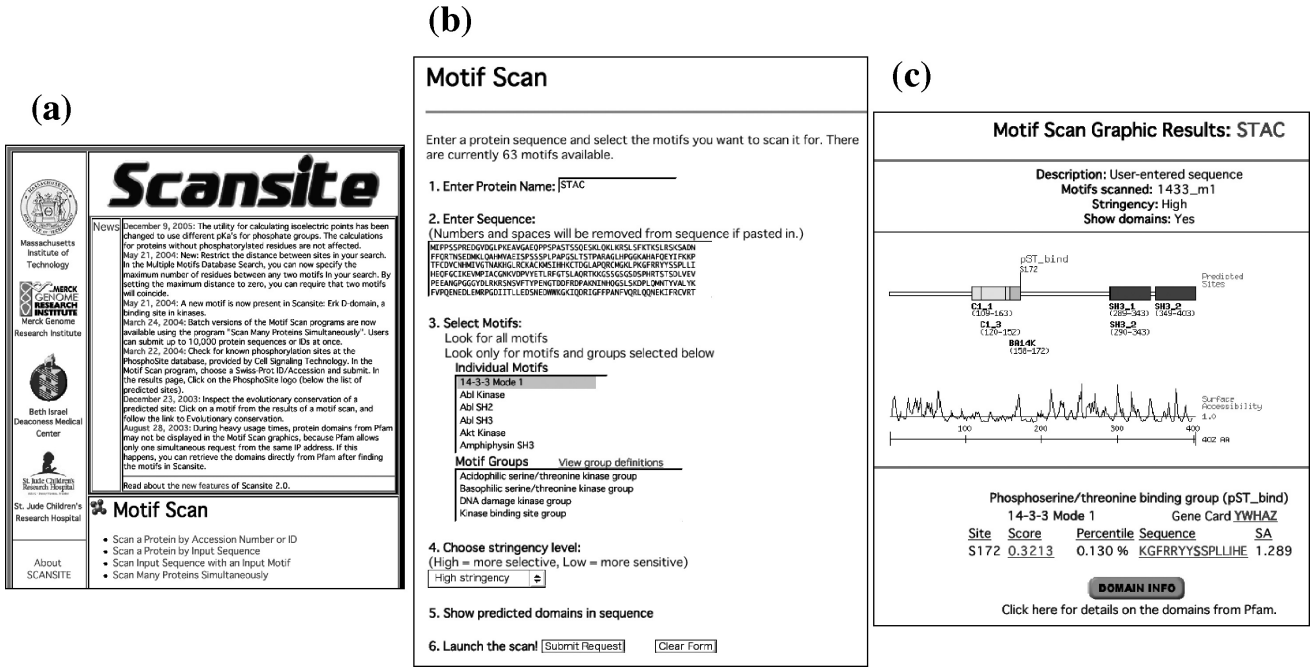


FIGURE 13.3 The Scansite Motif Scanner. The 14-3-3-binding consensus motif mode I (RSXpSXP) located in target proteins was surveyed by the Scansite Motif Scanner which assesses the probability of a site matching the candidate motif under high, medium or low stringent conditions. (a) The opening menu (cited from the website of http://scansite.mi.edu). (b) The motif scan menu. After entering the protein sequence of STAC, 63 different motifs including 14-3-3 mode I could be processed for searching. (c) The graphic view of the results. The candidate for the 14-3-3-binding motif of STAC under the high stringent condition (pST_bind S172) is indicated on the predicted domain structure of the protein (upper panel), accompanied by the relative scores for the interaction of pS172 domain: KGFRRYYSSPLLIHE with 14-3-3 protein (lower panel). A plot of the surface accessibility suggests the residues located near the protein surface with a capacity to interact with target proteins (upper panel).

3. *Methionine aminopeptidase 2 (METAP2; NM_006838).* This is a 67-kDa protein (MW 52,894; cytoplasmic) that interacts with eukaryotic initiation factor-2 (eIF-2) and regulates protein synthesis [62]. It contains two mode I motifs with low stringency; pT113 (KRGPKVQTDPPSVPI) and pS152 (TAAWRTTSEEKKALD).
4. *Melanoma antigen family B, 4 (MAGEB4; NM_002367).* This is a member of the MAGEB family (MW 38,925; nuclear) expressed abundantly in testis whose function remains unknown.[63] It contains three mode I motifs; T18 (AREKRQRTRGQTQDL) with medium stringency, and pT194 (GNQSSAWTLPRNGLL) and pS339 (SAYSRATSSSSSQPM) with low stringency.
5. *Chondroitin 4 sulfotransferase 11 (CHST11; NM_018413).* This is a member of the HNK1 sulfotransferase family GalNAc 4-*O*-sulfotransferase (MW 41,557; endoplasmic reticulum and mitochondria) that plays a role in chondroitin sulfate and dermatan sulfate biosynthesis.[64] It contains three mode I motifs; pS93 (TDTCRANSATSRKRR) with medium stringency, and pS56 (DICCRKGSRSPLQEL) and S194 (EPFERLVSAYRNKFT) with low stringency.
6. *Zinc finger, C3HC-type containing 1 (ZC3HC1; NM_016478).* This is a 60-kDa protein (MW 55,258; nuclear) that interacts with anaplastic lymphoma kinase (ALK) and plays an antiapoptotic role in nucleophosmin-ALK signaling event.[65] The 14-3-3-binding consensus motif mode I is not found.
7. *Minichromosome maintenance deficient 10 (MCM10; NM_018518).* This is a key component of the pre-replication complex (pre-RC) (MW 98,188; nuclear) essential for the initiation of DNA replication.[66] It contains five mode I motifs; pS90 (AQPPRTGSEFPRLEG) with medium stringency, and pS35 (KPAIKSISASALLKQ) S55 (LEMRRRKSEEIQKRF), pS302 (PCGNRSISLDRLPNK), and T329 (DGMLKEKTGPKIGGE) with low stringency.
8. *DEAD box polypeptide 54 (DDX54; NM_024072)* (Figure 13.2c). This is a 97-kDa RNA helicase (DP97) (MW 98,601; nuclear) that interacts with estrogen receptor (ER) and represses the transcription of ER-regulated genes.[67] It contains two mode I motifs with low stringency; pT95 (EDKKKIKTESGRYIS) and pS102 (TESGRYISSSYKRDL).
9. *Heterogeneous nuclear ribonucleoprotein C (HNPRC; NM_004500).* This is a member of heterogeneous nuclear ribonucleoproteins (hnRNPs) (MW 33,291; nuclear) involved in pre-mRNA processing, mRNA metabolism and transport.[68] It contains four mode I motifs; pS125 (DYYDRMYSYPARVPP) with high stringency, and pS158 (NTSRRGKSGFNSKSG), pS170 (KSGQRGSSKSGKLKG), and pS240 (ETNVKMESEGGADDS) with low stringency.
10. *Fibroblast growth factor 12 (FGF12; NM_004113).* This is a member of the FGF family (MW 27,401; nuclear) that plays a role in nervous system development and function.[69] It contains two mode I motifs with low

stringency; pS150 (VCMYREQSLHEIGEK) and pS165 (QGRSRKSSGTPTMNG).

11. *Glutathione S-transferase M3 (GSTM3; BC030253).* This is a cytoplasmic glutathione *S*-transferase of the mu class (MW 26,561; cytoplasmic) that plays a role in detoxification of carcinogens, therapeutic drugs, environmental toxins, and products of oxidative stress.[70] It contains one mode I motif with low stringency; pS64 (GIKLRSFSV).
12. *Src homology three (SH3) and cysteine rich domain (STAC; NM_003149)* (Figure 13.2d). This is a 47-kDa protein containing a SH3 domain and a cysteine-rich domain (MW 44,556; nuclear) that plays a role in the neuron-specific signal transduction pathway.[71] It contains seven mode I motifs; pS172 (KGFRRYYSSPLLIHE) with high stringency (Figure 13.3c), pS56 (TKSLRSKSADNFFQR) and pS255 (DLRKRSNSVFTYPEN) with medium stringency, and pS46 (QKLKRSLSFKTKSLR), pS51 (SLSFKTKSLRSKSAD), pS66 (NFFQRTNSEDMKLQA), and pS253 (GYDLRKRSNSVFTYP) with low stringency.
13. *ATPase, H^+ transporting, lysosomal, 21 kD, V0 subunit C" (ATP6V0B; NM_004047).* This is a 23-kDa component of vacuolar ATPase (MW 21,408; endoplasmic reticulum) that mediates acidification of intracellular organelles.[72] The 14-3-3-binding consensus motif mode I is not found.

Immunoprecipitation Analysis Validated the Specific Binding to 14-3-3

EAP30, DDX54, and STAC were selected to verify the results of microarray analysis in view of higher Z-Score values.[56] The recombinant proteins were expressed in HEK293 cells that constitutively express a substantial amount of endogenous 14-3-3 protein. The cells were homogenized in the lysis buffer either with inclusion of phosphatase inhibitors or with inclusion of recombinant protein phosphatase-1 (PP1) instead of phosphatase inhibitors. Total cell lysate was processed for immunoprecipitation (IP) with rabbit anti-14-3-3 protein antibody (K-19) or with normal rabbit IgG. K19 coimmunoprecipitated 14-3-3 and STAC from the lysate of HEK293 cells that express the recombinant STAC protein, whereas normal rabbit IgG did not pull down these proteins (Figure 13.1). K-19 immunoprecipitated EAP30 and DDX54 from the lysate of HEK293 cells that express the recombinant EAP30 or DDX54 protein, respectively (Figure 13.1). These results indicate that EAP30, DDX54 and STAC interact with the endogenous 14-3-3 protein in HEK293 cells where the corresponding recombinant proteins were expressed.

STAC has the highly stringent 14-3-3-binding consensus motif RYYSSP in amino acid residues 169 to 174 (pS172) by the Scansite Motif Scanner search (Figure 13.3). Therefore, a possible involvement of this motif in binding to 14-3-3 was further investigated by IP analysis of a panel of mutant and truncated STAC proteins. K-19 immunoprecipitated the full-length wild-type (WT) STAC consisting of amino acid residues 2 to 402 (Figure 13.1). K-19 also pulled down the S172A mutant (SMT), and the S172A and S173A double mutant (DMT) from the lysate of HEK293 cells that express the corresponding recombinant proteins.[56] K-19 immunoprecipitated

the N-terminal half (NTF; amino acid residues 2 to 233) but not the C-terminal half (CTF; amino acid residues 234 to 402) of STAC (Figure 13.1). These observations indicate that the RYYSSP motif is not involved in binding of STAC to 14-3-3. This was confirmed by the observations that K-19 immunoprecipitated the truncated STAC protein lacking the RYYSSP motif (TRA; amino acid residues 2 to 164)[56] and the shortest form lacking both the RYYSSP sequence and the cysteine-rich domain (CRD) (TRB; amino acid residues 2 to 105) (Figure 13.1). Finally, K-19 pulled down the full-length WT STAC, EAP30, and DDX54 under the dephosphorylated condition (PP1) (Figure 13.1). These observations indicate that the 14-3-3-interacting domain is located in the N-terminal segment spanning amino acid residues 2 to 105 of STAC. The interaction of 14-3-3 with STAC, EAP30, and DDX54 is independent of serine/threonine-phosphorylation of the binding domains.

DISCUSSION

Protein Microarray Analysis Effectively Identifies 14-3-3-Binding Proteins

Protein microarrays provide a valuable tool for global proteome analysis with a wide range of applications, particularly to identification and characterization of protein function and molecular pathways closely associated with disease markers and therapeutic targets.[39–43] The great advantage of this technology exists in low reagent and sample consumption, rapid interpretation of the results, and the ability to easily manipulate experimental conditions.

The present study was designed to identify 14-3-3-binding proteins by using a high-density human protein microarray. The array contains 1752 proteins derived from multiple gene families of biological importance, including cell-signaling proteins, kinases, membrane-associated proteins, and metabolic proteins. The entire procedure could be accomplished within five hours after we obtain a specific probe. By probing with V5-tagged 14-3-3ε, we identified twenty 14-3-3 interactors, most of which were previously unreported except for glutathione *S*-transferase M3 (GSTM3) that was reported previously.[36] Unexpectedly, the highly stringent 14-3-3-binding consensus motifs (STAC and HNPRC) were identified only in two by the Scansite Motif Scanner search. The specific binding to 14-3-3 of EAP30, DDX54 and STAC was validated by coimmunoprecipitation analysis of the recombinant proteins expressed in HEK293 cells. These results indicate that protein microarray is an effective tool for the rapid and systematic identification of protein–protein interactions, including those not predicted by the Database searching.

Potential Problems Remain to Be Solved in the Present Study

In general, protein microarray has its own limitations associated with the efficient expression and purification of native target proteins.[40,41] The target proteins spotted on the microarray we utilized were expressed by a baculovirus expression system and purified under non-denaturating conditions to maximize the preservation of native folding, posttranslational modifications, and proper functionality. In contrast,

bacterially expressed proteins lack glycosylation and phosphorylation moieties, and are often misfolded during purification. Post-translational modifications play a pivotal role in a range of protein–protein interactions. Immuno-labeling with anti-phosphotyrosine (pTyr) antibody showed that approximately 10 to 20% of the proteins on the array are phosphorylated (Invitrogen, unpublished data). When it was utilized for kinase substrate identification, most of known kinases immobilized on the array are enzymatically active with the capacity of autophosphorylation, suggesting that they are certainly phosphorylated on tyrosine residues, probably on serine and threonine residues (Invitrogen application note). However, we could not currently validate the precise level of serine and threonine phosphorylation of individual target proteins due to a lack of anti-phosphoserine (pSer) and anti-phosphothreonine (pThr) antibodies suitable for detection on glass slides.

The protein microarray we utilized includes 11 known 14-3-3-binding proteins, such as PCTAIRE protein kinase 1 (PCTK1),[73] protein kinase C zeta (PRKCZ),[74] keratin 18 (KRT18),[75] myosin light polypeptide kinase (MYLK),[76] v-abl Abelson murine leukemia viral oncogene homolog 1 (ABL1),[77] v-akt murine thymoma viral oncogene homolog 1 (AKT1),[78] epidermal growth factor receptor (EGFR),[79] cell division cycle 2 (CDC2),[80] mitogen-activated protein kinase kinase kinase 1 (MAP3K1),[81] mitogen-activated protein kinase-activated protein kinase 2 (MAPKAPK2),[82] and stratifin (SFN).[37] However, none of these were identified as positive. Therefore, there exists the possibility that some 14-3-3 binding partners were not detected due to imperfect phosphorylation of target proteins, inaccessibility by a sterical hindrance of epitope tags,[83] or a 14-3-3 isoform-specific binding ability. Calmodulin, another known 14-3-3 interactor,[84] is included as a negative control on the array. It was found as negative in the present study, because the calcium-dependent interaction between 14-3-3 and calmodulin could not be detected under the calcium-free conditions we employed. Recently, by using two dimensional (2-D)-gel electrophoresis and mass spectrometry, we showed that vimentin, an intermediate filament protein, interacts with 14-3-3ε in cultured human astrocytes.[20] More recently, we found that heat shock protein Hsp60 and the cellular prion protein PrPC interact with 14-3-3ζ in human neurons in culture and brain tissues.[85] Unfortunately, the protein microarray we examined here includes neither vimentin, Hsp60 nor prion protein.

Recent evidence indicates that 14-3-3-binding phosphorylation sites do not exactly fit the consensus motif,[1,25,75] and an accessory site is required to enhance a stable 14-3-3-target interaction.[4,86] Furthermore, 14-3-3 interacts with a set of target proteins in a phosphorylation-independent manner.[26–29] We found that the interaction is independent of serine/threonine-phosphorylation of the binding sites of EAP30, DDX54 and STAC, supporting this possibility.

Biological Roles of 14-3-3-Interacting Proteins

Among the 14-3-3 interactors we identified, several proteins are categorized as a component of multimolecular complexes involved in transcriptional regulation. ELL is a human oncogene encoding a RNA polymerase II (Pol II) transcription factor that promotes transcription elongation. EAP30 is a component of the ELL complex where EAP30 mediates derepression of transcription by Pol II,[59] although the PSORT

II search suggests that its putative location is cytoplasmic. A recent study showed that EAP30 interacts with the tumor susceptibility gene TSG101 product, a cellular factor that mediates packaging of HIV virions.[87] DDX54 is a RNA helicase that interacts with estrogen receptor (ER) and represses the transcription of ER-regulated genes.[67] A chromatin immunoprecipitation (ChIP) assay showed that hepatocyte nuclear factor 4-alpha (HNF4α), a master regulator of hepatocyte gene expression, interacts with the DDX54 gene promoter, together with Pol II.[88] HNPRC belongs to a member of heterogeneous nuclear ribonucleoproteins (hnRNPs) involved in pre-mRNA processing, mRNA metabolism and transport.[68] Increasing evidence indicates that the 14-3-3 protein and its targets are widely distributed in nearly all subcellular compartments, including the nucleus.[3,35]

STAC has a cysteine-rich domain (CRD) of the protein kinase C family in the N-terminal half (NTF) and a src homology three (SH3) domain in the C-terminal half (CTF), suggesting its role as an adapter on which divergent signaling pathways converge.[71,89] STAC is expressed predominantly in the brain with the distribution in a defined population of neurons.[71] IP analysis of mutant and truncated forms of STAC argued against an active involvement of the most stringent motif RYYSSP (pS172) in its binding to 14-3-3, and indicated that the interacting motif is located in the N-terminal amino acid residues 2 to 105 without requirement of serine/threonine phosphorylation.

FUTURE DIRECTIONS

Protein microarrays are a powerful tool for the rapid and systematic identification of protein–protein and other biomolecule interactions. However, they are still under development in methodological aspects. The strict quality controls of analytical procedures,[90] validation of the results by different methods, and evaluation of enormous data by bioinformatic approaches are highly important. The applications of protein microarrays include characterization of antibody specificity and autoantibody repertoire, and identification of novel biomarkers and molecular targets associated with disease type, stage and progression, leading to establishment of personalized medicine.[44–51] Theoretically, this technology could determine all of the binding partners at once, consisting of "the whole interactome" in a subset of cells responding to specific treatment. It would open up a new avenue of drug discovery research. Development of an ultrahigh-density protein microarray containing all spliced variants of target proteins could facilitate achievement of this purpose. A cell-free transcription and translation-coupled system might provide an effective tool for producing ideal proteins.[83] At present, the most advanced version of human protein microarray contains approximately 8000 GST-tagged proteins, commercially available from Invitrogen (ProtoArray v4.0), accompanied by an upgraded version of the analytical software (ProtoArray Prospector). It seems highly efficient to screen a large number of protein–protein interactions in human cells, including those unrecognized by the conventional methods such as Y2H.[91,92] However, when faced with a huge amount of data, bioinformatic and statistical analyses become crucial (visit the useful website of Pathguide for a comprehensive pathway resource list; cbio.mskcc.org/prl). Recently, an ultrahigh sensitive detection method armed with

silicon-nanowire field-effect sensors has come into use with its application to protein microarray analysis.[93] This promising technology could detect the low-femtomolar range of interacting proteins, and greatly increase the detection sensitivity and specificity.

SUMMARY AND CONCLUSIONS

The 14-3-3 protein family consists of acidic 30-kDa proteins composed of seven isoforms in mammalian cells, expressed abundantly in neurons and glial cells of the CNS. The 14-3-3 isoforms form a dimer that acts as a molecular adaptor interacting with key signaling components involved in cell proliferation, transformation, and apoptosis. Until present, more than 300 proteins have been identified as 14-3-3-binding partners, although most of previous studies focused on a limited range of 14-3-3-interacting proteins. In this chapter we describe a comprehensive profile of 14-3-3-binding proteins by analyzing a high-density protein microarray (1752 proteins; ProtoArray v1.0) using recombinant human 14-3-3ε protein as a probe. We identified twenty 14-3-3 interactors, most of which were previously unreported 14-3-3-binding partners, although eleven known 14-3-3-binding proteins on the array, including KRT18 and MAPKAPK2, were undetected. The assay required less than five hours. Unexpectedly, highly stringent 14-3-3-binding consensus motifs, such as STAC and HNPRC, were identified only in two proteins by the Scansite Motif Scanner search. The specific binding to 14-3-3 of EAP30, DDX54 and STAC was verified by coimmunoprecipitation analysis of the recombinant proteins expressed in HEK293 cells. These results suggest that protein microarray is a valuable tool for rapid and comprehensive profiling of 14-3-3-binding proteins.

ACKNOWLEDGMENTS

This work was supported by grants from Research on Psychiatric and Neurological Diseases and Mental Health, the Ministry of Health, Labour and Welfare of Japan (H17-020), Research on Health Sciences Focusing on Drug Innovation, the Japan Health Sciences Foundation (KH21101), and the Grant-in-Aid for Scientific Research, the Ministry of Education, Science, Sports and Culture (B2-15390280 and PA007-16017320).

REFERENCES

1. Fu, H., Subramanian, R.R., and Masters, S.C., 14-3-3 proteins: Structure, function, and regulation, *Annu. Rev. Pharmacol. Toxicol.*, 40, 617, 2000.
2. van Hemert, M.J., Steensma, H.Y., and van Heusden, G.P.H., 14-3-3 proteins: Key regulators of cell division, signaling and apoptosis, *Bioessays*, 23, 936, 2001.
3. Dougherty, M.K. and Morrison, D.K., Unlocking the code of 14-3-3, *J. Cell Sci.*, 117, 1875, 2004.
4. MacKintosh, C., Dynamic interactions between 14-3-3 proteins and phosphoproteins regulate diverse cellular processes, *Biochem. J.*, 381, 329, 2004.

5. Boston, P.F., Jackson, P., and Thompson, R.J., Human 14-3-3 protein: Radioimmunoassay, tissue distribution, and cerebrospinal fluid levels in patients with neurological disorders, *J. Neurochem.*, 38, 1475, 1982.
6. Berg, D., Holzmann, C., and Riess, O., 14-3-3 proteins in the nervous system, *Nature Rev. Neurosci.*, 4, 752, 2002.
7. Hsich, G. et al., The 14-3-3 brain protein in cerebrospinal fluid as a marker for transmissible spongiform encephalopathies, *N. Engl. J. Med.*, 335, 924, 1996.
8. Zerr, I. et al., Detection of 14-3-3 protein in the cerebrospinal fluid supports the diagnosis of Creutzfeldt-Jakob disease, *Ann. Neurol.*, 43, 32, 1998.
9. Richard, M. et al., Immunohistochemical localization of 14.3.3 ζ protein in amyloid plaques in human spongiform encephalopathies, *Acta Neuropathol.,* 105, 296, 2003.
10. Layfield, R. et al., Neurofibrillary tangles of Alzheimer's disease brains contain 14-3-3 proteins, *Neurosci. Lett.*, 209, 57, 1996.
11. Umahara, T. et al., Immunolocalization of 14-3-3 isoforms in brains with Pick body disease, *Neurosci. Lett.*, 371, 215, 2004.
12. Kawamoto, Y. et al., 14-3-3 proteins in Lewy bodies in Parkinson disease and diffuse Lewy body disease brains, *J. Neuropathol. Exp. Neurol.*, 61, 245, 2002.
13. Berg, D., Riess, O., and Bornemann, A., Specification of 14-3-3 proteins in Lewy bodies, *Ann. Neurol.*, 54, 135, 2003.
14. Kawamoto, Y. et al., Accumulation of 14-3-3 proteins in glial cytoplasmic inclusions in multiple system atrophy, *Ann. Neurol.*, 52, 722, 2002.
15. Komori, T. et al., Immunoexpression of 14-3-3 proteins in glial cytoplasmic inclusions of multiple system atrophy, *Acta Neuropathol.*, 106, 66, 2003.
16. Chen, H.-K. et al., Interaction of Akt-phosphorylated ataxin-1 with 14-3-3 mediates neurodegeneration in spinocerebellar ataxia type 1, *Cell*, 113, 457, 2003.
17. Malaspina, A., Kaushik, N., and de Belleroche, J., A 14-3-3 mRNA is up-regulated in amyotrophic lateral sclerosis spinal cord, *J. Neurochem.*, 75, 2511, 2000.
18. Toyo-oka, K. et al., 14-3-3ε is important for neuronal migration by binding to NUDEL: A molecular explanation for Miller-Dieker syndrome, *Nature Genet.*, 34, 274, 2003.
19. Satoh, J. et al., Detection of the 14-3-3 protein in the cerebrospinal fluid of Japanese multiple sclerosis patients presenting with severe myelitis, *J. Neurol. Sci.*, 212, 11, 2003.
20. Satoh, J., Yamamura, T., and Arima, K., The 14-3-3 protein ε isoform expressed in reactive astrocytes in demyelinating lesions of multiple sclerosis binds to vimentin and glial fibrillary acidic protein in cultured human astrocytes, *Am. J. Pathol.*, 165, 577, 2004.
21. Satoh, J. et al., The 14-3-3 protein detectable in the cerebrospinal fluid of patients with prion-unrelated neurological diseases is expressed constitutively in neurons and glial cells in culture, *Eur. Neurol.*, 41, 216, 1999.
22. Fujii, K. et al., Detection of 14-3-3 protein in the cerebrospinal fluid in mitochondrial encephalopathy with lactic acidosis and stroke-like episodes, *J. Neurol. Sci.*, 239, 115, 2005.
23. Tzivion, G. and Avruch, J., 14-3-3 proteins: Active cofactors in cellular regulation by serine/threonine phosphorylation, *J. Biol. Chem.,* 277, 3061, 2002.
24. Ganguly, S. et al., Melatonin synthesis: 14-3-3-dependent activation and inhibition of arylalkylamine *N*-acetyltransferase mediated by phosphoserine-205, *Proc. Natl. Acad. Sci. U.S.A.*, 102, 1222, 2005.
25. Aitken, A. et al., 14-3-3 proteins in cell regulation, *Biochem. Soc. Trans.*, 30, 351, 2002.

26. Zhai, J. et al., Identification of a novel interaction of 14-3-3 with p190RhoGEF, *J. Biol. Chem.*, 276, 41318, 2001.
27. Henriksson, M.L. et al., A nonphosphorylated 14-3-3 binding motif on exoenzyme S that is functional *in vivo*, *Eur. J. Biochem.*, 269, 4921, 2002.
28. Dai, J.-G. and Murakami, K., Constitutively and autonomously active protein kinase C associated with 14-3-3 ζ in the rodent brain, *J. Neurochem.*, 84, 23, 2003.
29. Fuglsang, A.T. et al., The binding site for regulatory 14-3-3 protein in plant plasma membrane H^+-ATPase. Involvement of a region promoting phosphorylation-independent interaction in addition to the phosphorylation-dependent C-terminal end, *J. Biol. Chem.*, 278, 42266, 2003.
30. Uetz, P. et al., A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*, *Nature*, 403, 623, 2000.
31. Ito, T. et al., A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc. Natl. Acad. Sci. U.S.A.*, 98, 4569–4574, 2001.
32. Vidalain, P.O. et al., Increasing specificity in high-throughput yeast two-hybrid experiments, *Methods*, 32, 363, 2004.
33. Zhang, L.V. et al., Predicting co-complexed protein pairs using genomic and proteomic data integration, *BMC Bioinformatics*, 5, 38, 2004.
34. Jin, J. et al., Proteomic, functional, and domain-based analysis of *in vivo* 14-3-3 binding proteins involved in cytoskeletal regulation and cellular organization, *Curr. Biol.*, 14, 1436, 2004.
35. Meek, S.E.M., Lane, W.S., and Piwnica-Worms, H., Comprehensive proteomic analysis of interphase and mitotic 14-3-3-binding proteins, *J. Biol. Chem.*, 279, 32046, 2004.
36. Pozuelo Rubio, M. et al., 14-3-3-affinity purification of over 200 human phosphoproteins reveals new links to regulation of cellular metabolism, proliferation and trafficking, *Biochem. J.*, 379, 395, 2004.
37. Benzinger, A. et al., Targeted proteomic analysis of 14-3-3 sigma, a p53 effector commonly silenced in cancer, *Mol. Cell. Proteomics*, 4, 785, 2005.
38. von Mering, C. et al., Comparative assessment of large-scale data sets of protein-protein interactions, *Nature*, 417, 399, 2002.
39. MacBeath, G., Protein microarrays and proteomics, *Nature Genet.*, 32 Suppl., 526, 2002.
40. Schweitzer, B., Predki, P., and Snyder M., Microarrays to characterize protein interactions on a whole-proteome scale, *Proteomics*, 3, 2190, 2003.
41. Bertone, P. and Snyder, M., Advances in functional protein microarray technology, *FEBS J.*, 272, 5400, 2005.
42. Mattoon, D. et al., Biomarker discovery using protein microarray technology platforms: Antibody-antigen complex profiling, *Expert. Rev. Proteomics*, 2, 879, 2005.
43. Zanger, R.C., Varnum, S.M., and Bollinger, N., Studying cellular processes and detecting disease with protein microarrays, *Drug Metab. Rev.*, 37, 487, 2005.
44. MacBeath, G. and Schreiber, S.L., Printing proteins as microarrays for high-throughput function determination, *Science*, 289, 1760, 2000.
45. Zhu, H. et al., Global analysis of protein activities using proteome chips, *Science*, 293, 2101, 2001.
46. Robinson, W.H. et al., Autoantigen microarrays for multiplex characterization of autoantibody responses, *Nature Med.*, 8, 295, 2002.
47. Michaud, G.A. et al., Analyzing antibody specificity with whole proteome microarrays, *Nature Biotechnol.*, 21, 1509, 2003.
48. Newman, J.R.S. and Keating A.E., Comprehensive identification of human bZIP interactions with coiled-coil arrays, *Science*, 300, 2097, 2003.

49. Robinson, W.H. et al., Protein microarrays guide tolerizing DNA vaccine treatment of autoimmune encephalomyelitis, *Nature Biotechnol.*, 21, 1033, 2003.
50. Chan, S.M. et al., Protein microarrays for multiplex analysis of signal transduction pathways, *Nature Med.*, 10, 1390, 2004.
51. Quintana, F.J. et al., Functional immunomics: Microarray analysis of IgG autoantibody repertoires predicts the future response of mice to induced diabetes, *Proc. Natl. Acad. Sci. U.S.A.*, 101, 14615, 2004.
52. Espejo, A. et al., A protein-domain microarray identifies novel protein-protein interactions, *Biochem. J.*, 367, 697, 2002.
53. Jones, R.B. et al., A quantitative protein interaction network for the ErbB receptors using protein microarrays, *Nature*, advance online publication, Nov. 2005.
54. Satoh, J. and Yamamura, T., Gene expression profile following stable expression of the cellular prion protein, *Cell. Mol. Neurobiol.*, 24, 793, 2004.
55. Satoh, J. and Kuroda, Y., Differential gene expression between human neurons and neuronal progenitor cells in culture: An analysis of arrayed cDNA clones in NTera2 human embryonal carcinoma cell line as a model system, *J. Neurosci. Methods*, 94, 155, 2000.
56. Satoh, J., Nanri, Y., and Yamamura, T., Rapid identification of 14-3-3-binding proteins by protein microarray analysis, *J. Neurosci. Methods*, 152, 278, 2005.
57. Ichimura, T. et al., 14-3-3 proteins modulate the expression of epithelial Na^+ channels by phosphorylation-dependent interaction with Nedd4-2 ubiquitin ligase, *J. Biol. Chem.*, 280, 13187, 2005.
58. Obenauer, J.C., Cantley, L.C., and Yaffe, M.B., Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs, *Nucleic Acids Res.*, 31, 3635, 2003.
59. Schmidt, A.E. et al., Cloning and characterization of the EAP30 subunit of the ELL complex that confers derepression of transcription by RNA polymerase II, *J. Biol. Chem.*, 274, 21981, 1999.
60. Hierro, A. et al., Structure of the ESCRT-II endosomal trafficking complex, *Nature*, 431, 221, 2004.
61. Motto, D.G. et al., Implication of the GRB2-associated phosphoprotein SLP-76 in T cell receptor-mediated interleukin 2 production, *J. Exp. Med.*, 183, 1937, 1996.
62. Wu, S. et al., Cloning and characterization of complementary DNA encoding the eukaryotic initiation factor 2-associated 67-kDa protein (p^{67}), *J. Biol. Chem.*, 268, 10796, 1993.
63. Lurquin, C. et al., Two members of the human *MAGEB* gene family located in Xp21.3 are expressed in tumors of various histological origins, *Genomics*, 46, 397, 1997.
64. Mikami, T. et al., Specificities of three distinct human chondroitin/dermatan *N* acetylgalactosamine 4-*O*-sulfotransferases demonstrated using partially desulfated dermatan sulfate as an acceptor. Implication of differential roles in dermatan sulfate biosynthesis, *J. Biol. Chem.*, 278, 36115, 2003.
65. Ouyang, T. et al., Identification and characterization of a nuclear interacting partner of anaplastic lymphoma kinase (NIPA), *J. Biol. Chem.*, 278, 30028, 2003.
66. Yoshida, K. and Inoue, I., Expression of MCM10 and TopBP1 is regulated by cell proliferation and UV irradiation via the E2F transcription factor, *Oncogene*, 23, 6250, 2004.
67. Rajendran, R.R. et al., Regulation of nuclear receptor transcriptional activity by a novel DEAD box RNA helicase (DP97), *J. Biol. Chem.*, 278, 4628, 2003.

68. Nakagawa, T.Y. et al., Molecular cloning of cDNA for the nuclear ribonucleoprotein particle C proteins: A conserved gene family, *Proc. Natl. Acad. Sci. U.S.A.*, 83, 2007, 1986.
69. Smallwood, P.M. et al., Fibroblast growth factor (FGF) homologous factors: New members of the FGF family implicated in nervous system development, *Proc. Natl. Acad. Sci. U.S.A.*, 93, 9850, 1996.
70. Campbell, E. et al., A distinct human testis and brain μ-class glutathione *S*-transferase. Molecular cloning and characterization of a form present even in individuals lacking hepatic type μ isoenzymes, *J. Biol. Chem.*, 265, 9188, 1990.
71. Suzuki, H. et al., Stac, a novel neuron-specific protein with cysteine-rich and SH3 domains, *Biochem. Biophys. Res. Commun.*, 229, 902, 1996.
72. Oka, T., Yamamoto, R., and Futai, M., Three *vha* genes encode proteolipids of *Caenorhabditis elegans* vacuolar-type ATPase. Gene structures and preferential expression in an H-shaped excretory cell and rectal cells, *J. Biol. Chem.*, 272, 24387, 1997.
73. Graeser, R. et al., Regulation of the CDK-related protein kinase PCTAIRE-1 and its possible role in neurite outgrowth in Neuro-2A cells, *J. Cell Sci.*, 115, 3479, 2002.
74. van der Hoeven, P.C.J. et al., Protein kinase C activation by acidic proteins including 14-3-3, *Biochem. J.*, 347, 781, 2000.
75. Ku, N.-O., Liao, J., and Omary, M.B., Phosphorylation of human keratin 18 serine 33 regulates binding to 14-3-3 proteins, *EMBO J.*, 17, 1892, 1998.
76. Haydon, C.E. et al., Identification of a phosphorylation site on skeletal muscle myosin light chain kinase that becomes phosphorylated during muscle contraction, *Arch. Biochem. Biophys.*, 397, 224, 2002.
77. Yoshida, K. et al., JNK phosphorylation of 14-3-3 proteins regulates nuclear targeting of c-Abl in the apoptotic response to DNA damage, *Nature Cell Biol.*, 7, 278, 2005.
78. Powell, D.W. et al., Identification of 14-3-3ζ as a protein kinase B/Akt substrate, *J. Biol. Chem.*, 277, 21639, 2002.
79. Oksvold, M.P., Huitfeldt, H.S., and Langdon, W.Y., Identification of 14-3-3ζ as an EGF receptor interacting protein, *FEBS Lett.*, 569, 207, 2004.
80. Chan, T.A. et al., 14-3-3σ is required to prevent mitotic catastrophe after DNA damage, *Nature*, 401, 616, 1999.
81. Fanger, G.R. et al., 14-3-3 proteins interact with specific MEK kinases, *J. Biol. Chem.*, 273, 3476, 1998.
82. Powell, D.W. et al., Proteomic identification of 14-3-3ζ as a mitogen-activated protein kinase-activated protein kinase 2 substrate: Role in dimmer formation and ligand binding, *Mol. Cell. Biol.*, 23, 5376, 2003.
83. Ramachandran, N. et al., Self-assembling protein microarrays, *Science*, 305, 86, 2004.
84. Luk, S.C.W. et al., *In vivo* and *in vitro* association of 14-3-3 epsilon isoform with calmodulin: Implication for signal transduction and cell proliferation, *J. Cell. Biochem.*, 73, 31, 1999.
85. Satoh, J. et al., The 14-3-3 protein forms a molecular complex with heat shock protein Hsp60 and cellular prion protein, *J. Neuropathol. Exp. Neurol.*, 64, 858, 2005.
86. Yaffe, M.B., How do 14-3-3 proteins work? Gatekeeper phosphorylation and the molecular anvil hypothesis, *FEBS Lett.*, 513, 53, 2002.
87. von Schwedler, U.K. et al., The protein network of HIV budding, *Cell*, 114, 701, 2003.
88. Odom, D.T. et al., Control of pancreas and liver gene expression by HNF transcription factors, *Science*, 303, 1378, 2004.
89. Hardy, K. et al., Transcriptional networks and cellular senescence in human mammary fibroblasts, *Mol. Biol. Cell*, 16, 943, 2005.

90. Kricka, L.J. and Master SR., Validation and quality control of protein microarray-based analytical methods, *Methods Mol. Med.*, 114, 233, 2005.
91. Stelzl, U. et al., A human protein-protein interaction network: A resource for annotating the proteome, *Cell*, 122, 957, 2005.
92. Rual, J.F. et al., Towards a proteome-scale map of the human protein-protein interaction network, *Nature*, 437, 1173, 2005.
93. Zheng, G. et al., Multiplexed electrical detection of cancer markers with nanowire sensor arrays, *Nature Biotechnol.*, 23, 1294, 2005.

14 A Combined Force of Chemical Genetics and Protein Microarrays

Heng Zhu and Jing Huang

CONTENTS

INTRODUCTION

Proteome Microarray Technology

To date, the genomic sequences of over 200 organisms have been determined (http://www.ncib.nlm.nih.gov). The science of genomics has revolutionized both basic research and drug discovery. DNA microarray technology, in particular, has become a routine tool to profile the expression of thousands of genes and even the entire gene repertoire of an organism. Similar approaches have been applied to identify genes that are differentially expressed in response to drug treatments, which in turn facilitates genomics-based drug discovery and disease classification. However, a major drawback to such approaches is that differences in gene expression profiles usually do not provide direct links to the causative elements (drugs) and, in some cases, may not be related to them at all. Further, cellular functions are mostly executed by genes that encode proteins, whose activities are often controlled, modified and regulated by other proteins. Thus, determining the biochemical activities of each protein, how they might assemble together to carry out the biochemical

reactions and cellular events, and also how they may function in a sequential pathway or collaborative network are crucial to elucidating the molecular basis of complex processes.

Protein microarray technology may have the greatest potential for providing direct information on protein functions and drug targets. It has been shown as a flexible platform to analyze the biochemical activities of proteins.[1,2] Protein chips are miniature grids that contain small amounts of purified proteins in a high-density format.[2] Based on their applications, protein microarrays can be categorized into two classes: functional and analytical protein microarrays.[3–7] Analytical protein arrays can be used for monitoring protein expression levels, for protein profiling and for clinical diagnostics. Functional protein microarrays can be screened in a high-throughput fashion for biochemical activities, protein–protein, protein–DNA, protein–RNA, and protein–ligand interactions.[2,8–10] One of our major contributions in the field was the fabrication of a high-density protein chip containing >5800 purified yeast proteins (>90% of the yeast proteome).[2] First, we demonstrated that such proteome chips could be applied to identify protein–protein and protein–lipid interactions. Later, we developed additional assays to identify DNA- and small molecule- binding proteins.[9,10] Most recently, we accomplished a large-scale screening for *in vitro* kinase substrates and identified >4192 kinase-substrate interactions.[11]

This approach can be further extended in several different ways. Binding can be studied in real-time by use of a surface plasmon resonance (SPR) biosensor surface with 64 individual immobilized sites in a single flow cell, which can be scaled to 400 assays per day (Biacore).[12] Peptides can also be analyzed using microarrays. Recently, a monolayer-coated gold chip was shown to be useful for immobilization of peptides for biochemical analysis using detection by a phosphorimager, SPR, and fluorescence microscopy.[13,14] Synthesis of peptide microarrays may become more practical with the development of methods for *in situ* synthesis of high-density peptide microarrays, using photolithography or light-directed synthesis. Recently, LaBaer's group has reported using an *in situ* approach to fabricate protein microarrays.[15] Carbohydrate and small-molecule microarrays have also shown great potential for characterizing protein-small molecule binding activities.[16–20]

Protein Microarray Fabrication and Assay Development

A major challenge of constructing a proteome-wide protein microarray is to convert genomic sequence information into thousands of pure, functional proteins that are immobilized on a solid surface. The basic technologies necessary for this process, such as recombination-based cDNA cloning, affinity tag-based protein expression and purification, and even the array technology, have been developed for a number of years. However, only recently have we seen rapid progress in the integration of these technologies in a high-throughput format for proteome microarray fabrication.

Snyder and colleagues were among the first to develop functional proteome microarrays.[2] Using budding yeast as a model, we have developed high-throughput protein purification protocols to purify >5800 yeast proteins from yeast cells. Because of biochemical diversity of proteins, it has been skeptical as how to develop a technology that can efficiently immobilize proteins while keeping them active

across an entire proteome. However, biochemists have been effectively cross-linking proteins to surface support for decades. To accomplish this kind of immobilization, we and others have tested a great variety of surface chemistries, including but not limited to surfaces that are either coated or grafted with polyvinylidene difluride,[21] agarose,[22] polyacrylamide gel pads,[3] nitrocellulose, polylysine,[3] aldehyde, epoxy, or a homofunctional cross-linker,[23] and see Zhu and Snyder[3] for review. All of the above approaches resulted in protein immobilization in a random fashion to the surfaces. In principle, attachment of proteins in an oriented fashion via a commonly shared affinity tag on each protein molecule is a more desirable strategy because immobilization through the spare handle generally does not interfere with the protein conformation and therefore, maintain the immobilized proteins in their native forms. We have achieved such affinity-based immobilization by printing N-terminal Hisx6 tagged yeast proteins on nickel-NTA coated glass slides, and the results showed that the signals were 10-times stronger relative to random attachment methods.[3] Similarly, Lesaicherre et al.[24] reported a new strategy for site-directed immobilization for fusion proteins onto streptavidin-coated glass slides. They produced fusion proteins containing an intein-tag with a chitin-binding domain, which are used to purify the fusion proteins and specifically add biotin to their C-terminal ends. Since no systematic and parallel studies have been reported to compare the experimental results carried out on different surfaces, it is not possible at this time to determine which, if any, of the above surfaces or surface chemistries are likely to be best suited for various types of assays.

Based on literature and our own experiences, the success of developing a new type of assay on protein microarrays requires careful pilot experiments to identify the optimal surface chemistry and reaction conditions. Surface chemistry, implementation of stringent quality controls, reaction and washing conditions, and detection methods play equally important roles. In theory, protein microarrays fabricated via affinity-based protein immobilization should generally be better than other types of surfaces or surface chemistries. However, this is not always true. For example, in the recently accomplished "phosphorylome" project,[11] we found that the surface chemistry could dramatically affect the outcome of the kinase assays: FullMoon slides provided us with signal-to-noise ratios that were far superior to Ni-NTA and other types of surfaces, such as aldehyde and epoxy grafted glass slides. Because of a lack of specific antibodies to detect phospho-serine or -threonine residues or an efficient fluorescence-based labeling method, radioisotope-labeled ^{33}P-ATP was the only choice for detection. As a charged small molecule, ^{33}P-ATP can nonspecifically bind to porous surfaces, such as nitrocellulose. We and others indeed observed much higher background in kinase assays on nitrocellulose-coated slides, although it has been considered as a better surface to conduct protein–protein interactions or serum profiling.[11,23,25] To further reduce the background and to remove signal from the binding of autophosphorylated kinase proteins to the surface, we tested various washing conditions covering a great range of stringency. We finally determined to apply 0.5% SDS in washing buffers to ensure the complete removal of any nonspecific signals. When background is too low, it becomes extremely difficult to align the spot-calling grid to the images for the identification of the positives. To solve this problem, we implemented a human kinase that is known to strongly autophosphorylate

and is stable during the process of protein microarray fabrication such that this particular protein was spotted at each corner of all 48-protein blocks on the slides. Therefore, after kinase reactions, they will be labeled and can serve as landmarks for grid alignment. This kind of kinome platform should be extremely useful for screening and evaluating potential small molecule kinase inhibitors. Selective kinase inhibitors are considered to be of great therapeutic importance, and up to this date their target specificity has been examined largely on an *ad hoc* basis.

Compared to higher eukaryotes, fabrication of yeast proteome arrays was relatively simple because most yeast open reading frames (ORFs) are not interrupted by introns. For humans and other higher eukaryotes, a successful ORF cloning is highly dependent on the availability of a full-length cDNA for the gene of interest. Nevertheless, ambitious efforts are currently underway to generate nonredundant and sequence-verified clone collections in higher eukaryotic organisms, ranging from the *C. elegans* ORFeome project to the various public efforts to generate human full-length cDNA collections, such as the UniGene set,[26] the Full-length Expression (FLEX) Gene repository,[27] the Integrated Molecular Analysis of Genomes and their Expression (I.M.A.G.E.) cDNA collection and the associated Mammalian Gene Collection (MGC).[28,29] Commercial efforts are also being undertaken by companies such as Invitrogen, Genecopoeia, and Origene.

Although complete whole human proteome microarrays are still far from realization, comprehensive subsets of the proteome also have great utility. Fabrication of gene family-specific, tissue-specific and disease-specific protein microarrays will certainly facilitate rapid characterization of protein function, disease pathways for drug and biomarker identification, validation, selection, etc.

Quantitative Analysis on Protein Microarrays

One of the most significant contributions of DNA microarray technology is its ability to measure quantitatively the expression levels of every single mRNA species in a complex mixture. Unlike the simple chemistry of hybridization between DNA-DNA and DNA-RNA molecules, interactions among proteins are much more complex and can vary dramatically. In the past few years, the reported applications of protein microarrays have been mostly qualitative rather than quantitative. However, this is about to change. In a recent paper, MacBeath and colleagues reported a quantitative measurement of protein-peptide interactions using protein microarrays.[30] First, a protein microarray containing virtually 102 Src homology 2 (SH2) and 41 phosphotyrosine binding (PTB) domains was constructed. Based on literature search, the authors identified 12, 6, and 11 sites that can be found on human proteins, EGFR, ErB2, and ErB3, respectively. They also included four predicted peptide sequences in their binding assays. After synthesizing 17–19 residue, phosphotyrosine-containing peptides, the authors probed the protein microarrays with eight concentrations of each peptide, ranging from 10 nM to 5 μM. The interactions have yielded binding curves for each peptide-protein pair, which can be used to calculate the equilibrium dissociation constants (off-rates) of each pair. Thus, the protein-peptide interactions can be visualized as a quantitative and systematic network. By varying the thresholds for composing such networks, the authors have observed surprising differences

between the receptors. For example, EGFR and ErbB2 became more promiscuous as the threshold was lowered, whereas ErB3 did not. When comparing the observed interactions with those predicted by an algorithm, Scansite 2.0, they found that the predicted interactions were biased: some matched closely to the microarray data for several domains, but much less for the others. Furthermore, many previously unknown interactions were uncovered in the low K_d ranges (< 2 μM). Overall, their studies have demonstrated that quantitative measurement can be achieved on protein microarrays and our knowledge of signaling pathways is still limited even with those that have been intensively studied in recent years.

APPLICATIONS OF PROTEIN MICROARRAYS IN DRUG DISCOVERY

Perhaps the most exciting prospect for the next generation of protein microarrays lies in the application of these assays to drug discovery, drug target identification, and clinical prognosis and diagnosis. Here we will discuss our exploratory work in this area through an example of chemical genetics discovery using proteome microarrays combined with other functional genomic, genetic, and molecular and cell biological tools (Figure 14.1).

Chemical Genetics

By combining the elements of chemistry and genetics, chemical genetics aims to elucidate biological mechanisms through small-molecule perturbation of gene/protein function in a conditional, specific, and systematic manner.[31] The attractiveness of this approach also lies in the fact that it directly explores a potential therapeutic solution.[32]

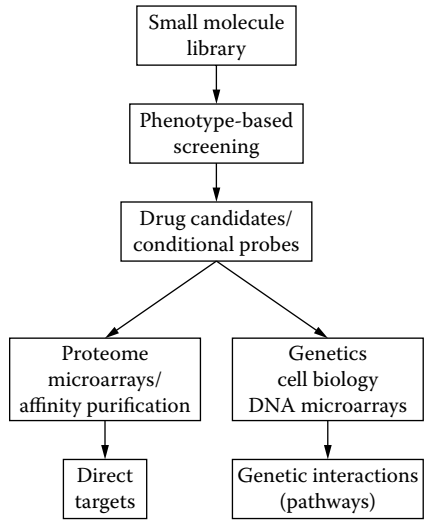


FIGURE 14.1 Flow chart of an integrated approach using chemical genetics and proteome microarrays.

A basic requirement in chemical genetics is that small molecule libraries need to be generated and screened rapidly and efficiently. Development of strategies to produce small molecules that can effectively intervene with complex biological systems is a big challenge by itself. Nature has provided numerous inspirations, especially in the areas of antibiotics and anticancer therapeutics.[33,34] Expanding the chemical toolbox beyond Nature has been of great interest to both chemists and biologists.[35–44] Given the focus of this chapter, we will not cover these areas and will refer interested readers to the original accounts above. We will also not discuss other areas of chemical genetics or chemical biology that aim to generate perfectly orthogonal scenarios which achieve absolute selectivity in targeting engineered biological systems for basic research, including signaling, target validation, and imaging studies,[45–47] except to point out that the elegance of these approaches highlights the challenge in obtaining extremely specific and selective probes for endogenous biological systems. On the other hand, we postulate that absolute selectivity is likely not required in every case for a drug (or cellular probe to some extent) to be useful, due to inherent properties of biological networks.

Classic studies by Barabasi and others indicate that protein networks are best modeled as scale-free networks, in which the majority of nodes have only a few neighbors while a small number of "hub" nodes have many.[48] Such scale-free networks have the double-edged property of being both vulnerable and robust at the same time.[49] The highly connected nodes are vulnerable whereas the low-degree nodes can be eliminated to a large extend without compromising the whole system (attack tolerance). In this light, a kinase inhibitor that targets a local high-degree protein kinase as its intended target and some (unintended) low-degree kinases may be comparable to an inhibitor that only inhibits the intended target. On the other hand, an inhibitor that inhibits a few high-degree kinases simultaneously might be too nonspecific or lethal altogether. We suggest that these principles be taken into consideration by the drug development community. As discussed below, proteome microarrays provide an ideal technological platform to evaluate both issues: node connectivity through analysis of protein–protein interactions (this can be complemented with yeast two-hybrid and other analyses)[50] and drug specificity through interrogating the drug molecule against the whole proteome as we have suggested.[10]

Small-Molecule Target Identification — A Bottleneck in the Practice of Chemical Genetics

In the past few years, phenotype-based chemical-genetic screens have been very successful in identifying conditional probes for dynamic cellular processes as well as potential leads to therapeutic drugs.[51–55] Subsequent target identification, however, poses a significant challenge and is currently the rate-limiting step in elucidating relevant biological pathways.[56,57] The problem is not necessarily less daunting when biased small molecule libraries are concerned. For example, substituted purine libraries, which are commonly used to discover kinase inhibitors, can potentially target non-kinase proteins as well.[58–61] Similar problems apply to small molecule hits that result from reverse chemical genetic screening, e.g., for direct binding to a target protein.[62] In most cases, testing of target specificity is performed on known

proteins within the pathway of interest, without interrogating the whole proteome in an unbiased fashion.

The recent launch of chemical genomics related initiatives by the NIH Roadmap (http://nihroadmap.nih.gov/) represents an organized effort towards providing an essential resource to the biomedical community. The next few years should witness an avalanche of small molecule probes identified from forward chemical genetic screens. Developing efficient target identification strategies is an urgent need.

Traditionally, small molecule targets are mainly identified through affinity chromatography.[63–70] This biochemical method relies mainly on the derivatization of radioisotope-labeled probes or affinity matrices for binding to cell lysates, followed by mass spectrometry to identify the bound protein(s). Besides being a relatively difficult and lengthy process, an inherent limitation to affinity chromatography is that it is biased against low-abundance proteins (many regulatory proteins are expressed at much lower levels than structural proteins), and biased towards high-abundance targets (which may not be the most biologically relevant targets). This issue is being improved by increased detection sensitivity of mass spectrometry techniques.

Genetics-based strategies, which may recover both direct targets and indirect targets (same or parallel pathways), can also be useful in drug target studies.[71–74] In addition, methods are developed that cleverly couple affinity with genetics, including cDNA display cloning,[75,76] the yeast three-hybrid system,[77] and a magnetic nanoprobe strategy.[78] Unbiased biochemical purification is also a powerful approach.[79,80] although fractionations are technically difficult and time-consuming to perform.

Proteome Microarrays as a Novel Target Identification Tool

As an alternative approach to purify and identify the small molecule target, we have developed a proteomics approach. Like affinity chromatography, small molecules were first labeled such that their physical presence and/or location can be followed. Common labels used for affinity chromatography can also be used to probe proteome chips, including affinity tags (e.g., biotin), fluorescence tags, photochemical tags, and radioisotopes.[81,82] The only requirement (as in affinity chromatography) is that the specific label can be incorporated into the molecule without abolishing its biological activity. A key advantage is that now the whole proteome can be probed with the biotin-labeled compounds in a microarray format. Because all the proteins immobilized on the chips are addressable, the drug-interacting proteins are readily identified using, for instance, Cy3-labeled streptavidin. This technology is thus also far more efficient and much easier than affinity chromatography as we can now identify targets within a few hours (instead of months or years).

Using this approach, we found a number of candidate proteins that could bind (with various affinities) to SMIR4, a small molecule inhibitor of rapamycin that we had identified from a yeast chemical genetics screen.[10] To determine whether these candidate proteins were indeed the *in vivo* targets of SMIR4, genetic and cell biology approaches were employed to validate the results. Eventually, we identified a yeast protein of previously unknown function (encoded by *YBR077*c) as a target for SMIR4 and a new component in TOR signaling.[10]

We will use SMIR4 as an example to illustrate the general principles and our thinking behind the important process of small molecule target identification. We expect a *bonafide* SMIR4 target *in vivo* to ideally satisfy two conditions upon its elimination from the cell. First, it may alter the cell's sensitivity to rapamycin (chemical effect equals genetic effect: it is important to note that the two effects are not necessarily equal, depending on the specificity of each). Second, it should reduce the ability of SMIR4 to suppress rapamycin (effect should disappear upon removal of target). We will address the second condition first because it is more straightforward; we found that in yeast cells deleted of *YBR077*c, SMIR4 can no longer suppress rapamycin's growth inhibitory effect at concentrations effective for wild type cells. To address the first condition, we tested the rapamycin sensitivity of yeast strains with deletions in each of the candidate proteins identified by the proteome chip. Of the 30 SMIR4 binders, only one exhibited an altered sensitivity to rapamycin: the Ybr077c deletion was hypersensitive to rapamycin. This suggests that SMIR4 likely causes a gain of function in Ybr077c by increasing rapamycin resistance of the cell. Consistently, forced expression of *YBR077*c confers rapamycin resistance in wild-type cells. Furthermore, a transcript profile of untreated *ybr077c* cells is strikingly similar to rapamycin-treated wild-type cells, indicating a requirement for YBR077c in TOR function. That Ybr077c functions in TOR signaling is consistent with its cellular localization. We know from database searching (http://yeastgfp.ucsf.edu) that both Ybr077c and the recently discovered TOR complex 1 (TORC1) component Kog1[83,84] are localized to the vacuolar membrane.[85] Other genetic and cell biological evidence also supports the idea that Ybr077c is a new component of the TOR signaling network (we named the protein Nir1, for new in rapamycin-sensitive signaling).

The role of Ybr077c in rapamycin-sensitive TOR signaling was also identified in an elegant study by Claudio De Virgilio's group.[86] In this study, Ybr077c (which they named Ego3) has been shown to function as a subunit of the EGO protein complex that is localized to the vacuolar membrane and regulates microautophagy, a process critical for recovery from rapamycin treatment (and presumably also from the natural starvation state). Although the exact mechanism of Ybr077c in microautophagy remains unknown, these studies have opened up a new avenue for regulating cellular sensitivity to rapamycin. Intriguingly, Ybr077c was detected to bind PI(3,4)P2 *in vitro*,[2] suggesting a possible involvement of PIs in regulating TOR pathway activity.[10] Consistent with this idea, a different line of investigation by Scott Emr's group[87] identified YBR077c (which they named SLM4) in a synthetic lethal screen with MSS4, which encodes a PI4P 5-kinase. Phospholipid involvement is likely a general theme in regulating TOR pathway activity.[88]

An added layer of complexity in using small molecules as probes is that the number and extent of proteins targeted by a small molecule may vary depending on the dosage of the small molecule. At low micromolar concentrations (identical to those required to confer rapamycin-resistant phenotype in wild-type cells), SMIR4 is unable to rescue ybr077c deleted cells (while it readily rescues other deletions that are more sensitive to rapamycin than ybr077c deletion), suggesting Ybr077c to be the *in vivo* target. As SMIR4 concentrations increase, however, the dependence on Ybr077c for rescue is reduced. That Ybr077c/Nir1/Slm4/Ego3 is unlikely the only target for SMIR4 is also suggested by DNA microarray data with at different SMIR4 concentrations.[10] Last but

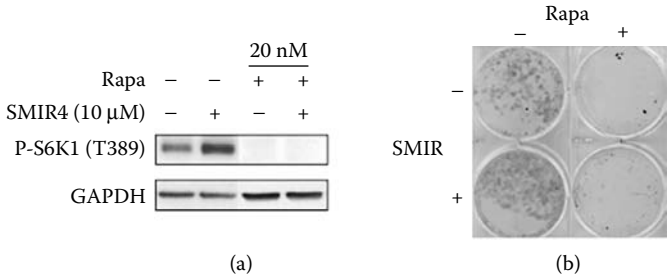


FIGURE 14.2 (a) SMIR4 effect on S6K1 detected by Western blot analysis using the phosphorylation status of Thr-389 as a readout. (b) SMIR effect on adipogenesis. 3T3-L1 cells are treated with a differentiation cocktail (insulin + IBMX + dexamethasone) in the presence and absence of SMIR, and adipogenesis is assayed by a simple staining method using oil red O, which stains the lipid droplets in differentiated adipocytes.[95] See color insert following page 236.

not least, whereas Ybr077c itself does not have an obvious mammalian homolog, SMIR4 has shown various activities in mammalian systems. In addition to the preliminary Jurkat cell transcript profiling data described in our original paper, we found (i) that SMIR4-treated cells exhibit hyperphosphorylation of S6K1, a direct target of mammalian TOR,[84] and (ii) that SMIR4 enhances pre-adipocyte differentiation (Raymond Wu, Fulai Jin, and J.H., unpublished results; see Figure 14.2 for an example) in the 3T3-L1 cell model.[89] Both these effects are opposite to those elicited by the TOR inhibitor rapamycin (although SMIR4 does not reverse rapamycin's effect efficiently in this system when both compounds are added together),[90,91] consistent with a positive effect of SMIR4 on the mammalian TOR pathway activity.

These results illustrate that small molecules discovered by using the yeast TOR model system are translatable for potential use in mammalian cells. In this case, the pro-adipogenesis phenotype gives high hope that SMIR4 (or SMIR4-like molecules) may exert antidiabetic effect *in vivo* or serve as leads for the development of potent diabetes drugs.

We hope that our preliminary success with proteome microarrays in target identification will encourage others to explore this promising new target identification platform. The advance should greatly facilitate the identification of protein targets modulated by drugs or small molecules obtained from combinatorial chemical syntheses and chemical-genetic screens. Furthermore, since nearly *all* possible targets are examined on the proteome chip, "off-targets" will also be revealed, which can help anticipate and avoid therapeutic side effects, as well as reveal entirely novel effects of known compounds. The proteome microarray technology thus has the potential to revolutionize the field of drug target identification and mechanism studies.

LIMITATIONS AND FUTURE CHALLENGES

A major concern of the protein chip technology is whether the immobilized proteins on a solid surface retain their native conformation and maintain their functionality. Since about a third of the yeast proteome contains proteins of unknown function, it

is not feasible to determine the percentage of functional proteins on a chip. Nevertheless, we and others have performed a wide variety of biochemical assays on chips, including protein–protein,[2] protein-phospholipid,[2] protein–DNA,[9] protein–small molecule,[10] and antigen-antibodies interactions (Zhu et al., unpublished),[23] as well as enzymatic reactions to identify substrates of kinases,[1,11] and phosphatases (unpublished), from which a plethora of known and unknown activities of immobilized proteins have been identified. For example, 150 proteins on the yeast proteome chips demonstrated specific binding activities to five different phosphatidylinositdes (PIPs); ~80% of the immobilized protein kinases showed significant autophosphorylation activities; >200 proteins on chips could bind to genomic DNA; 14 proteins were specifically recognized by a small molecule identified from a chemical genetic screen. These results indicate that a significant portion of the immobilized proteins is functional on the chip. Another concern is that the position of an affinity tag used for purification may interfere with the function of a given protein. This problem has been addressed by the construction of two collections of fusion proteins, each tagged at the N- or C-terminus, respectively.[2,92] Therefore, we believe that by using site-directed immobilization coupled with N- and C-terminal fusion proteins, there is a good chance that the majority of immobilized proteins on glass surfaces should be active.

If no targets are identified for a small molecule, one possible reason is that the target may not be represented on the proteome chip or the particular protein on the chip may be nonfunctional. In that case, affinity chromatography using biotin-labeled compounds may be required. It is also possible that a particular small molecule may have targets other than proteins, such as RNA molecules, lipids, carbohydrates, etc. In that case, affinity chromatography followed by mass spectrometry analysis should give us clues about the nature of the target.

Besides attaching affinity tags to small molecules of interest, label-free detection is expected to be a powerful tool in this realm in the future. Surface plasma resonance (SPR)[12] has emerged as an important means of label-free detection on glass surfaces coated with gold. In SPR, analytes are first immobilized on the gold-coated chip, probes are then loaded to the surface, and the interactions are detected as function of the change in reflection of light caused by the interactions. SPR has been shown useful in measuring various kinds of interactions, including small molecule–protein interactions. However, this technology cannot yet reach the throughput comparable to that of the protein microarray technology. One of the promising alternatives is the so-called Epic technology developed by Corning in recent years. It utilizes a 384 well microplate with optical biosensors and attachment surface chemistry inside each well, an optical reader detection instrument with liquid handling, and label-independent assay protocols. The Epic optical reader is capable of generating binding data in real time, which is useful for assay development, as well as medium-throughput reading of 384-well plates. A major drawback of the Epic system is that it currently requires fairly large amount (e.g., in a milligram range) of purified proteins to be immobilized in 384-well plates, and it is likely to be several years before it can be applied to screen an entire eukaryotic proteome. Developments in label-free detection using nanowire sensors and other approaches[93,94] are especially promising for ultrasensitive parallel applications.

REFERENCES

1. Zhu, H. et al., Analysis of yeast protein kinases using protein chips, *Nat. Genet.,* 26, 283, 2000.
2. Zhu, H. et al., Global analysis of protein activities using proteome chips, *Science*, 293, 2101, 2001.
3. Zhu, H. and Snyder, M., Protein arrays and microarrays, *Curr. Opin. Chem. Biol.*, 5, 40, 2001.
4. Zhu, H. and Snyder, M., Protein chip technology, *Curr. Opin. Chem. Biol.*, 7, 55, 2003.
5. Phizicky, E. et al., Protein analysis on a proteomic scale, *Nature*, 422, 208, 2003.
6. Joos, T.O. et al., A microarray enzyme-linked immunosorbent assay for autoimmune diagnostics, *Electrophoresis*, 21, 2641, 2000.
7. Templin, M.F. et al., Protein microarrays: Promising tools for proteomic research, *Proteomics*, 3, 2155, 2003.
8. MacBeath, G. and Schreiber, S.L., Printing proteins as microarrays for high-throughput function determination, *Science*, 289, 1760, 2000.
9. Hall, D.A. et al., Regulation of gene expression by a metabolic enzyme, *Science*, 306, 482, 2004.
10. Huang, J. et al., Finding new components of the target of rapamycin (TOR) signaling network through chemical genetics and proteome chips, *Proc. Natl. Acad. Sci. USA*, 101, 16594, 2004.
11. Ptacek, J. et al., Global analysis of protein phosphorylation in yeast, *Nature*, 438, 679, 2005.
12. Myszka, D.G. and Rich, R.L., Implementing surface plasmon resonance biosensors in drug discovery, *Pharm. Sci. Technol. Today*, 3, 310, 2000.
13. Houseman, B.T. et al., Peptide chips for the quantitative evaluation of protein kinase activity, *Nat. Biotechnol.*, 20, 270, 2002.
14. LeProust, E. et al., Digital light-directed synthesis. A microarray platform that permits rapid reaction optimization on a combinatorial basis*, J. Comb. Chem.*, 2, 349, 2000.
15. Ramachandran, N. et al., Self-assembling protein microarrays, *Science*, 305, 86, 2004.
16. Wang, D. et al., Carbohydrate microarrays for the recognition of cross-reactive molecular markers of microbes and host cells, *Nat. Biotechnol.*, 20, 275, 2002.
17. Ratner, D.M. et al., Probing protein-carbohydrate interactions with microarrays of synthetic oligosaccharides, *Chembiochem*, 5, 379, 2004.
18. MacBeath, G., Koehler, A.N. and Schreiber, S.L., Printing small molecules as microarrays and detecting protein-ligand interactions en masse, *J. Am. Chem. Soc.*, 121, 7967, 1999.
19. Hergenrother, P.J., Depew, K.M., and Schreiber, S.L., Small-molecule microarrays: Covalent attachment and screening of alcohol-containing small molecules on glass slides, *J. Am. Chem. Soc.*, 122, 7849, 2000.
20. Barnes-Seeman, D. et al., Expanding the functional group compatibility of small-molecule microarrays: Discovery of novel calmodulin ligands, *Angew. Chem. Int. Ed. Engl.*, 42, 2376, 2003.
21. Lueking, A. et al., Protein microarrays for gene expression and antibody screening, *Anal. Biochem.*, 270, 103, 1999.
22. Arenkov, P. et al., Protein microchips: Use for immunoassay and enzymatic reactions. *Anal. Biochem.*, 278, 123, 2000.
23. Michaud, G.A. et al., Analyzing antibody specificity with whole proteome microarrays, *Nat. Biotechnol.*, 21, 1509, 2003.
24. Lesaicherre, M.L. et al., Intein-mediated biotinylation of proteins and its application in a protein microarray, *J. Am. Chem. Soc.*, 124, 8768, 2002.

25. Schweitzer, B., Predki, P., and Snyder, M., Microarrays to characterize protein interactions on a whole-proteome scale, *Proteomics*, 3, 2190, 2003.
26. Bussow, K. et al., A human cDNA library for high-throughput protein expression screening, *Genomics*, 65, 1, 2000.
27. Brizuela, L. et al., The FLEXGene repository: Exploiting the fruits of the genome projects by creating a needed resource to face the challenges of the post-genomic era, *Arch. Med. Res.*, 33, 318, 2002.
28. Strausberg, R.L. et al., The mammalian gene collection, *Science*, 286, 455, 1999.
29. Strausberg, R.L. et al., Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences, *Proc. Natl. Acad. Sci. USA*, 99, 16899, 2002.
30. Jones, R.B. et al., A quantitative protein interaction network for the ErbB receptors using protein microarrays, *Nature*, 439, 168, 2006.
31. Schreiber, S.L., Chemical genetics resulting from a passion for synthetic organic chemistry, *Bioorg. Med. Chem.*, 6, 1127, 1998.
32. Strausberg, R.L. and Schreiber, S.L., From knowing to controlling: A path from genomics to drugs using small molecule probes, *Science*, 300, 294, 2003.
33. Demain, A.L., Prescription for an ailing pharmaceutical industry, *Nat. Biotechnol.*, 20, 331, 2002.
34. Clardy, J. and Walsh, C., Lessons from natural molecules, *Nature*, 432, 829, 2004.
35. Corey, E.J. and Cheng, X.-M., *The Logic of Chemical Synthesis*, John Wiley, New York, 1989.
36. Schreiber, S.L., Target-oriented and diversity-oriented organic synthesis in drug discovery, *Science*, 287, 1964, 2000.
37. Gray, N.S., Combinatorial libraries and biological discovery, *Curr. Opin. Neurobiol.*, 11, 608, 2001.
38. Spring, D.R., Diversity-oriented synthesis; a challenge for synthetic chemists, *Org. Biomol. Chem.*, 1, 3867, 2003.
39. Burke, M.D. and Schreiber, S.L., A planning strategy for diversity-oriented synthesis, *Angew. Chem. Int. Ed. Engl.*, 43, 46, 2004.
40. Khosla, C. and Keasling, J.D., Metabolic engineering for drug discovery and development, *Nat. Rev. Drug Discov.*, 2, 1019, 2003.
41. Dervan, P.B., Molecular recognition of DNA by small molecules, *Bioorg. Med. Chem.*, 9, 2215, 2001.
42. Halpin, D.R. and Harbury, P.B., DNA display II. Genetic manipulation of combinatorial chemistry libraries for small-molecule evolution, *PLoS Biol.*, 2, E174, 2004.
43. Li, X. and Liu, D.R., DNA-templated organic synthesis: Nature's strategy for controlling chemical reactivity applied to synthetic molecules, *Angew. Chem. Int. Ed. Engl.*, 43, 4848, 2004.
44. Tan, D.S., Diversity-oriented synthesis: Exploring the intersections between chemistry and biology, *Nat. Chem. Biol.*, 1, 74, 2005.
45. Bishop, A. et al., Unnatural ligands for engineered proteins: New tools for chemical genetics, *Annu. Rev. Biophys. Biomol. Struct.*, 29, 577, 2000.
46. Zhang, J. et al., Creating new fluorescent probes for cell biology, *Nat. Rev. Mol. Cell. Biol.*, 3, 906, 2002.
47. Wang, L. and Schultz, P.G., Expanding the genetic code, *Angew. Chem. Int. Ed. Engl.*, 44, 34, 2004.
48. Jeong, H. et al., Lethality and centrality in protein networks, *Nature*, 411, 41, 2001.
49. Albert, R., Jeong, H. and Barabasi, A.L., Error and attack tolerance of complex networks, *Nature*, 406, 378, 2000.

50. Jin, F. et al., A pooling-deconvolution strategy for biological network elucidation, *Nat. Methods*, 3, 183, 2006.
51. Mayer, T.U. et al., Small molecule inhibitor of mitotic spindle bipolarity identified in a phenotype-based screen, *Science*, 286, 971, 1999.
52. Koh, B. and Crews, C.M., Chemical genetics: A small molecule approach to neurobiology, *Neuron*, 36, 563, 2002.
53. Koeller, K.M. et al., Chemical genetic modifier screens: Small molecule trichostatin suppressors as probes of intracellular histone and tubulin acetylation, *Chem. Biol.*, 10, 397, 2003.
54. Zhang, X. et al., A potent small molecule inhibits polyglutamine aggregation in Huntington's disease neurons and suppresses neurodegeneration *in vivo*, *Proc. Natl. Acad. Sci. USA*, 102, 892, 2005.
55. Smukste, I. and Stockwell, B.R., Advances in chemical genetics, *Annu. Rev. Genom. Hum. Genet.*, 261, 2005.
56. Burdine, L. and Kodadek, T., Target identification in chemical genetics: The (often) missing link, *Chem. Biol.*, 11, 593, 2004.
57. Tochtrop, G.P. and King, R.W., Target identification strategies in chemical genetics, *Comb. Chem. High Throughput Screen*, 7, 677, 2004.
58. Gray, N.S. et al., Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors, *Science*, 281, 533, 1998.
59. Armstrong, J.I. et al., Discovery of carbohydrate sulfotransferase inhibitors from a kinase-directed library, *Angew. Chem. Int. Ed. Engl.*, 39, 1303, 2000.
60. Wu, X. et al., Purmorphamine induces osteogenesis by activation of the hedgehog signaling pathway, *Chem. Biol.*, 11, 1229, 2004.
61. Sinha, S. and Chen, J.K., Purmorphamine activates the Hedgehog pathway by targeting Smoothened, *Nat. Chem. Biol.*, 2, 29, 2006.
62. Kuruvilla, F.G. et al., Dissecting glucose signalling with diversity-oriented synthesis and small-molecule microarrays, *Nature*, 416, 653, 2002.
63. Harding, M.W. et al., A receptor for the immunosuppressant FK506 is a cis-trans peptidyl-prolyl isomerase, *Nature*, 341, 758, 1989.
64. Liu, J. et al., Calcineurin is a common target of cyclophilin-cyclosporin A and FKBP-FK506 complexes, *Cell*, 66, 807, 1991.
65. Sabatini, D.M. et al., RAFT1: A mammalian protein that binds to FKBP12 in a rapamycin-dependent fashion and is homologous to yeast TORs, *Cell*, 78, 35, 1994.
66. Brown, E.J. et al., A mammalian protein targeted by G1-arresting rapamycin-receptor complex, *Nature*, 369, 756, 1994.
67. Fenteany, G. et al., Inhibition of proteasome activities and subunit-specific amino-terminal threonine modification by lactacystin, *Science*, 268, 726, 1995.
68. Taunton, J., Hassig, C.A., and Schreiber, S.L., A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p, *Science*, 272, 408, 1996.
69. Ding, S. et al., Synthetic small molecules that control stem cell fate, *Proc. Natl. Acad. Sci. USA*, 100, 7632, 2003.
70. Wan, Y. et al., Synthesis and target identification of hymenialdisine analogs, *Chem. Biol.*, 11, 247, 2004.
71. Heitman, J., Movva, N.R., and Hall, M.N., Targets for cell cycle arrest by the immunosuppressant rapamycin in yeast, *Science*, 253, 905, 1991.
72. Giaever, G. et al., Genomic profiling of drug sensitivities via induced haploinsufficiency, *Nat. Genet.*, 21, 278, 1999.
73. Luesch, H. et al., A genome-wide overexpression screen in yeast for small-molecule target identification, *Chem. Biol.*, 12, 55, 2005.

74. Butcher, R.A. et al., Microarray-based method for monitoring yeast overexpression strains reveals small-molecule targets in TOR pathway, *Nat. Chem. Biol.*, 2, 103, 2006.
75. Sche, P.P. et al., Display cloning: Functional identification of natural product receptors using cDNA-phage display, *Chem. Biol.*, 6, 707, 1999.
76. McKenzie, K.M. et al., Simultaneous identification of multiple protein targets by using complementary-DNA phage display and a natural-product-mimetic probe, *Angew. Chem. Int. Ed. Engl.*, 43, 4052, 2004.
77. Licitra, E.J. and Liu, J.O., A three-hybrid system for detecting small ligand-protein receptor interactions, *Proc. Natl. Acad. Sci. USA*, 93, 12817, 1996.
78. Won, J. et al., A magnetic nanoprobe technology for detecting molecular interactions in live cells, *Science*, 309, 121, 2005.
79. Jiang, X. et al., Distinctive roles of PHAP proteins and prothymosin-alpha in a death regulatory pathway, *Science*, 299, 223, 2003.
80. Verma, R. et al., Ubistatins inhibit proteasome-dependent degradation by binding the ubiquitin chain, *Science*, 306, 117, 2004.
81. Mitsopoulos, G., Walsh, D.P., and Chang, Y.T., Tagged library approach to chemical genomics and proteomics, *Curr. Opin. Chem. Biol.*, 8, 26, 2004.
82. Colca, J.R. and Harrigan, G.G., Photo-affinity labeling strategies in identifying the protein ligands of bioactive small molecules: Examples of targeted synthesis of drug analog photoprobes, *Comb. Chem. High Throughput Screen,* 7, 699, 2004.
83. Loewith, R. et al., Two TOR complexes, only one of which is rapamycin sensitive, have distinct roles in cell growth control, *Mol. Cell.*, 10, 457, 2002.
84. Kim, D.H. and Sabatini, D.M., Raptor and mTOR: Subunits of a nutrient-sensitive complex, *Curr. Top. Microbiol. Immunol.*, 279, 259, 2004.
85. Huh, W.K. et al., Global analysis of protein localization in budding yeast, *Nature*, 425, 686, 2003.
86. Dubouloz, F. et al, The TOR and EGO Protein Complexes Orchestrate Microautophagy in Yeast, *Mol. Cell.*, 19, 15, 2005.
87. Audhya, A. et al, Genome-wide lethality screen identifies new PI4,5P2 effectors that regulate the actin cytoskeleton, *Embo. J.*, 23, 3747, 2004.
88. Fang, Y. et al., Phosphatidic acid-mediated mitogenic activation of mTOR signaling, *Science*, 294, 1942, 2001.
89. Rosen, E.D. and Spiegelman, B.M., Molecular regulation of adipogenesis, *Annu. Rev. Cell. Dev. Biol.*, 16, 145, 2000.
90. Yeh, W.C., Bierer, B.E., and McKnight, S.L., Rapamycin inhibits clonal expansion and adipogenic differentiation of 3T3-L1 cells, *Proc. Natl. Acad. Sci. USA*, 92, 11086, 1995.
91. Kim, J.E. and Chen, J., Regulation of peroxisome proliferator-activated receptor-gamma activity by mammalian target of rapamycin and amino acids in adipogenesis, *Diabetes*, 53, 2748, 2004.
92. Gelperin, D.M. et al., Biochemical and genetic analysis of the yeast proteome with a movable ORF collection, *Genes Dev.*, 19, 2816, 2005.
93. Ramachandran, N. et al., Emerging tools for real-time label-free detection of interactions on functional protein microarrays, *Febs. J.*, 272, 5412, 2005.
94. Wang, W.U. et al., Label-free detection of small-molecule-protein interactions by using nanowire nanosensors, *Proc. Natl. Acad. Sci. USA*, 102, 3208, 2005.
95. Green, H. and Kehinde, O., An established preadipose cell line and its differentiation in culture. II. Factors affecting the adipose conversion, *Cell,* 5, 19, 1975.

15 Antibody Profiling for Protein Drug Development and Clinical Development

Steve H. Herrmann

CONTENTS

INTRODUCTION

The goal of medicinal therapy is to improve patients' health and quality of life. Optimal medicinal therapy should be safe, effective, judiciously chosen and cost-effective. There should be equity of access to medicinal care and an accurate and up-to-date information base, meeting the needs of patients and providers (http://www.wma.net/e/policy/ m33.htm). Protein drugs, also referred to as biologics or biotherapeutics, have become an important part of medicinal therapy and the fastest growing class of biotherapeutics are antibodies and receptor Fc's.[1–3] The commonality between these two protein types are a binding site and the Fc, or

crystallizable portion, of immunoglobulin G (IgG). The ability of the Fc region to greatly increase the half-life by utilization of a receptor was first suggested by Brambell[4] and later shown to correspond with the neonatal Fc receptor, or FcRn. In addition to transporting maternal Ig to the fetus and infant, FcRn in the adult can prolong Ig half-life by exocytosis of endocytosed Ig.[5–11] The long half-life (one to three weeks) and precise specificity of a bio-therapeutic for its antigen/ligand are the two most important aspects differentiating this class of medicinals from the classical ligand small-molecule drugs (such as modern-day statins). Half-life can be easily measured for large and small molecule medicinals alike. On the other hand, demonstrating absolute target specificity of the target for both classes of therapeutics is something often claimed but almost never conclusively proven. It is often stated that proteins, due to their large size, are much more specific compared to small molecules and, compared with small molecules, thus much less likely to suffer from serious or off-target toxic effects seen with some small-molecule drags. Government regulatory agencies are unlikely to take this statement as fact.

FROM GENE TO THERAPEUTIC ...

Once a gene is discovered, a number of approaches can be used to ask if the protein encoded by the gene is linked to a disease. One obvious attribute of a biotherapeutic approach is to test the linkage between the target protein and disease through an antibody or R-Fc. Protein drugs can be generated to behave as agonists or antagonists for nearly any protein that normally binds to a cell surface receptor and delivers a signal. By testing *in vitro*, usually cell culture, confirmation of stimulation or inhibition of the target can be shown for the drug candidate. Preclinical or animal model testing to support efficacy for the chosen indication follows this. This path can result in rapid development of therapeutics that are next tested clinically in humans. This is great from the standpoint of bringing new drugs to market rapidly. However, this speed and the ability of the biologic to specifically effect the target, can trigger phase I trials before a full understanding of the biology, or consequence of the protein drug acting on its given target, is understood. Without a complete understanding of the biology of the target and the consequences of altering its normal *in vivo* behavior, toxicity of the protein therapeutic is a possibility, just as it is for small molecule drug targets. Without a sufficient understanding of the biology, genetic or disease differences between patients can result in the drug's showing efficacy in a subset of the population, and failed clinical trials. Our imprecise understanding of biology and the individual differences in the human population mean some very successful drugs work in some but not all patients. If we don't fully understand at the molecular level why these successful drugs are efficacious, it is not surprising we frequently enter clinical trials with biological uncertainty. Consequently, predictions of success may reflect project management's hopes more than a detailed understanding of the biology. There is an attempt to increase our chances of quickly understanding if a drug is working by measuring bio-markers,[12–15] which are typically changes in serum or blood components that correlate with drug efficacy and the switch from disease to health.

WHAT MAKES A SUCCESSFUL BIOTHERAPEUTIC?

Currently two monoclonal antibodies (mAbs) and one R-Fc, all able to antagonize TNF, have been approved for treatment of rheumatoid arthritis (RA). [(Enbrel (etanercept) approved in 1998, Remicade, (infliximab) approved in 2001, Humira, (adalimumab) approved 2003)].[1,3] These are successful drugs because they are effective at reducing the RA-related symptoms. However, they do not work in all patients and there is not agreement upon why this approach fails in some patients and whether efficacy or side effects are due to how a certain biotherapeutic blocks TNF.[16–23] While these therapeutics are not without side effects, in a segment of the population these biologicals work extremely well to reverse the symptoms of rheumatoid arthritis in addition to several other autoimmune diseases. Currently several new TNF inhibitory proteins are undergoing clinical trials. It is not clear if these will be better from a cost, efficacy, or side effect standpoint.

Some of the non-antibody-based protein therapeutics are native proteins, or similar to native proteins, that normally circulate in the blood such as Neupogen (filgrastim) and the PEG modified long half-life form, Neulasta (pegfilgrastim), as well as a variety of erythropoietins; PROCRIT (Epoetin alfa), Epogen (Epoetin alfa) and the modified longer half-life form, Aranesp (darbepoetin). These drugs have enabled oncologists to give a much higher level of myelosuppressive chemotherapy to patients while maintaining neutrophile and red blood cell levels. These growth factors have been widely used and while safe are not completely free of adverse or toxic side effects as discussed below.

SIDE EFFECTS OF BIOTHERAPEUTICS?

For any drug, when toxic side effects occur, this may be due to the biology of the intended target modified by the therapeutic in an expected or unexpected manner, or the observed toxicity may be due to the therapeutic acting upon an unintended target. Thus proteins as well as small molecules may bind to an unintended target and cause unintended results. The clinical experience with several protein drugs will be used to look at the spectrum of potential toxic effects of biotherapeutics.

The severity seen in this first example, an antibody targeting CD28, was unprecedented, resulting in widespread coverage in the media and some scientific journals. This antibody was supposed to activate suppressor T cells, also known as regulatory T cells, and the anticipated result was a dampening of the immune response. This was being developed for patients with autoimmune disease. In phase I trials, healthy individuals (simultaneously) received the anti-CD28 antibody and all experienced an immediate life threatening response. This resulted in halting all work with this target and triggered editorials calling for transparency in the pharmaceutical industry; a call for a panels of experts empowered to approve or block any novel therapy; and questions as to the specificity of this antibody.[24–29] The exact cause of this anti-CD28 related toxic effect has not been established. Questions exist regarding the purity or nature of the antibody itself as well as the wisdom of generating a therapeutic against this target and the level of reactivity of the antibody against the target in preclinical animal models. If the side effects were a direct action of the antibody binding to

the CD28, T cell costimulatory molecule, this underscores the power of biotherapeutics to modify the biology and health of the patient. The response seen may have been driven by inappropriate binding of the antibody, to CD28 and other cell surface proteins, or due to some excipients in the drug product or simply binding was specific but triggered an unanticipated 'super' response.

The second example involves a clinical trial using an antibody able to block an integrin, alpha 4, and being developed for treatment of multiple sclerosis. Initial clinical trials supported efficacy and the antibody was approved, but, then put on hold when it became clear that the progressive multifocal leukoencephalopathy (PML), seen in some patients, most likely due to infection with JC-virus, were linked to the biotherapeutic.[1,30–32] There are several theories attempting to link the action of this anti-integrin with the resulting PML, including some evidence that this antibody can block the trafficking of lymphocytes into the cerebral spinal fluid.[33] At this time, the drug is being reviewed, the risks/benefits ratio. These clinical studies have shown that blocking the immune system, while possibly beneficial to the MS patient, may allow virus normally controlled by the immune system to expand. As already discussed, several autoimmune diseases are controlled by biotherapeutics blocking TNF. This level of inhibition of the immune system can lead to bacterial related side effects such as re-emergence of tuberculosis in a previously infected individual.[34–36] This has led to prescreening patients for tuberculosis and other infectious agents. Most agree that the incidence of infection associated with blocking TNF is an acceptable risk. The toxicity with blocking the cell surface integrin receptor or blocking TNF is unlikely to be due to off-target binding but due to blocking different immune response processes.

A third example of toxicity related to the therapeutic is an antibody that binds a tyrosine kinase linked receptor over-expressed on certain types of malignant cells. While targeting this ErbB2 receptor appears to be beneficial for treatment of certain types of tumors[37–39] there is a concern with cardiac toxicity in some patients.[40–42] This cardiac toxicity may be due to off-target binding or a role played by ErbB2 in cardiac function.[41] This anti-Her2 antibody is used for cancer therapy, an area where most of the small molecule therapeutics are associated with a much higher level of toxic side effects. For this therapeutic it may be the only choice some patients have for survival and the side effects, while important, are clearly judged acceptable.

In use since approval in 1993, erythropoietin (EPO) is a protein primarily expressed in the kidney in response to hypoxia that controls red blood cell (RBC) production. Regarded as an excellent drug and viewed as extremely safe, it can still cause side effects under the certain conditions. Currently there are at least 4 different marketed versions of EPO physicians can prescribe to increase a patients' level of RBC. Thousands of patients have been treated and responded by an increase in red blood cells. However, there have been some toxic side effects seen in a small percentage of these patients receiving product from one vendor. These side effects came about as a result of the patients' immune system responding to the EPO as if it were a foreign protein and generating an immune response against the drug product.[43–46] Since this therapeutic is a normal factor involved in the regulation of red blood cell production the anti-drug product response cross-reacted with endogenous EPO and resulted in red cell aplasia in some patients.[46–48] For some patients, this response was believed to be due to the

drug product containing some materials that acted as an adjuvant to induce the immune response.[47,48] Thus even safe protein drugs can cause toxic side effects if they are not in a native form or contain materials that can stimulate an inflammatory response.

Biotherapeutics, like small molecules, are not exempt from side effects and the physician must weigh the risks with the benefits. The challenge facing the drug industry is how to put into practice approaches that will generate safe new drugs — drugs that are effective and at the same time free of side effects. This requires a through understanding of the biology along with an understanding of the specificity of the biotherapeutic. Antibodies are poised to provide the dominant biotherapeutic platform for multiple disease. This is possible due to the short time frame for generation, combined with the perceived specificity and tight binding to the target. Key to this success will be maintaining the patient population's trust by bring forth only safe drugs where the risk is known and far lower than the medicinal benefit.

GENERATION OF THERAPEUTIC ANTIBODIES

There are at least four basic approaches (not counting anti-sera) used to generate therapeutic antibodies. The earliest monoclonal antibodies (mAb) were generated by immunizing mice with the human protein and select hybridomas expressing antibodies that bind to the target of interest. The inability of the mouse Fc to bind the human FcRn[5,6,10,49] and the rapid human anti-mouse antibody (HAMA) response, resulted in a rapid clearance of the mouse antibody. Chimeric molecules, having a mouse variable region and a human Fc region improved the half-life and decreased the HAMA. The chimeric antibody Rituxan (rituximab) is one such example of a very successful biotherapeutic.[50,51]

A modification of the chimeric approach termed humanization starts with a mouse or other nonhuman mAb and while maintaining the variable or antigen binding and specificity region, the rest of the molecule, the Fc region and heavy and light chain framework, is reformatted to have a human framework sequences. Currently several different approaches to humanization are practiced, all striving to end up with the majority of the sequence matching some human sequence. This can be done by keeping the variable or antigen binding region of the mouse antibody, also termed the complementarity determining regions (CDRs), or keeping only those amino acids in the CDRs that are involved in ligand binding.[52–55] Frequently, this requires affinity optimization of the resulting antibody to regain binding strength for the target lost during the humanization process.

A third approach uses genetically modified mice — animals engineered by replacing the mouse germline immunoglobulins with human immunoglobulins.[56–58] On the surface this seems straightforward, requiring only the immunization and selection of monoclonal antibodies in parallel to generation of mouse mAb from wild type mice. In practice, the immune response of these "xeno" mice is not robust, requiring multiple immunizations and large number of mice to find the desired antibodies. During generation of an immune response there are likely self-reactive antibodies generated as well as those against the specific immunogen.[59–62] Is the selection or central tolerance that takes place in the mouse going to yield a cross-reactive tolerance in man?[59]

Still another approach is the generation of a human antibody library by PCR-mediated isolation of heavy chain and light chain variable regions from multiple individuals. Random combinations of these heavy and light chain variable regions, usually with a linking segment between the variable heavy chain and light chain regions, forms a single-chain Fv antibody able to bind antigen. These libraries consist of very large numbers ($>10^{10}$) of unique antibody species which are then selected *in vitro* using either phage display or ribosome display approaches.[63–69] The resulting scFv antibodies from these libraries frequently need to be optimized to achieve sufficient binding to neutralize the target. Similar libraries can also be generated by the random combination of the heavy chain and light chain variable regions in the form of a Fab binding unit.[70] One advantage of the Fab selection process is the ability to readily convert the selected fab binder to a full length antibody from the starting Fab binder.

SELECTION CRITERIA FOR BIOTHERAPEUTICS

Regardless of how it is generated, there are a number of properties that the therapeutic antibody leads must possess in order to advance into the clinic. Most important is the continued evidence that the antibody or receptor-Fc is active in animal models predictive of human disease and demonstration of continued high specificity binding to the human target. This is an issue, as many human targets are not identical to the animal orthologue. If one does not develop a cross-reactive antibody or R-Fc then a surrogate is needed that binds to the exact same epitope on the target in the animal model and has the same properties as the lead. If the biotherapeutic is being developed for a chronic disease it will usually need to be given on a continual basis. This requires stable protein that can be concentrated to a high level and injected by the patient in their home. Ideally, the amount given by sub-cutaneous injection should not be more than 1 ml and this amount should contain enough protein to modify the disease state for a minimum of two weeks to a month. The binding constant for cytokines and other growth factors to its endogenous receptor is generally in the picomolar to nanomolar range. If an antibody or receptor-Fc is to effectively block this type of interaction it must be able to out-compete the receptor(s) for binding to the soluble factor. The ideal biotherapeutic needs a high binding affinity for the target, properties that allow high expression in mammalian cell lines, stability at concentrations of several hundred mg/ml, cross-reaction with the target protein in one or more animal models, and be absolutely specific for the target. In some instances if the therapeutic candidate antibody is found to lack sufficient binding affinity or has expression or solubility issues the protein is improved by a process of maturation or optimization also called *in vitro* evolution. This is usually done employing random and site directed mutagenesis, using phage or ribosome display approaches.[63,68,71] Throughout these selection and development phases, the biotherapeutic needs to be repeatedly tested to show consistent stability and specificity.

ONE ANTIBODY, ONE TARGET?

Are antibodies specific? We are attempting to develop antibodies that can cross-react with the animal model orthologues and after generation of the initial antibody, mouse or human, the protein must then be subjected to a process to humanize and/or optimize.

Can we assume perfect specificity? For antibodies generated from human libraries, even if the library was generated from normal non-autoimmune individuals, the approach of randomly combining heavy and light chains may result in generation of variable regions that are not specific for the target. There are indications that generation of an antibody able to recognize one antigen can result in an antibody that recognizes a second antigen as well.[72–75] Others examining a cholera toxin mAb have found it to recognize two other antigen epitopes unrelated to the cholera epitope[76] and the authors have suggested that mAbs may be poly specific. One mAb binding to multiple epitopes has also been indicated when two mAb to the same collagen epitope were shown to bind different peptides in a peptide phage display library.[77]

Another source of self-reactive antibodies from a normal donor would be natural self-reactive antibodies that have been identified against a host of antigens including Factor VIII, A-beta, mitochondrial proteins and vascular proteins.[78–82] There is also the suggestion that some antibodies can have cofactors that influence their binding as well as their biological properties.[83,84]

What makes a mAb or a R-Fc specific? Immunization of humans via nasal delivery of an influenza vaccine, Nasaflu, an inactivated virosome-formulated subunit vaccine, resulted in a low number of patients (11/1526) with facial paralysis.[85] Although the facial paralysis reversed in all but one patient, this side effect resulted in removal of this vaccine from the market in 2001. Is this 0.7% incidence due to some cross-reactive response of the antibodies generated in patients to the influenza protein that resulted in facial paralysis by cross-reactivity with a self antigen? Data to support this was not provided in this study and a follow up study support the use of a different virosome formulated vaccine.[86] There have been other reported side effects that may be due to an inappropriate antibody response causing symptoms termed oculo-respiratory syndrome (0.05% of patients in 2000) following intramuscular vaccination with inactivated split-virion influenza vaccine The symptoms were not linked to a hyperimmune response but appear to be due to the generation of a response in a small population that give rise to the syndrome. The vaccine literature contains many examples of side effects that appear to be due to some cross-reactive response. So, the generation of an immune response against one antigen may result in antibodies that bind to a different, nonrelated antigen. In the situation of vaccination, where each patient will generate a range of different antibodies it may be near impossible to formulate a vaccine so no single individual generates an autoimmune response. There is growing evidence for a link between infection with certain viral or bacterial pathogens and the development of human autoimmune disease including rheumatoid arthritis.[79,88–92] While this link is not understood at the molecular level it is in line with side effects following vaccination with one antigen generating an immune response against the immunizing antigen as well as other self antigens leading to an auto-reactive or autoimmune antibody response in humans. Immunization of nonhuman species to generate a therapeutic antibody candidate or selection of antibodies from immunoglobulin library may contain antibodies that can react with self proteins. Starting with mouse derived or human derived antibody does not rule out the possibility of obtaining an antibody that binds to the intended target as well as another self protein resulting in a an autoimmune like response. Those developing the biotherapeutic should recognize the possibility of a therapeutic

protein cross-reacting with a self-determinant. In addition to the primary concern, the patient's health, a thorough testing of a biotherapeutic for specificity allows one to link any side effects of a therapeutic to the biology of the therapeutic rather than some off-target effect. The challenge: how to develop the best process for testing specificity? The methodology for generation and selection of a therapeutic antibody may allow for the antibody to bind more than then intended antigen. But in contrast to vaccines, we can test different candidate biotherapeutics and select those that have the highest specificity and lowest risk of off-target binding.

While there have not been indications of toxicities or off-target response for Rituxan, the CD20 specific mAb, two recent publication using peptide libraries to define the binding epitopes[93,94] show the ability of peptides nonrelated to CD20 to be bound by this antibody. These studies identified the ANPS motif using a 7-mer cyclic library in agreement with earlier work[95] which corresponds to the major extracellular portion of CD20. They also found an epitope mapped with linear peptide libraries, WPxWLE, that do not map to any CD20 sequence (or any other known sequence). The CD20 related peptide and the WPxWLE peptides were shown to cross-block each other, compete with Rituxan for binding to cell surface CD20 and the peptides could be used to generate CD20 reactive mouse antisera.[93] This reactivity with a sequence nonrelated peptide underscores the potential for antibodies to recognize structural determents on unrelated proteins. Even more this work highlights our level of understanding protein epitopes by showing that the non-CD20 related sequence was able to compete with Rituxan both as the forward sequence as well as the reverse sequence. Some but not all structural biologists agree that the inverse protein will give the same overall structure[96] and thus the same epitope is possible for two sequences with the same series of amino acids in the forward and reverse direction. Peptide mapping is frequently used attempting to map the epitope for a given antibody. While this approach does at times give an insight into the binding epitope it is not always accurate and one must be aware of a number of issues with this approach.[97] However, the more recent paper mapping the CD20 epitope does seem to bridge the gap between nonconserved peptide epitope and antigen site. These authors believe the WWEWS/T epitope they identified from the phage display, somewhat different from that above, is mimicking the YCYSI segment of CD20 and the real epitope of Rituxan is a discontinuous one consisting of ANPSI and YCYSI and dependent upon a disulfide bond between C167 and C186.[94] While this appears correct, the data from these reports show that non-antigen-related amino acid segments will frequently bind to a given antibody and if this epitope is present in humans then there may well be cross-reactive binding.

HOW TO IDENTIFY CROSS-REACTIVE BINDING FOR PROTEIN THERAPEUTICS

Another step in the pathway to developing biotherapeutics is PK studies. These studies, if carried out in an animal expressing the target antigen, will give an idea of distribution, uptake, and half-life for the therapeutic. If the target is not present the antibody half-life should be similar to a control or previously tested antibody in

a given species. Deviation from a "normal" half-life may indicate some issue with the biotherapeutic such as protease degradation, depletion due to sequestration by the target expressed at high levels or a tip off that there is some level of off-target binding.[98]

Proteins on track for clinical development must undergo a series of tests for specificity. One of these tests is immunohistochemistry, usually performed first on the animal species being used for toxicology studies. A panel of tissues from different organs is examined. In some cases cross-reactive binding to some cells in a certain organ or to certain cellular proteins is observed. When a faint background staining is seen it is difficult to determine if this is due to the presence of the antigen in the tissue or if it represents some type of cross-reactive or off-target binding? Sometimes a repeat of the IHC with different tissue samples will not give the same background staining. Was the first time a false positive? Are the new tissue sections giving the correct response? Replicate experiments searching for evidence of off-target binding may then follow. Questions arise as to the value of seeing or not seeing your therapeutic binding to tissue that is fixed, denatured, or frozen, and clearly not representing native human protein. Sometimes, due to the tissue preparation, the protein therapeutic does not bind to tissue containing the antigen. Clearly regulatory bodies will continue to require IHC. But is this nonuniform sampling of tissue the best approach to search for cross-reactive protein? When there appears to be cross-reactive binding that is persistent, the antibody or receptor-Fc usually leave the race. If the backup antibody also gives some cross-reactive binding this may send the team back to the earliest stages to generate another panel of antibodies or to work toward determining if what they are seeing by IHC is binding to a similar epitope on another protein or the presence of the target antigen in an unanticipated tissue with very low message. These types of studies are usually done late in the development process due to the cost. There are always concerns here, such as; will looking at binding to fixed nonhuman tissue really give us a good insight into how a therapeutic will work in humans? There are also issues with the overall approach. The apparent nonspecific binding may not involve the variable region of the antibody or the binding region of the receptor on the R-Fc. The apparent interaction may be chemical in nature, due to the fixative or due to free thiols in the tissue section and a reactive thiol like amino acid in the biotherapeutic.[99]

USE OF A PROTEIN ARRAY TO TEST FOR SPECIFICITY?

Could screening of potential protein therapeutics using an array consisting of a large number of different, primarily extracellular proteins (say 5000), allow one to select the best? Is it possible to use an array of different proteins on a chip and select the most stable biotherapeutic candidate by comparing the binding (or failure to bind), of each lead before and after subjecting to reversible denaturing conditions? What properties would the array need to possess to help select the best therapeutic proteins? What properties would an array of proteins need to possess to detect off-target or cross-reactive binding and satisfy the criteria of project teams so they would be confident the candidates selected based upon array profiling data were the best future biotherapeutics?

TABLE 15.1
Validation Criteria for a Protein Array

Reproducible profiles/fingerprints for sera from normal patients.
Compare clinical IVIG with normal profile
Demonstrate a change in profile during an acute disease state (viral infection)
Reproducible profile/fingerprint for sera from patient with chronic autoimmune disease (e.g. ie, rheumatoid arthritis)
Correlate change in profile with vaccination and specific reactivity toward appropriate pathogenic antigen
Correlate change in profile with induction of autoimmune state or occurrence of cancer

The first line of a list of properties or qualities most teams would demand would likely be reproducibility (Table 15.1). If off-target binding occurs with a protein under one set of conditions, this should be seen every time these conditions are used. Every time arrays are used under standard conditions the same binding profile should be obtained using the same protein. There are commercial arrays available and these do appear to satisfy the reproducibility requirement.[100–103] Another required property is that the proteins on the array should be in a native form and absolutely pure. Ideally the purified proteins must match the *in vivo* proteins they are representing so that the candidate therapeutic found binding to or failing to bind to a specific protein would correlate with what happens once injected into the patient. The proteins present on the array should include the same spectrum of proteins that an injected therapeutic should encounter once administered. Proteins should have native chemistry, be properly folded, associated with subunits, have proper *in vivo* disulfide bonding and normal post-translational modifications. The oxidation state should be such that free sulfhydryl groups would not be present on a protein, since this could give apparent reactivity similar to the disulfide interaction observed during antibody binding to protein in tissue sections during immunohistochemistry.[99] If a protein is a single pass transmembrane protein, the extracellular portion could be represented on the array without the trans membrane and cytoplasmic region. While some may argue against the need for cytoplasmic proteins on the array, the finding that many autoimmune determinants are provided by cytoplasmic determinants[62] and the ability of these antibodies to bind cytoplasmic antigen and support the generation of autoreactive T cells[104] appears to be one of the hall-marks of autoimmune disease. So in addition to all secreted and membrane proteins, a panel of intracellular proteins should also be present (Table 15.2).

If we look at these basic requirements it is unlikely current protein arrays can fulfill these ideal characteristics. Generating thousands of proteins and getting to an acceptable level of purity is a complex undertaking but possible with current technology. However, there are issues with proving any protein is in native form. What are criteria for native protein and who decides? Will a membrane protein expressed at high density on the cell surface have the same exact conformation as a soluble protein expressed without the transmembrane region and without other interacting

TABLE 15.2
Properties and Content of an Ideal Array for Biotherapeutic Profiling

Reproducability, ease of use, cost
Native Abundant extracellular proteins in active form
Native extracellular membrane proteins including heteromeric proteins and methods to demonstrate activity
Functional antibody interactive proteins such as Fc-receptors.
Native intracellular proteins that are exposed to antibody during the endocytic and FcRn mediated process
Vascular, Lymphatic and Hepatic associated proteins
Antigens recognized by "Natural" antibodies
Common auto-antibody epitopes, intracellular and extracellular protein and non-protein
Antigens associated with viral and bacterial infections such as EBV and TB
Antigens associated with autoimmune disease
Antigens associated with cancer
Antigens associated with cardiac or liver disease
Common food and environmental allergeins
Vaccine related antigens
Antigens associated with adverse events

membrane proteins? What is the effect upon the protein of spotting at a high protein concentration on a chip or glass slide? Even if the protein is shown to be fully active or possess all known functions in solution how does one verify these properties are maintained once printed on the solid support? If the protein of interest is expressed and isolated as a fusion protein with a generic tag, for purification or stability purposes, should this tag be at the carboxyl terminal end or amino terminal end? What are the ranges of *in vitro* conditions for salt, pH and redox level that proteins should be subjected to during or before testing? For these and other issues there is not a definitive path to a solution.

Using antibodies generated against native protein it should be possible to show binding to cell expressed native proteins as well as array proteins. Fc receptors on the array should be functional and able to define antibody and other proteins that may interact with these Fc binding proteins. Proteins that are only functional as a two or multichain entity should if present on the chip be in this multichain form. For example, having on the array the heavy chain of one of the representative major histocompatibility complex class I proteins without beta 2-microglobulin and stabilizing peptide would be expected to yield an unfolded protein. There are many examples of heterodimer and heterotrimer receptors such as adhesion molecules, LFA-1 or CD11a + CD18, and many cytokine receptors consisting of multiple subunits such as the ß IL-2 receptors. How likely is it that a protein array will contain all heteromeric proteins in native form?

What should be on the array? An array used as an early prescreen to remove any self-reactive antibodies must contain the dominant proteins an antibody would come in contact with. Thus, the major protein components of plasma and cerebral

spinal fluid should be present. The dominant cell surface proteins found on blood cells need be present — red blood cells, white blood cells and platelets. Antibodies will encounter proteins expressed on endothelial cells of the vascular wall and lymphatic compartments. Extracellular components, membrane proteins, and extracellular matrix proteins of major organs, liver, muscle, kidney, that the biotherapeutic will encounter should be present on the array.

Experimental data showing an antibody half-life of 7 to 10 days after injection into a normal mouse compared to less than 24 hours when injected into a mouse without a functional FcRn indicate up to 50% of injected antibody is taken up by reticuloendothelial cell pinocytosis and bound to cytoplasmic FcRn and transported back out of the cell. Having on the array proteins from the reticuloendothelial cells that the internalized antibody would come in contact with might allow one to correlate half-life differences with the level of interaction by biotherapeutic lead with these proteins. Treatment of a chronic disease will require subcutaneous (sc) injection. Depending upon dosage and the desire to have the sc injection in a small volume, if a protein is given at 3 mg/kg then a 70 kg patient would receive 210 mg of protein in 1 ml, the therapeutic injected at 420 mg/ml. Thus, having proteins on the array that represent what the therapeutic would encounter once injected at a fairly high concentration could give insight into bioavailability and injection site reactivity.

Density of Protein on the Array and Posttranslational Modification

Is there an ideal level for protein copy number per unit area? If one sees binding at a concentration or density of protein that will never occur in nature, is this type of binding information helpful? Is there value in binding data where the therapeutic lead is incubated with the array at 100s of mg/ml? Is it helpful to establish non-native conditions where non-specific protein–protein interactions will occur? Mentioned already is the issue of the oxidation state of the protein and the background that can occur due to free sulfhydryls.[99] What about redox conditions for both the antibody and the array? Recent work examining several classes of auto reactive antibodies has found that antibody reactivity often depends upon redox conditions.[105–107] One group believes that disease related nitrosylation of tyrosine residues on the antibody binding sites (CDRs) correlates with the antibody binding to self-determinants. The argument can be made that the defect of autoimmune disease is not the inability of the immune system to regulate itself but rather metabolic errors in regulating the redox state.[105] Turning this around, if a misregulated redox state is responsible for modification of antibodies, would this not also modify tissue determinants? Should one use different redox states for proteins on the array to test for cross-reactivity? If McIntyre's aberrant redox theory is correct, generating a biotherapeutic to correct an autoimmune disease, one should test the therapeutic under conditions it is likely to find once injected. Also the recent work linking antibodies reactive with citrullinated protein and disease state[108–112] suggest that some dominant autoimmune epitopes such as proteins with arginine converted into citrulline[112] should be present on the array.

Is There a Path for Using the Protein Arrays in Their Current State?

Let's look at this question another way. What questions would one be addressing using a protein array? All "human" antibodies currently in the clinic are humanized mouse antibody, or antibodies isolated from a human phage display library. Should these antibodies generated in a mouse or from a library of human antibodies not already be selected for specificity and lack any cross-reactivity with self-protein?

Let's look at mouse antibodies first. Most projects start with a target in mind that needs to be validated. This will generally occur by generating an antibody that will react with the target in an animal model. The best approach is to generate an antibody in mouse that reacts with the mouse (or other animal model) target and that also cross-reacts with the human target. To generate a mouse anti-mouse mAb usually requires the use of multiple antigen injections with some form of adjuvant or other approach to boost immunoreactivity. These antibodies are being generated with the idea in mind to bind a self protein target (and cross-react with the human target). The manner in which they are generated may encourage self-reactive, non-target antibodies be present. For the generation of human antibodies from a display library these are constructed by PCR isolation of heavy and light chain regions of existing antibodies from pools of many donors. These human heavy chain and light chain variable regions are then randomly paired to generate the library. Will this random pairing not provide a pairing of CDRs that might cross-react with a non-intended self-determinant? The antibodies selected from this library could be cross-reactive because the heavy and light chain variable regions are combined randomly. Could there be self-reactive mAb within this population? In addition to this, the prevalence of auto-reactive antibodies in the "healthy" human population has been well documented for numerous self proteins as pointed out earlier. There is also the issue of "natural" antibodies that may be involved in normal responses following tissue damage.[81,82,113] Thus, in addition to the potential that the human donors for the antibody libraries may have been in a predisease autoimmune state, there is potential for antibodies to be present that recognize self-determinants. More recent iterations of the phage display library include synthetically randomizing one or more of the CDRs and optimization of binding by random mutagenesis of the variable regions. These end products cannot be assumed to have gone through a normal selection process to remove self-reactive antibodies. Also, the data above on Rituxan and other mAb mapping the binding site using peptide phage display libraries suggest that structural epitopes other than the intended one can be bound. It follows that we cannot assume the therapeutic candidate will not cross-react to some unwanted determinant in the patient. This risk of cross-reactive binding should be low because the selection process consists of numerous secondary screens to test binding to targets related by sequence and these leads discarded upstream of PK and Tox studies. This does not completely rule out the ability to bind to some unanticipated structurally similar target but unexpected PK data will tip the team off to cross-reactive issues. As pointed out recently we have been focused on finding molecular mimicry to link the immunological side effects following vaccination or virus infection. Mimicry based on sequence similarities between virus or vaccine and the rare but some times

serious subsequent autoimmune diseases, has not been readily shown. However, instead of sequence mimicry it is likely we should be looking for a structural mimicry.[114]

One possible path to benefit from protein arrays is to attempt to use them to identify cross-reactive antibodies that we know would be harmful. Just like the small molecule drugs if the protein therapeutic binds to something other than what it is intended there can be severe toxic side effects. Many autoimmune associated antibodies have been identified.[8,24–37,62,115–128] The ability to ameliorate autoimmune symptoms by injection of high levels of pooled human immunoglobulin (IVIG) suggests auto-antibodies (some we have not identified) are involved in autoimmune disease.[129,130] An ever increasing list of antigens that can mediate autoimmune disease are being defined and some of these appear to have inflammatory properties themselves.[131] How to use this information to select biotherapeutic proteins as well as proteins on a protein array?

Worse case scenario? Assuming a biotherapeutic, with a functional FcR-binding domain, cross-reacts with an identified auto-antigen. In a pre-disposed individual this could potentially lead to the generation of an autoimmune disease by delivery of the cross-reactive antigen to antigen presenting cells.[60,61] By using arrays containing all known auto-antigens, one could monitor the biotherapeutic for cross-relativities most likely to be harmful. The list of known auto-antigens is long, containing the most common or abundant, such as, IgG, beta 2-glycoprotein I,[132] nuclear antigens[62,104,112,119,126,132–134] and natural antibody targets[81,82,113,135] as well as antigens that have been shown able to initiate an auto-immune response through their ability to also serve as chemotactic factors.[131,136] Many of these antigens have posttranslational modifications needed for their antigenicity or have modifications, such as oxidized LDL, dependent upon the person's metabolic state.[137–139]

INTERIM APPROACH?

Given the process of generating and validating the many identified auto-antigens, accomplishing this is going to be an evolving process. Without a guarantee that academic labs and industry alike will use and pay for these arrays, development of the ideal array will be slow. Needed is an interim approach that if successful would build confidence in the value of the application. Taking what we know, that even normal individuals occasionally show auto-antibodies, current arrays can be tested with normal antibodies to show that they give a fairly consistent background binding for the 10 mg/ml normal IgG that is circulating.[5] The ability of Fc receptors on the array and other proteins known to bind to the constant or Fc region of the therapeutic should be demonstrated.[140,141]

A first step would be to develop arrays, using a large enough number of normal patients to establish the binding pattern or signature and demonstrate that any binding seen to array proteins can be demonstrated for the same proteins in solution. Perhaps the starting normal Ig should come from different lots of IVIG, pooled human IgG that is used as a therapeutic.[107,129,142,143] Not only have there been reports of antibodies such as anti-A-beta reactive Ig in IVIG, and believed to be protective,[78,144] but also the success or failure of the disease being treated may depend upon the specific lot

of IVIG.[145,146] Thus, obtaining a "normal" profile for any protein array using clinical grade IVIG would result in obtaining the normal pattern as well as potentially being able to show subtle differences that could correlate with efficacy. Donor plasma for transfusions could also be screened using protein arrays to match the immunoglobulin specificities with the need.

Work has already been done, primarily from academic research laboratories, to construct arrays that contain known auto-reactive target antigens with the goal of distinguishing one autoimmune disease from another.[104,112,147–149] Antibody reactivity demonstrated for patients with rheumatoid arthritis include reactivity to collagen,[126,150] keratin, filaggrin, citrullin-modified proteins,[108,109,112,151,152] and bacterial antigens.[88,153,154] Rheumatoid Factor is usually an IgM response directed against the Fc region of IgG, however both IgA and IgG anti-Fc antibodies have been identified.[155] There is also a profiling of antibodies that recognize a range of antigens for multiple sclerosis.[156–158] Databases of antigenic sites can be cross referenced and there is even an epitope database for some autoimmune diseases that can be cross referenced.[159]

Thus, to generate a profile of which proteins will be most commonly recognized by rheumatoid arthritis patients, it would be important to have on the array the antigens already demonstrated to be recognized by most RA patients. The intensity of binding to these identified antigens could then be compared with the binding to other proteins on the protein array. The binding pattern or fingerprint of a autoimmune patient sera would indicate either some type of cross-reactive binding of antibodies to targets already identified, or indicate the response to one auto-antigen leads to reactivity against a second antigen; termed epitope spreading.[147,160,161] There is accumulating evidence that autoantibodies can be demonstrated before the onset of symptoms in type I diabetic patients.[162–164] Examining a large enough number of patients with a specific autoimmune disease would result in a standard disease related profile or fingerprint of the known antigens bound as well as other not yet identified protein targets. This profiling using autoimmune patient sera from autoimmune diseases including type 1 diabetes, Lupus, Multiple Sclerosis, psoriasis, inflammatory bowel disease and RA would establish a profile or fingerprint for each disease state. By using this approach to compare autoimmune patients with those who are predisposed to develop autoimmune disease, such as type I diabetes, it is likely that a predisease pattern could be recognized using a large protein array. The patient antibody profile will likely indicate antibodies that recognize some antigens in common as well as a specific fingerprint pattern of binding seen only for a specific autoimmune disease. This would also generate the background data set needed to use the array for profiling biotherapeutic proteins and demonstrating specificity of binding.

Thus, by carefully mapping the binding profile of IgG from the autoimmune patient, one can establish a database for comparing the binding intensity known RA autoantigens with binding to other less well known auto-antigens. Array profiles may represent binding of antibody to one antigenic epitope or antibody binding to a crossreactive sequence or structurally identical epitope.[114,159] Arrays followed over time may indicate antigen spread well documented in multiple sclerosis.[61,62] A unique profile or fingerprint may help determine whether the subject is in the early stages of a specific disease or define and follow the chronic autoimmune disease.

Subtle differences in disease can be followed by using secondary antibodies that are isotype specific to determine which Ig subclass the antibody belongs. Profiling would also identify new antigens that RA associated antibodies are binding and help track the progression of disease.

There is also emerging work using autoreactive or tumor reactive antibodies to detect cancer.[127,165,168] Most of this work has focused on trying to show tumor specific antibodies. By using a large protein array to obtain a pattern of binding for the patient with a specific cancer it should be possible to determine common antibody specificities and look for these in the population at risk for a specific cancer.

Once multiple IgG samples from normal healthy patients, different IVIG lots, patients with a specific autoimmune disease and patients with specific cancers have been profiled on arrays, what emerges may be a normal profile, a profile indicative of a specific autoimmune disease and a profile specific for a cancer class or type. It is possible these disease associated fingerprints could serve as a diagnostic in patients without symptoms of disease. With this background data, biotherapeutic hits, leads and candidate antibodies could be tested for binding to these arrays. Those antibodies giving a binding profile identical to that seen for the normal healthy patient would be selected to advance. Those candidate biotherapeutics giving a profile similar to that of an autoimmune patient would be viewed as cross-reactive, or able to bind to an antigen linked to disease and would be omitted from further development. Therapeutic leads that give a profile similar to that of normal healthy patients would be brought forward. As an antibody or R-Fc are optimized the new protein could be tested on the array to make certain the binding to the intended antigen is maintained while not acquiring the ability to bind to other determinants.

In addition to this straight-forward approach there are several other ways the array can be used in conjunction with clinical studies to develop biotherapeutics. Starting with the antibody binding profile for an autoimmune patient, such as a person with rheumatoid arthritis, the profile before and after treatment with the therapeutic could be compared. The response to specific auto antigens such as rheumatoid factor and citrullinated peptides could be followed in a quantitative manner. This would serve two purposes. If the therapeutic is working there should be a decrease in the autoimmune reactive antibodies over time and this could serve as a biomarker for efficacy. Unanticipated off-target binding with continued therapy may lead to the generation of a response against this target that could show up as a change in the overall binding pattern of the patients' Ig to the array. The initial therapeutic protein may not cross-react with a protein on the array but the cross-reactivity with a protein in the human and presentation or targeting to antigen presentation cells could lead to the *in vivo* generation of antibodies against this cross-reactive protein. This amplified response could show up on the array due to a polyclonal response or even epitope spreading. Having the therapeutic on the array could detect any immune response by the patient against the therapeutic.

By using the array to profile serum IgG following not only initial injections of a biotherapeutic but following months or years of injection of the protein therapeutic, the clinician will have a picture of what the patients' immune system is doing. If there is an immune response against the biotherapeutic this will show up. If the

therapeutic protein binds only the intended antigen and there is no immune response against this or the therapeutic then the profile should be identical to that of a healthy normal individual. If the person treated has an autoimmune disease, such as rheumatoid arthritis, then the treatment with the therapeutic should modify the initial disease profile and over time change it to reflect the normal state. The array could also be used to follow individuals immunized for influence looking for the intended response but also unintended or cross-reactive responses against self proteins. Thus in the individuals immunized for *influenza* and demonstrating facial paralysis,[85] would this correlate with a change in the antibody binding pattern of the array? While this would be expensive to do for all immunized patients, clinical trials monitoring the response of patients to the vaccine could identify any cross-reactive response. If this could be linked to some patient specific quality such as MHC type then this population of patients could be more closely followed.

It should also be possible to configure the array for a specific class of therapeutics. It is known that RA patients treated with TNF inhibitors are more susceptible to infection by some agents including intracellular pathogens such as tuberculosis.[34,36] Thus, present on the array could be markers for generation of antibodies against these agents. Given the results with the VLA-4 inhibitor,[33] it would be prudent to include viral antigens from JC and others on the array to detect patients' immune status. Over time, if the cost of an array profile is compatible with insurance company metrics, this approach could be used to examine patients receiving immunosuppressive therapy for signs of antibody responses to pathogenic agents.

NEXT STEPS

First step, health status: using the protein array to demonstrate a stable Immunoglobulin fingerprint profile for disease-free healthy patients. In addition to the proteins on current array this array should include those antigens most of us should respond against: influenza, polio, tetanus, mumps, diphtheria, whooping cough, pneumococcal polysaccharide, Epstein Barr Virus, JC virus, and some most of us should not be responding against, HCV, HIV, Herpes 6. Using different lots of clinical IVIG compared with disease-free individuals. Demonstrating stable fingerprints over time for one individual and a range of values for a number of normal individuals would also be necessary to establish stability parameters.

Second step, health status: compared with the consistent fingerprint for healthy individuals, demonstrate an altered fingerprint for patients with chronic illness such as RA. Next, or in parallel develop an array containing antigenic determinants involved in the chronic autoimmune disease being examined. Establish how this IgG binding fingerprint is different from the fingerprint for normal individuals. Different individuals with a specific disease such as RA may not have an identical profile but there should be commonalities and the stability of their profile established. Profiling autoantibodies before and after a specific therapy may lead to a segregation of the responders from the nonresponders based upon their profile. Success for one chronic disease would clearly lead to examining other chronic autoimmune diseases as well as cancer. The array could indicate the stage of the disease as well as the success or failure of a therapeutic.

First-step biotherapeutics: The array could demonstrate cross-reactivity of the therapeutic or provide evidence against cross-reactivity. Using existing biotherapeutics and comparing the profile with that of the disease-free patients would provide a base line for examining therapeutic leads. Using the array and comparing the fingerprint of a candidate therapeutic with the fingerprint of existing therapeutic antibodies as well as the profile of normal disease-free patients would allow one to weed out those antibodies that fall outside this "normal" pattern.

Second-step biotherapeutic: A protein array could be used to follow patients once injected with the biotherapeutic. If an antibody therapeutic is able to reverse an autoimmune disease, this should be reflected in a the IgG binding pattern of the patient. Injection of the therapeutic into a normal healthy individual should not cause a significant change in the patient IgG profile. Early detection of a change in the normal profile could indicate a therapeutic associated side effect. Looking at the patient Ig binding to a number of viral epitopes may have alerted clinical investigators that Natalizumab (Tsabri) was preventing the immune system from its normal checkmate of JC virus. If the existing arrays can show the difference between a healthy immune system and a chronically diseased one this should establish value. Linking a good biotherapeutic to one binding profile and a bad biotherapeutic to a different binding profile will also establish value. The costly process of developing new vaccines may be greatly streamlined if one can, by using the array, determine off-target effects in relatively small trials. Parallel use of the array in multiple areas will rapidly lead to improvements. This initial usage of the array, if successful, would lead to the upgrade and the generation of proteins more reflective of the *in vivo* situation. Over time, removing proteins that are never recognized and replacing proteins that may not reflect *in vivo* folded proteins will result in an array able to reflect the system. This may require a number of approaches such as expression of heterodimer proteins and the use of model membranes for the expression of multipass cell or membrane surface proteins. The exciting aspect of such a system is that it has the potential of generating better drugs by telling us which proteins to take forward and then once in the clinic the array could report quickly on who is responding and who is not responding. If a protein therapeutic is not efficacious because of an inappropriate response of the biotherapeutic cross-reacting with another target, the protein array may be able to alert us to this. The antibody fingerprint of a patient may allow not only the early detection of a disease state before serious symptoms but also allow one to stratify the patients into different groups, each responding to a different biotherapeutic.

REFERENCES

1. Carter, P.J., Potent antibody therapeutics by design, *Nat. Rev. Immunol.*, 6, 343, 2006.
2. Pollard, L. and Choy, E., Rheumatoid arthritis: Non-tumor necrosis factor targets, *Curr. Opin. Rheumatol.*, 17, 242, 2005.
3. Puppo, F. et al., Emerging biologic drugs for the treatment of rheumatoid arthritis, *Autoimmun. Rev.*, 4, 537, 2005.
4. Brambell, F.W., The transmission of immunity from mother to young and the catabolism of immunoglobulins, *Lancet*, 2, 1087, 1966.

5. Manz, R.A. et al., Maintenance of serum antibody levels, *Annu. Rev. Immunol.*, 23, 367, 2005.
6. Roopenian, D.C. et al., The MHC class I-like IgG receptor controls perinatal IgG transport, IgG homeostasis, and fate of IgG-Fc-coupled drugs, *J. Immunol.*, 170, 3528, 2003.
7. Yoshida, M. et al., Human neonatal Fc receptor mediates transport of IgG into luminal secretions for delivery of antigens to mucosal dendritic cells, *Immunity*, 20, 769, 2004.
8. Lencer, W.I. and Blumberg, R.S., A passionate kiss, then run: Exocytosis and recycling of IgG by FcRn, *Trends Cell. Biol.*, 15, 5, 2005.
9. Ghetie, V. and Ward, E.S., Transcytosis and catabolism of antibody, *Immunol. Res.*, 25, 97, 2002.
10. Vaccaro, C. et al., Engineering the Fc region of immunoglobulin G to modulate *in vivo* antibody levels, *Nat. Biotechnol.*, 23, 1283, 2005.
11. Simister, N.E., Placental transport of immunoglobulin G, *Vaccine*, 21, 3365, 2003.
12. Bonnick, S.L. and Shulman, L., Monitoring osteoporosis therapy: Bone mineral density, bone turnover markers, or both?, *Am. J. Med.*, 119, S25, 2006.
13. Vermeire, S., Van Assche, G., and Rutgeerts, P., Laboratory markers in IBD: Useful, magic, or unnecessary toys?, *Gut*, 55, 426, 2006.
14. Burczynski, M.E. and Dorner, A.J., Transcriptional profiling of peripheral blood cells in clinical pharmacogenomic studies, *Pharmacogenomics*, 7, 187, 2006.
15. Bild, A.H. et al., Oncogenic pathway signatures in human cancers as a guide to targeted therapies, *Nature*, 439, 353, 2006.
16. Catrina, A.I. et al., Evidence that anti-tumor necrosis factor therapy with both etanercept and infliximab induces apoptosis in macrophages, but not lymphocytes, in rheumatoid arthritis joints: Extended report, *Arthritis Rheum.*, 52, 61, 2005.
17. Eriksson, C. et al., Autoantibody formation in patients with rheumatoid arthritis treated with anti-TNF alpha, *Ann. Rheum. Dis.*, 64, 403, 2005.
18. De Rycke, L. et al., Tumor necrosis factor alpha blockade treatment down-modulates the increased systemic and local expression of Toll-like receptor 2 and Toll-like receptor 4 in spondylarthropathy, *Arthritis Rheum.*, 52, 2146, 2005.
19. De Rycke, L. et al., Infliximab, but not etanercept, induces IgM anti-double-stranded DNA autoantibodies as main antinuclear reactivity: Biologic and clinical implications in autoimmune arthritis, *Arthritis Rheum.*, 52, 2192, 2005.
20. Kruithof, E. et al., Immunomodulatory effects of etanercept on peripheral joint synovitis in the spondylarthropathies, *Arthritis Rheum.*, 52, 3898, 2005.
21. Buch, M.H. et al., C-reactive protein as a predictor of infliximab treatment outcome in patients with rheumatoid arthritis: Defining subtypes of nonresponse and subsequent response to etanercept, *Arthritis Rheum.*, 52, 42, 2005.
22. Aeberli, D. et al., Increase of peripheral CXCR3 positive T lymphocytes upon treatment of RA patients with TNF-alpha inhibitors, *Rheumatology* (Oxford), 44, 172, 2005.
23. Jarand, J. et al., Neurological Complications of Infliximab, *J. Rheumatol.*, 2006.
24. Wadman, M., London's disastrous drug trial has serious side effects for research, *Nature*, 440, 388, 2006.
25. Self, C.H. and Thompson, S., How specific are therapeutic monoclonal antibodies?, *Lancet*, 367, 1038, 2006.
26. Mayor, S., Severe adverse reactions prompt call for trial design changes, *BMJ*, 332, 683, 2006.
27. Mayor, S., Inquiry into adverse events in trial blames drug, not study design, *BMJ*, 332, 870, 2006.

28. Marshall, E., Clinical medicine. Accident prompts a closer look at antibody trials, *Science*, 312, 172, 2006.
29. Goodyear, M., Learning from the TGN1412 trial, *BMJ*, 332, 677, 2006.
30. Hauser, S.L. and Weiner, H.L., Natalizumab: Immune effects and implications for therapy, *Ann. Neurol.*, 59, 731, 2006.
31. Langer-Gould, A. and Steinman, L., Progressive multifocal leukoencephalopathy and multiple sclerosis: Lessons from natalizumab, *Curr. Neurol. Neurosci. Rep.*, 6, 253, 2006.
32. Langer-Gould, A. and Steinman, L., What went wrong in the natalizumab trials?, *Lancet*, 367, 708, 2006.
33. Stuve, O. et al., Immune surveillance in multiple sclerosis patients treated with natalizumab, *Ann. Neurol.*, 59, 743, 2006.
34. Symmons, D.P. and Silman, A.J., The world of biologics, *Lupus*, 15, 122, 2006.
35. Nahar, I.K. et al., Infliximab treatment of rheumatoid arthritis and Crohn's disease, *Ann. Pharmacother.*, 37, 1256, 2003.
36. Listing, J. et al., Infections in patients with rheumatoid arthritis treated with biologic agents, *Arthritis Rheum.*, 52, 3403, 2005.
37. Slamon, D.J. et al., Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2, *N. Engl. J. Med.*, 344, 783, 2001.
38. Vogel, C. et al., First-line, single-agent Herceptin(trastuzumab) in metastatic breast cancer: A preliminary report, *Eur. J. Cancer*, 37 (Suppl. 1), S25, 2001.
39. Romond, E.H. et al., Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer, *N. Engl. J. Med.*, 353, 1673, 2005.
40. Tan-Chiu, E. et al., Assessment of cardiac dysfunction in a randomized trial comparing doxorubicin and cyclophosphamide followed by paclitaxel, with or without trastuzumab as adjuvant therapy in node-positive, human epidermal growth factor receptor 2-over-expressing breast cancer: NSABP B-31, *J. Clin. Oncol.*, 23, 7811, 2005.
41. Levine, M.N., Trastuzumab cardiac side effects: Only time will tell, *J. Clin. Oncol.*, 23, 7775, 2005.
42. Bryant, J. and Geyer, C.E., Trastuzumab for early breast cancer, *Lancet*, 367, 728, 2006.
43. Casadevall, N., What is antibody-mediated pure red cell aplasia (PRCA)?, *Nephrol. Dial. Transplant*, 20 Suppl 4, iv3, 2005.
44. Shinohara, K. et al., Pure red-cell aplasia caused by the antibody to recombinant erythropoietin, epoetin-beta, in a Japanese patient with chronic renal failure, *Am. J. Hematol.*, 78, 15, 2005.
45. Hermeling, S. et al., Reaction to the paper: Interaction of polysorbate 80 with erythropoietin: A case study in protein-surfactant interactions, *Pharm. Res.*, 23, 641, 2006.
46. Casadevall, N., Eckardt, K.U., and Rossert, J., Epoetin-induced autoimmune pure red cell aplasia, *J. Am. Soc. Nephrol.*, 16 (Suppl. 1), S67, 2005.
47. Villalobos, A.P., Gunturi, S.R., and Heavner, G.A., Interaction of polysorbate 80 with erythropoietin: A case study in protein-surfactant interactions, *Pharm. Res.*, 22, 1186, 2005.
48. Rossert, J., Erythropoietin-induced, antibody-mediated pure red cell aplasia, *Eur. J. Clin. Invest.*, 35 Suppl 3, 95, 2005.
49. Ghetie, V. and Ward, E.S., Multiple roles for the major histocompatibility complex class I- related receptor FcRn, *Annu. Rev. Immunol.*, 18, 739, 2000.
50. Lamanna, N. et al., Pentostatin, cyclophosphamide, and rituximab is an active, well-tolerated regimen for patients with previously treated chronic lymphocytic leukemia, *J. Clin. Oncol.*, 24, 1575, 2006.

51. Lamanna, N., Advances in the treatment of chronic lymphocytic leukemia, *Curr. Oncol. Rep.*, 7, 333, 2005.
52. Kashmiri, S.V. et al., SDR grafting — a new approach to antibody humanization, *Methods*, 36, 25, 2005.
53. Hwang, W.Y. et al., Use of human germline genes in a CDR homology-based approach to antibody humanization, *Methods*, 36, 35, 2005.
54. Presta, L.G., Selection, design, and engineering of therapeutic antibodies, *J. Allergy Clin. Immunol.*, 116, 731, 2005.
55. Hwang, W.Y. and Foote, J., Immunogenicity of engineered antibodies, *Methods*, 36, 3, 2005.
56. Rathanaswami, P. et al., Demonstration of an *in vivo* generated sub-picomolar affinity fully human monoclonal antibody to interleukin-8, *Biochem. Biophys. Res. Commun.*, 334, 1004, 2005.
57. Neuberger, M., Generating high-avidity human Mabs in mice, *Nat. Biotechnol.*, 14, 826, 1996.
58. Lonberg, N., Human antibodies from transgenic animals, *Nat. Biotechnol.*, 23, 1117, 2005.
59. Kyewski, B. and Klein, L., A central role for central tolerance, *Annu. Rev. Immunol.*, 24, 571, 2006.
60. Ferry, H. et al., B-cell tolerance, *Transplantation*, 81, 308, 2006.
61. Mamula, M.J., Farber, D.L., and Tsokos, G.C., Autoimmune odyssey on the Aegean Sea, *Nat. Immunol.*, 7, 219, 2006.
62. Lim, P.L. and Zouali, M., Pathogenic autoantibodies: Emerging insights into tissue injury, *Immunol. Lett.*, 103, 17, 2006.
63. Schaffitzel, C. et al., Ribosome display: An *in vitro* method for selection and evolution of antibodies from libraries, *J. Immunol. Methods*, 231, 119, 1999.
64. Hoogenboom, H.R., Selecting and screening recombinant antibody libraries, *Nat. Biotechnol.*, 23, 1105, 2005.
65. Hoet, R.M. et al., Generation of high-affinity human antibodies by combining donor-derived and synthetic complementarity-determining-region diversity, *Nat. Biotechnol.*, 23, 344, 2005.
66. Carmen, S. and Jermutus, L., Concepts in antibody phage display, *Brief Funct. Genomic Proteomic*, 1, 189, 2002.
67. Jermutus, L., Phage Display Technologies — SMi Conference. 23-24 January 2002, London, U.K., *IDrugs*, 5, 203, 2002.
68. Groves, M.A. and Osbourn, J.K., Applications of ribosome display to antibody drug discovery, *Expert Opin. Biol. Ther.*, 5, 125, 2005.
69. Lennard, S., Standard protocols for the construction of scFv libraries, *Methods Mol. Biol.*, 178, 59, 2002.
70. Jostock, T. et al., Rapid generation of functional human IgG antibodies derived from Fab-on-phage display libraries, *J. Immunol. Methods*, 289, 65, 2004.
71. Hanes, J. et al., Ribosome display efficiently selects and evolves high-affinity antibodies *in vitro* from immune libraries, *Proc. Natl. Acad. Sci. USA*, 95, 14130, 1998.
72. Tamaoka, A. et al., Antibodies to amyloid beta protein (A beta) crossreact with glyceraldehyde-3-phosphate dehydrogenase (GAPDH), *Neurobiol. Aging*, 17, 405, 1996.
73. Liu, B. et al., Cross-reactivity of C219 anti-p170(mdr-1) antibody with p185(c-erbB2) in breast cancer cells: Cautions on evaluating p170(mdr-1), *J. Natl. Cancer Inst.*, 89, 1524, 1997.
74. Schuermann, J.P. et al., Structure of an anti-DNA fab complexed with a non-DNA ligand provides insights into cross-reactivity and molecular mimicry, *Proteins*, 57, 269, 2004.

75. Spellerberg, M.B. et al., Dual recognition of lipid A and DNA by human antibodies encoded by the VH4-21 gene: A possible link between infection and lupus, *Hum. Antibodies Hybridomas*, 6, 52, 1995.
76. Otte, L. et al., Molecular basis for the binding polyspecificity of an anti-cholera toxin peptide 3 monoclonal antibody, *J. Mol. Recognit.*, 19, 49, 2006.
77. Xu, Y. et al., Two monoclonal antibodies to precisely the same epitope of type II collagen select non-crossreactive phage clones by phage display: Implications for autoimmunity and molecular mimicry, *Mol. Immunol.*, 41, 411, 2004.
78. Geylis, V. et al., Human monoclonal antibodies against amyloid-beta from healthy adults, *Neurobiol. Aging*, 26, 597, 2005.
79. Czompoly, T. et al., A possible new bridge between innate and adaptive immunity: Are the anti-mitochondrial citrate synthase autoantibodies components of the natural antibody network?, *Mol. Immunol.*, 43, 1761, 2006.
80. Franchini, M. et al., Acquired hemophilia A: A concise review, *Am. J. Hematol.*, 80, 55, 2005.
81. Fleming, S.D. and Tsokos, G.C., Complement, natural antibodies, autoantibodies and tissue injury, *Autoimmun. Rev.*, 5, 89, 2006.
82. Zhang, M. et al., Identification of a specific self-reactive IgM antibody that initiates intestinal ischemia/reperfusion injury, *Proc. Natl. Acad. Sci. USA*, 101, 3886, 2004.
83. Zhu, X. et al., Cofactor-containing antibodies: Crystal structure of the original yellow antibody, *Proc. Natl. Acad. Sci. USA*, 103, 3581, 2006.
84. Nieva, J. et al., Immunoglobulins can utilize riboflavin (Vitamin B2) to activate the antibody-catalyzed water oxidation pathway, *Immunol. Lett.*, 103, 33, 2006.
85. Sendi, P. et al., Intranasal influenza vaccine in a working population, *Clin. Infect. Dis.*, 38, 974, 2004.
86. de Bruijn, I.A. et al., Clinical experience with inactivated, virosomal influenza vaccine, *Vaccine*, 23 (Suppl. 1), S39, 2005.
87. De Serres, G. et al., Oculo-respiratory syndrome after influenza vaccination: Trends over four influenza seasons, *Vaccine*, 23, 3726, 2005.
88. Ebringer, A. and Rashid, T., Rheumatoid arthritis is an autoimmune disease triggered by Proteus urinary tract infection, *Clin. Dev. Immunol.*, 13, 41, 2006.
89. Chiavaroli, C. and Moore, A., An hypothesis to link the opposing immunological effects induced by the bacterial lysate OM-89 in urinary tract infection and rheumatoid arthritis, *BioDrugs*, 20, 141, 2006.
90. Soderberg-Naucler, C., Does cytomegalovirus play a causative role in the development of various inflammatory diseases and cancer?, *J. Intern. Med.*, 259, 219, 2006.
91. Anders, H.J. et al., Molecular mechanisms of autoimmunity triggered by microbial infection, *Arthritis Res. Ther*, 7, 215, 2005.
92. Patole, P.S. et al., Viral double-stranded RNA aggravates lupus nephritis through Toll-like receptor 3 on glomerular mesangial cells and antigen-presenting cells, *J. Am. Soc. Nephrol.*, 16, 1326, 2005.
93. Perosa, F. et al., Generation of biologically active linear and cyclic peptides has revealed a unique fine specificity of rituximab and its possible cross-reactivity with acid sphingomyelinase-like phosphodiesterase 3b precursor, *Blood*, 107, 1070, 2006.
94. Binder, M. et al., The epitope recognized by rituximab, *Blood*, 2006.
95. Polyak, M.J. and Deans, J.P., Alanine-170 and proline-172 are critical determinants for extracellular CD20 epitopes; heterogeneity in the fine specificity of CD20 monoclonal antibodies is defined by additional requirements imposed by both amino acid sequence and quaternary structure, *Blood*, 99, 3256, 2002.

96. Preissner, R. et al., Inverse sequence similarity in proteins and its relation to the three-dimensional fold, *FEBS Lett.*, 414, 425, 1997.
97. Menendez, A. and Scott, J.K., The nature of target-unrelated peptides recovered in the screening of phage-displayed random peptide libraries with antibodies, *Anal. Biochem.*, 336, 145, 2005.
98. Tabrizi, M.A., Tseng, C.M., and Roskos, L.K., Elimination mechanisms of therapeutic monoclonal antibodies, *Drug Discov. Today*, 11, 81, 2006.
99. Rogers, A.B., Cormier, K.S., and Fox, J.G., Thiol-reactive compounds prevent non-specific antibody binding in immunohistochemistry, *Lab. Invest.*, 86, 526, 2006.
100. Michaud, G.A. et al., Analyzing antibody specificity with whole proteome microarrays, *Nat. Biotechnol.*, 21, 1509, 2003.
101. Merkel, J.S. et al., Functional protein microarrays: Just how functional are they?, *Curr. Opin. Biotechnol.*, 16, 447, 2005.
102. Bangham, R. et al., Protein microarray-based screening of antibody specificity, *Methods Mol. Med.*, 114, 173, 2005.
103. Predki, P.F. et al., Protein microarrays: A new tool for profiling antibody cross-reactivity, *Hum. Antibodies*, 14, 7, 2005.
104. Boscolo, S. et al., Detection of anti-brain serum antibodies using a semi-quantitative immunohistological method, *J. Immunol. Methods*, 309, 139, 2006.
105. McIntyre, J.A., Wagenknecht, D.R., and Faulk, W.P., Redox-reactive autoantibodies: Detection and physiological relevance, *Autoimmun. Rev.*, 5, 76, 2006.
106. McIntyre, J.A., Wagenknecht, D.R., and Faulk, W.P., Autoantibodies unmasked by redox reactions, *J. Autoimmun.*, 24, 311, 2005.
107. McIntyre, J.A., The appearance and disappearance of antiphospholipid autoantibodies subsequent to oxidation — reduction reactions, *Thromb. Res.*, 114, 579, 2004.
108. Agrawal, S., Misra, R., and Aggarwal, A., Autoantibodies in rheumatoid arthritis: Association with severity of disease in established RA, *Clin. Rheumatol.*, 2006.
109. Ates, A., Karaaslan, Y., and Aksaray, S., Predictive value of antibodies to cyclic citrullinated peptide in patients with early arthritis, *Clin. Rheumatol.*, 2006.
110. Caspi, D. et al., Synovial fluid levels of anti-cyclic citrullinated peptide antibodies and IgA rheumatoid factor in rheumatoid arthritis, psoriatic arthritis, and osteoarthritis, *Arthritis Rheum.*, 55, 53, 2006.
111. Sihvonen, S. et al., The predictive value of rheumatoid factor isotypes, anti-cyclic citrullinated peptide antibodies, and antineutrophil cytoplasmic antibodies for mortality in patients with rheumatoid arthritis, *J. Rheumatol.*, 32, 2089, 2005.
112. Vander Cruyssen, B. et al., Anti-citrullinated protein/peptide antibodies (ACPA) in rheumatoid arthritis: Specificity and relation with rheumatoid factor, *Autoimmun. Rev.*, 4, 468, 2005.
113. Zhang, M. et al., Identification of the target self-antigens in reperfusion injury, *J Exp Med*, 203, 141, 2006.
114. Westall, F.C., Molecular mimicry or structural mimicry?, *Mol. Immunol.*, 43, 1062, 2006.
115. Etienne, M. and Weimer, L.H., Immune-mediated autonomic neuropathies, *Curr. Neurol. Neurosci. Rep.*, 6, 57, 2006.
116. Hill, P.G. and McMillan, S.A., Anti-tissue transglutaminase antibodies and their role in the investigation of coeliac disease, *Ann. Clin. Biochem.*, 43, 105, 2006.
117. Agrup, C. and Luxon, L.M., Immune-mediated inner-ear disorders in neuro-otology, *Curr. Opin. Neurol.*, 19, 26, 2006.
118. Greidinger, E.L. et al., A murine model of mixed connective tissue disease induced with U1 small nuclear RNP autoantigen, *Arthritis Rheum.*, 54, 661, 2006.

119. Fujii, T., The mechanisms of antinuclear antibody production and its pathogenicity in systemic autoimmune diseases, *Nihon Rinsho Meneki Gakkai Kaishi*, 29, 57, 2006.
120. Ganor, Y. et al., Antibodies to glutamate receptor subtype 3 (GluR3) are found in some patients suffering from epilepsy as the main disease, but not in patients whose epilepsy accompanies antiphospholipid syndrome or Sneddon's syndrome, *Autoimmunity*, 38, 417, 2005.
121. Ganor, Y. et al., Autoimmune epilepsy: Distinct subpopulations of epilepsy patients harbor serum autoantibodies to either glutamate/AMPA receptor GluR3, glutamate/NMDA receptor subunit NR2A or double-stranded DNA, *Epilepsy Res.*, 65, 11, 2005.
122. Husebye, E.S. et al., Autoantibodies to a NR2A peptide of the glutamate/NMDA receptor in sera of patients with systemic lupus erythematosus, *Ann. Rheum. Dis.*, 64, 1210, 2005.
123. Fathman, C.G. et al., An array of possibilities for the study of autoimmunity, *Nature*, 435, 605, 2005.
124. Robinson, W.H., Antigen arrays for antibody profiling, *Curr. Opin. Chem. Biol.*, 10, 67, 2006.
125. Robinson, W.H. et al., Autoantigen microarrays for multiplex characterization of autoantibody responses, *Nat. Med.*, 8, 295, 2002.
126. Wang, X.P. et al., Distinct epitopes for anti-glomerular basement membrane alport alloantibodies and goodpasture autoantibodies within the noncollagenous domain of alpha3(IV) collagen: A janus-faced antigen, *J. Am. Soc. Nephrol.*, 16, 3563, 2005.
127. Wang, X. et al., Autoantibody signatures in prostate cancer, *N. Engl. J. Med.*, 353, 1224, 2005.
128. Du, H. et al., The prevalence of autoantibodies against cartilage intermediate layer protein, YKL-39, osteopontin, and cyclic citrullinated peptide in patients with early-stage knee osteoarthritis: Evidence of a variety of autoimmune processes, *Rheumatol. Int.*, 26, 35, 2005.
129. Trebst, C. and Stangel, M., Promotion of remyelination by immunoglobulins: Implications for the treatment of multiple sclerosis, *Curr. Pharm. Des.*, 12, 241, 2006.
130. Bayary, J. et al., Intravenous immunoglobulin in autoimmune disorders: An insight into the immunoregulatory mechanisms, *Int. Immunopharmacol.*, 6, 528, 2006.
131. Oppenheim, J.J. et al., Autoantigens act as tissue-specific chemoattractants, *J. Leukoc. Biol.*, 77, 854, 2005.
132. Yasuda, S. et al., Pathogenesis of antiphospholipid antibodies: Impairment of fibrinolysis and monocyte activation via the p38 mitogen-activated protein kinase pathway, *Immunobiology*, 210, 775, 2005.
133. Matsubayashi, H. et al., IgG-Antiphospholipid antibodies in follicular fluid of IVF-ET patients are related to low fertilization rate of their oocytes, *Am. J. Reprod. Immunol.*, 55, 341, 2006.
134. Borza, D.B. et al., Goodpasture autoantibodies unmask cryptic epitopes by selectively dissociating autoantigen complexes lacking structural reinforcement: Novel mechanisms for immune privilege and autoimmune pathogenesis, *J. Biol. Chem.*, 280, 27147, 2005.
135. Chan, R.K. et al., Attenuation of skeletal muscle reperfusion injury with intravenous 12 amino acid peptides that bind to pathogenic IgM, *Surgery*, 139, 236, 2006.
136. Howard, O.M. et al., Autoantigens signal through chemokine receptors: Uveitis antigens induce CXCR3- and CXCR5-expressing lymphocytes and immature dendritic cells to migrate, *Blood*, 105, 4207, 2005.

137. Hinagata, J. et al., Oxidized LDL receptor LOX-1 is involved in neointimal hyperplasia after balloon arterial injury in a rat model, *Cardiovasc. Res.*, 69, 263, 2006.
138. Awadallah, S.M. et al., Autoantibodies against oxidized LDL correlate with serum concentrations of ceruloplasmin in patients with cardiovascular disease, *Clin. Chim. Acta*, 365, 330, 2006.
139. Resch, U. et al., Reduction of oxidative stress and modulation of autoantibodies against modified low-density lipoprotein after rosuvastatin therapy, *Br. J. Clin. Pharmacol.*, 61, 262, 2006.
140. Nimmerjahn, F. and Ravetch, J.V., Fcgamma receptors: Old friends and new family members, *Immunity*, 24, 19, 2006.
141. Tsuchiya, N. and Kyogoku, C., Role of Fc gamma receptor IIb polymorphism in the genetic background of systemic lupus erythematosus: Insights from Asia, *Autoimmunity*, 38, 347, 2005.
142. Akilesh, S. et al., The MHC class I-like Fc receptor promotes humorally mediated autoimmune disease, *J. Clin. Invest.*, 113, 1328, 2004.
143. Li, N. et al., Complete FcRn dependence for intravenous Ig therapy in autoimmune skin blistering diseases, *J. Clin. Invest.*, 115, 3440, 2005.
144. Geylis, V. and Steinitz, M., Immunotherapy of Alzheimer's disease (AD): From murine models to anti-amyloid beta (Abeta) human monoclonal antibodies, *Autoimmun. Rev.*, 5, 33, 2006.
145. Tsai, M.H. et al., Clinical responses of patients with Kawasaki disease to different brands of intravenous immunoglobulin, *J. Pediatr.*, 148, 38, 2006.
146. Stiehm, E.R., Lessons from Kawasaki disease: All brands of IVIG are not equal, *J. Pediatr.*, 148, 6, 2006.
147. Robinson, W.H. et al., Protein microarrays guide tolerizing DNA vaccine treatment of autoimmune encephalomyelitis, *Nat. Biotechnol.*, 21, 1033, 2003.
148. Hueber, W. et al., Autoantibody profiling for the study and treatment of autoimmune disease, *Arthritis Res.*, 4, 290, 2002.
149. Balboni, I. et al., Multiplexed protein array platforms for analysis of autoimmune diseases, *Annu. Rev. Immunol.*, 24, 391, 2006.
150. Borza, D.B. et al., Recurrent Goodpasture's disease secondary to a monoclonal IgA1-kappa antibody autoreactive with the alpha1/alpha2 chains of type IV collagen, *Am. J. Kidney Dis.*, 45, 397, 2005.
151. van Gaalen, F. et al., The devil in the details: The emerging role of anticitrulline autoimmunity in rheumatoid arthritis, *J. Immunol.*, 175, 5575, 2005.
152. Huizinga, T.W. et al., Refining the complex rheumatoid arthritis phenotype based on specificity of the HLA-DRB1 shared epitope for antibodies to citrullinated proteins, *Arthritis Rheum.*, 52, 3433, 2005.
153. Lam, A., Toma, W., and Schlesinger, N., Mycobacterium marinum arthritis mimicking rheumatoid arthritis, *J. Rheumatol.*, 33, 817, 2006.
154. Gompels, L.L. et al., Single-blind randomized trial of combination antibiotic therapy in rheumatoid arthritis, *J. Rheumatol.*, 33, 224, 2006.
155. Mimura, Y. et al., Rheumatoid factor isotypes in mixed connective tissue disease, *Clin. Rheumatol.*, 1, 2006.
156. Frohman, E.M., Racke, M.K., and Raine, C.S., Multiple sclerosis — the plaque and its pathogenesis, *N. Engl. J. Med.*, 354, 942, 2006.
157. Kanter, J.L. et al., Lipid microarrays identify key mediators of autoimmune brain inflammation, *Nat. Med.*, 12, 138, 2006.

158. Lalive, P.H. et al., Antibodies to native myelin oligodendrocyte glycoprotein are serologic markers of early inflammation in multiple sclerosis, *Proc. Natl. Acad. Sci. USA*, 103, 2280, 2006.
159. Schlessinger, A. et al., Epitome: Database of structure-inferred antigenic epitopes, *Nucleic Acids Res.*, 34, D777, 2006.
160. Routsias, J.G., Vlachoyiannopoulos, P.G., and Tzioufas, A.G., Autoantibodies to intracellular autoantigens and their B-cell epitopes: Molecular probes to study the autoimmune response, *Crit. Rev. Clin. Lab. Sci.*, 43, 203, 2006.
161. Fontoura, P. and Steinman, L., Nogo in multiple sclerosis: Growing roles of a growth inhibitor, *J. Neurol. Sci.*, 2006.
162. Gianani, R. et al., Initial results of screening of nondiabetic organ donors for expression of islet autoantibodies, *J. Clin. Endocrinol. Metab.*, 91, 1855, 2006.
163. Pietropaolo, M. and Eisenbarth, G.S., Autoantibodies in human diabetes, *Curr. Dir. Autoimmun.*, 4, 252, 2001.
164. Fourlanos, S. et al., A clinical screening tool identifies autoimmune diabetes in adults, *Diabetes Care*, 29, 970, 2006.
165. Nesterova, M. et al., Autoantibody biomarker opens a new gateway for cancer diagnosis, *Biochim. Biophys. Acta*, 1762, 398, 2006.
166. Cho-Chung, Y.S., Autoantibody biomarkers in the detection of cancer, *Biochim. Biophys. Acta*, 1762, 587, 2006.
167. Bradford, T.J., Wang, X., and Chinnaiyan, A.M., Cancer immunomics: Using autoantibody signatures in the early detection of prostate cancer, *Urol. Oncol.*, 24, 237, 2006.
168. Mooney, C.J. et al., Identification of autoantibodies elicited in a patient with prostate cancer presenting as dermatomyositis, *Int. J. Urol.*, 13, 211, 2006.

16 Humoral Response Profiling Using Protein Microarrays

Arun Sreekumar, Barry S. Taylor, Xiaoju Wang, David M. Lubman, and Arul M. Chinnaiyan

CONTENTS

INTRODUCTION

Biomarker discovery has emerged as a major field of research in oncology. Conventional clinical methods for cancer detection involve either testing for single tumor antigens or relying on histopathology of tissue biopsies. While the former is successful and minimally invasive, it is limited in both sensitivity and specificity to the given cancer type. For example, prostate specific antigen (PSA) is being used routinely in clinical practice as a first level of detection for prostate cancer. Nevertheless, since high levels of PSA are found in nonmalignant samples, elevated PSA does not confirm the presence of cancer. Elevated PSA levels found in benign prostatic hyperplasia (BPH), prostatitis, and prostatic intraepithelial neoplasia (PIN) requires follow-up biopsy to rule out the incidence of cancer. However, the needle-biopsy is both invasive and suffers the limitation of frequently missing the physical site of cancer. Thus, it has become incumbent on the community to develop multiplex biomarker panels that in combination can provide a noninvasive platform for sensitive and specific detection of a given cancer.

Protein microarray or "biochip" technologies have the potential to revolutionize the analysis of human cancer. By simultaneously measuring the parallel expression or interaction of thousands of proteins in clinical specimens, a high-dimensional data set can be culled to form a molecular fingerprint of a disease process. Tumor markers are

proteins or substances that correlate with or causally determine malignancy and may represent alterations from a benign state to a neoplastic process. These can be detected in solid tumors, lymph nodes, bone marrow, stool and biofluids including serum and urine. Detecting or monitoring the levels of tumor markers may aid in diagnosis, staging, population screening, prognostic assessment, and assessing response to therapy or for the identification of metastatic or recurrent disease. The absolute presence or relative levels of tumor markers distinguishes benign from malignant states. One of the major confounding factors in identifying viable tumor biomarkers is their low abundance in biofluids compared to other house keeping and high abundance proteins. Thus, the dynamic range of detection methodologies are required to span as much 10^{10} orders of magnitude to reliably detect these markers in complex biofluids like plasma or serum. Unfortunately, none of the existing technologies and platforms offers such a broad dynamic range of detection without using pre-fractionation strategies like protein depletion. The latter could result in the loss or suppression of important biomarkers as many of the high-abundant proteins subject to depletion have been known to act as carriers for low-abundant biomarkers.

An alternate strategy to sift through this molecular noise without performing any of the above steps is to take advantage of the body's own immune mechanism whereby it produces autoantibodies to tumor antigens. The explicit link between cancer and inflammation was proposed as early as 1863 by Virchow. In a clinical setting, this is best typified by chronic inflammation of the gastrointestinal tract and the subsequent increase in colon cancer susceptibility. Further, the presence of an immune response to cancer in humans has been demonstrated by the screening of auto-antibodies against a number of intracellular antigens in patients with various tumor types. This phenomenon is known as the *humoral response* and the detection of such autoantibodies has been shown to be of great diagnostic and prognostic significance in the detection of cancer and the ability to predict the course of disease. For example, it has been shown that somatic alterations in the p53 gene elicit a humoral response in 30 to 40% of affected patients. Additionally, the detection of these anti-p53 antibodies can predate the diagnosis of cancer. In other work, 60% of patients with lung adenocarcinoma exhibited a humoral response to glycosylated annexins I and/or II whereas none of the noncancerous standards exhibited such a response. Similarly, autoantibodies to the proteasome as well as various antigens including prostate specific antigen (PSA), prostatic acid phosphatase (PAP), HER-2/neu, p53, alpha methylacyl-CoA racemase (AMACR) and GRP78 have been observed in the sera of prostate cancer patients. Furthermore, it has been shown that the majority of antigens from tumor cells that elicit this response are not just products of mutated genes. These proteins are often differentiation antigens or other proteins over-expressed or modified in cancer. Interestingly, the majority of results to date demonstrate that, in humoral response trials where there are a large number of patients tested, only a subset of patients with a specific tumor type will develop a response to a specific antigen. The reason for this phenomenon is not yet clear, but a number of factors may be responsible for influencing the humoral response in each type of tumor in each individual patient. Among the possible factors affecting this response is that proteins may become immunogenic after undergoing a post-translational

modification, a process that is subject to variability among tumors of a similar type. The result is that any protein may provide a humoral response for only a limited fraction of a patient population for a particular tumor, suppressing its sensitivity. Rather, several protein targets may be required to detect a tumor with broad coverage for a large number of people. Figure 16.1 represents, in the broadest sense, the flow of events that allow for detection of an immune response induced by tumor-specific antigens or cells in circulation. Diagrammatically speaking, better detection is enjoyed through immune system-driven amplification of the autoantibody response that promises higher sensitivity, specificity and reproducibility over the detection of low abundant biofluid-derived proteomic tumor markers. The genesis of this downstream immune response is seeded in the tumor promoting effects of chronic inflammation.

There have been a number of approaches used to interrogate humoral response in serum samples. These include protein-antigen array-based platforms like phage display or SEREX, two-dimensional electrophoretic platforms or more recently a combination of two-dimensional liquid phase fractionation and protein microarrays. Phage display and the SEREX approaches use recombinant proteins obtained from either phage display libraries. These methods demonstrate excellent sensitivity, which is sufficient for measurement of many clinically important proteins in patient blood and sera samples. Nonetheless, a limitation of these technologies is that they do not take advantage of the numerous post-translational modifications (PTMs) such as phosphorylations, glycosylations, and acetylations that *in vivo* proteins undergo. These PTMs, in addition to playing an important role during the neoplastic process, have been shown to play a major role in the generation of humoral response against various tumor antigens. This heightens the importance of using fractionated cellular proteins as baits to study the autoantibody response. Such methods have included the extraction of proteins from cells using either two-dimensional gel electrophoresis or liquid separation methods. Two-dimensional gel electrophoresis has been used to separate over a thousand individual cellular proteins from tumor tissue or cell lines. The separated proteins are then blotted onto a membrane. Sera from cancer patients are individually screened for antibodies reacting against the separated proteins by Western blot analysis. Proteins that react with sera from these cancer patients are subsequently sequenced and identified by mass spectrometry. This method has been used successfully to detect autoantibodies to annexins I and II in lung cancer, β-tubulin isoforms as tumor antigens in neuroblastomas, and Op18 isoforms in acute lymphocytic leukemia. Although the method allows for identification of autoantibodies in patient sera, it has several drawbacks. These include the lack of reproducibility of 2-D gels and the need for large starting quantities of serum as probes. Moreover, the method is labor intensive and in most cases lacks the sensitivity in identifying low abundance proteins in the cells.

Most of the drawbacks of 2-D gels can be overcome with liquid-phase separation of proteins in two dimensions. This involves separating intact proteins from cell lysates using a combination of chromatofocusing in the first dimension, and non-porous silica reverse-phase high pressure liquid chromatography (RP HPLC) in the second dimension. The result is a two-dimensional separation of proteins from a cell lysate where relatively pure proteins in the liquid phase are obtained. Using this

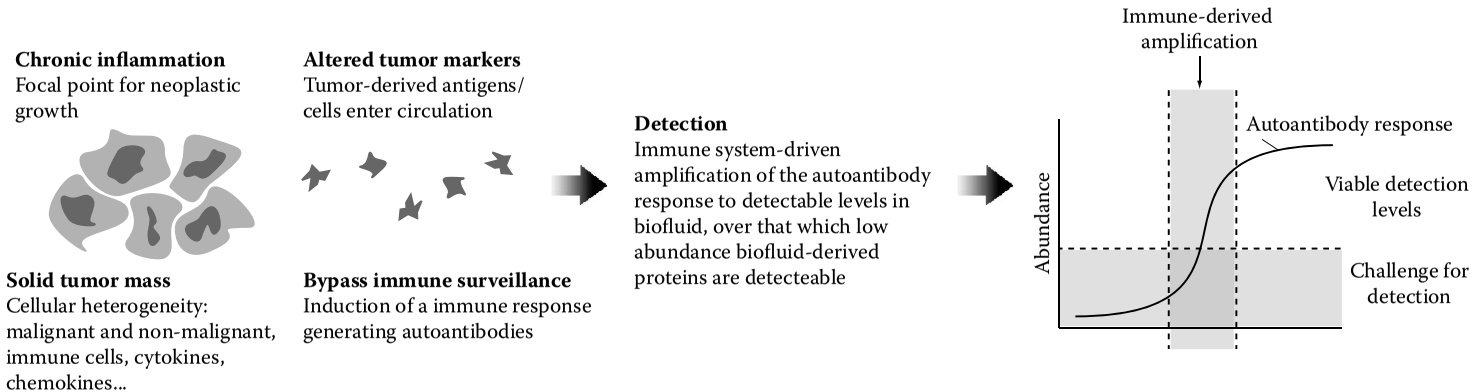


FIGURE 16.1 Schematic representation of autoantibody response detection upon induction by circulating tumor antigens and subsequent immune amplification.

method, hundreds of isolated proteins in the liquid phase can be collected for spotting on a microarray that can be used to interrogate humoral response. This method offers a means for comprehensive proteomic analysis of large numbers of purified proteins as expressed in cancer cells while maintaining their post-translational modifications, which are often critical to the generation of humoral response.

With an eye on the scope of the current chapter, we will discuss two methodologies in greater detail, namely protein fractionation in two-dimensions coupled to protein microarrays and phage-display coupled to protein microarrays.

TWO-DIMENSIONAL LIQUID-PHASE SEPARATION OF THE TUMOR PROTEOME COUPLED TO PROTEIN MICROARRAYS

Figure 16.2 shows the diagrammatic representation of the entire process involving the use of two-dimensional liquid phase fractionation and protein microarrays to interrogate humoral response in serum samples. This involves separating intact proteins from cell or tumor lysates in the first dimension using chromatofocusing based on pI or isoelectric point. Each pI fraction is then separated in a second dimension by nonporous silica reverse-phase high-pressure liquid chromatography (NPS-RP-HPLC). The result is a two-dimensional liquid-phase fractionation of greater than >2500 proteins from a given lysate at relatively high purity. The fractionated proteins can then be spotted on a nitrocellulose slide and used to study humoral response by exposing them to sera from cancer patients and normal controls. This method allows for comprehensive analysis of the cancer proteome using very small amounts of analyte obtained by fractionation.

Interrogating the humoral response involves blocking nonspecific sites, hybridization with serum, and data acquisition. In brief and by example, nitrocelluose slides containing spotted proteins are blocked in PBS containing 1% bovine serum albumin in 0.1% Tween-20 at 4°C overnight. The slides are then incubated with either serum from cancer patients or control individuals (1:400 diluted in probe buffer: containing 50 m*M* PBS, $MgCl_2$, DTT, Triton X100 and 1% BSA) in a hybridization bag for 2 h at 4°C. The hybridized slides are washed six times with probe buffer, each for 5 minutes and incubated with Alexa-647 conjugated anti-human IgG (1:2000, Invitrogen, Carlsbad, CA) for 1 h at 4°C. After washing the slides as above, they are dried and analyzed using a microarray scanner.

This platform has been used to interrogate humoral response to a variety of solid cancers including prostate, lung, colon, and many more. In prostate cancer, Fan et al. identified humoral response to mitochondrial creatine kinase in patients with prostate cancer. Using a similar strategy, Haab et al. identified a humoral response signature in prostate cancer sera. The use of two-dimensional liquid chromatography to fractionate proteins at the front-end offers many advantages. These include reproducibility, ability to start with high amounts of lysates for fractionation, low time-scale, ease of mass spectrometry-based identification of fractionated proteins and the ability to obtain and detect post-translational changes. Further, since the humoral response targets are native tumor-associated proteins, they could play a causative role during tumor development and progression.

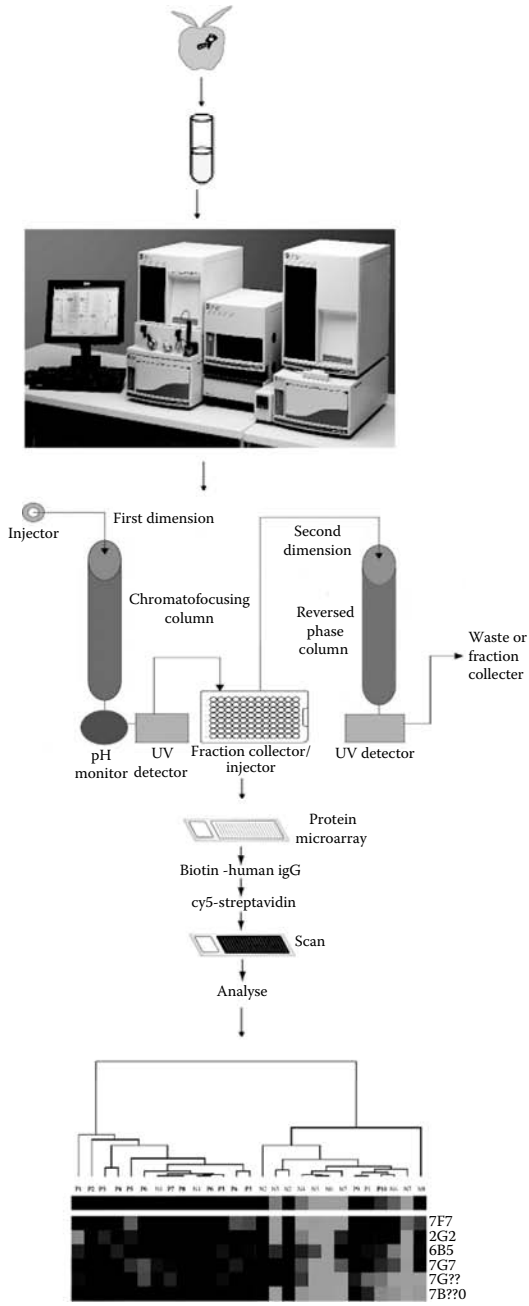


FIGURE 16.2 A workflow of protein fractionation in two dimensions coupled to subsequent protein microarray analysis. This includes lysate preparation, first dimension chromatofocusing, second dimension reverse-phase HPLC, fraction collection, spotting on array platforms and posterior computational analyses.

PHAGE MICROARRAYS TO INTERROGATE EPITOMIC SIGNATURES IN CANCER

Phage microarrays interrogate humoral immune response using phage peptide libraries. These libraries are created initially from tumor-derived mRNA and undergo a process of tumor-specific epitope selection termed *biopanning*. The use of phage display coupled to protein microarrays allows for combinatorial screening and high-throughput analysis of autoantibody repertoires and has lead to an emerging area of research, termed *cancer epitomics*, which allows for the global analysis of autoantibodies against antigens in a neoplasm.

The first reported use of filamentous phage was to display a random oligopeptide on the N-terminus of the viral pIII coat protein by inserting a stretch of random deoxyoligonucleotide into the pIII gene of filamentous phage. Since then this technique has been successfully applied in identifying peptides for various molecular targets. The possibility of displaying amino acid sequence on the surface of filamentous phages has proven to be a valuable tool for the selection of ligands to different targets. In contrast to other cloning strategies, phage display of peptides and proteins is amenable to affinity enrichment. It has been shown that display on the surface of filamentous phages is well-suited for the enrichment of serum antibody-binding ligands.

Procedurally, phage microarrays are constructed from a library of bacteriophage-displaying protein fragments expressed from a randomly fragmented cDNA library. This involves directed synthesis of cDNA from the tumor-derived transcripts using a combination of oligodT and random primers. Fragments of the cDNA library are ligated into phage vectors, such as M13, λ, or T7. The resulting phage library consists of phage that contain random peptide stretches that are fused to the phage envelope protein.

The random library is subsequently enriched for cancer-specific antibody recognition sites, also termed *epitopes*. Epitopes in this context are sequences of amino acids within proteins that react with the antibodies present in human serum. The humoral response in serum could be directed against either a linear stretch of amino acids, naturally termed *linear epitopes*, or toward specific secondary or tertiary conformations of small peptide segments, termed *conformational epitopes*. This process of epitope-selection, or biopanning to which we previously referred, involves iterative and powerful immunoaffinity-based enrichment steps. Said another way, it involves immunoselection of the phage library using cancer sera followed by amplification of the immuno-selected phage in bacteria. The number of selection cycles is a compromise between the necessity to eliminate the majority of background phage and the desire to keep the panning process as short as possible to maintain the ligand diversity of the original library, as well as to avoid enrichment of tighter binders or faster growers. Such an enrichment process allows for high sensitivity during humoral response screening. Further, to achieve greater specificity in detection of cancer-specific immune response, a pre-clearing step can be performed, which includes the removal of phage population that react to antibodies in control individuals.

Prior to humoral response screening, the phage enriched in cancer-specific epitopes are spotted as baits on nitrocellulose-backed slides. These slides are then exposed to serum from either cancer patients or healthy controls and the immunoreactivity is detected using a procedure similar to the one described in the previous section. However, just as in gene expression profiling and "pattern-recognition" serum-proteomics approaches, this method may bear the limitations of significant background signal, sample selection bias, and reproducibility. To mitigate these issues, immunoreactivity for each phage peptide is measured relative to an internal control signal detected by an antibody against phage capsid proteins.

One of the clear advantages of this platform, in addition to being sensitive and specific, is its ability to easily generate bulk quantities of peptides. Further, the expressed peptide can be sequenced at the nucleotide level, and then translated to give the amino acid sequence. This could then be mapped to regions of known proteins using various sequence alignment tools. This allows for the identification of proteins in neoplasm that elicit humoral response. Again, many of these proteins could be dysregulated in the tumor and hence may play a role during development/progression of the cancer. One of the drawbacks of a phage-based humoral response screening method is its insensitivity to post-translational modifications that play a major role during oncogenesis and cancer progression. These modifications could lead to the generation of cancer-specific autoantibody repertoires that would be missed when using this strategy.

Irrespective of this drawback, the epitomic profiling strategy has been widely implemented to study the humoral response profile in various cancers including prostate, breast, and ovarian cancers. In one such study, Wang et al. identified four prostate cancer-specific humoral targets namely BRD2, eIF4G1, RPL13a, and RPL22, all of which were dysregulated in prostate tumors. Similar profiling efforts in ovarian cancer by Chatterjee et al. detected autoantibodies against a number of interesting proteins that included RCAS1, signal recognition protein-19, AHNAK-related sequence, nuclear autoantogenic sperm protein, Nijmegen breakage syndrome 1 (Nibrin), ribosomal protein L4, Homo sapiens KIAA0419 gene product, eukaryotic initiation factor 5A, and casein kinase II.

BIOINFORMATIC APPROACHES TO PROTEIN MICROARRAY DATA

The computational challenges in appropriately mining and analyzing data generated from these platforms are significant, yet tractable. As is best illustrated by the extensive body of work in statistical and bioinformatics approaches to resolve DNA microarray data, there are diverse approaches in the protein microarray domain to issues of normalization, classification, and learning.

Given the variety of protein microarray platforms, the diversity of experimentation and labeling techniques, resulting data can be of many forms and differing quality. Arrays are most frequently in single or two-color format whose enumeration is of spot intensity, which is simply labeling fluorescence. Internal controls, both positive and negative, allow for anchoring and standardization per spot concentration, slide, and across samples. Normalization and selection is most often driven by the

experiment and composition of the arrays and often leverages methods borrowed from, but not identical, to the gene expression community including log-transformations, statistical implementations of the standardized form, and fitting a variety linear regression models. Signal-to-noise issues are of particular importance as they may vary significantly with established standards in other high-throughput environments. Additionally, in the context of serological studies, previously mentioned issues of variability of autoantibody profiles between patients within cohorts are posing a normalization challenge, which negates many of the assumptions that ease this corresponding burden in the DNA microarray domain. Specifically, assumptions of similar aggregate signal between samples and reliable same-spot similarities across patient classes over the majority of spots are often rendered erroneous by this variability. This stresses the need for analytical and computational strategies that absorb the heterogeneity demonstrated across samples of the same class, but concurrently maintains resistance to artifactual entities, whether biological, experimental, or statistical.

Where study motivation is molecular classification, there are a variety of algorithms in statistical and machine learning that have been implemented in class prediction problems on a variety of data types. Solutions derived from these range from relatively transparent and human readable to entirely opaque. Examples include naïve Bayes methods, decision trees, k-nearest neighbor algorithms, variations in discriminant analysis and relatively obfuscated decision functions such as artificial neural networks and support vector machines. While not nearly comprehensive, these algorithms vary between exploiting correlations between data and those treating them as discrete, independent features and will demonstrate differing performance. This is another important aspect of the marriage between the hypothesized and tested biological significance of a given experiment, and the computational method chosen for analysis. However, the choice of learning algorithm is also considered less relevant on ultimate classification and prediction performance than the process by which informative features are selected from the input space. This critical distinction is vital in the immune response profiling domain as significant variability in autoantibody responses of patients to given antigens can challenge the selection of class-dependent, stable, and informative autoantibodies for class prediction. In fact, the issue of statistical and probabilistic models applied to individual markers and their contribution to classification versus association, say with regard to clinical outcome, may differ. This affects its viability for the former and thus the criteria for feature selection that precedes classification.

Critically, to issues of over-fitting and dimensionality, the features on a given protein microarray platform are often orders of magnitude less than the current composition of a typical whole-genome DNA microarray. This is not necessarily for reasons of protein sequence paucity, but rather issues of real estate on a given array and of native protein conformation that speaks directly to the viability of the experiment and the quality of subsequent data. This reduction in dimensionality can be exploited by methods to generate parsimonious models. These are easier to understand, are far more accessible in the context of absolute biological meaning, and incur significantly less computational expense. However, the issue of over-fitting, in either clustering or classification is still a fundamental problem.

Complex and highly parameterized models can fit random variations in training data. However, these relationships do not represent functional effects at the biological level and will not exist in independent data, so the predictive utility of over-fit models is reduced. In the context of these classification problems, and independent of the protein microarray platform used, internal cross-validation as well as external validation with independent cohorts of samples and experiments are necessary to evaluate the performance of a given predictor and increase its analytical rigor. The former is an extensively studied area in the context of statistical learning and a particularly popular approach is that of *N*-fold cross validation. A specific variant of this technique is leave-one out cross validation (LOOCV), used for small sample sizes to measure the fraction of errors over the total number of training examples in a supervised learning situation. LOOCV repeatedly partitions the given data set, removing one sample from the training data, constructing the decision function on the basis of the remaining data and then testing it on the removed example. Its benefits are of disjoint training and test data sets, classifiers being tested on each sample exactly once, and it yields a relatively unbiased estimate of the classifier in question with increasing sample sizes. This can also be used in other contexts, one of which is wrapper-based feature selection when coupled to the learning problem.

The computational and bioinformatic tool set for managing protein microarray data is constantly expanding. As these methods mature in parallel with the high-throughput platforms on which they are applied, the production of viable and robust signatures of autoantibodies for early detection and assessing risk of disease susceptibility may become reality and the goals of clinical applicability met.

ACKNOWLEDGMENT

This work is supported in part by the National Cancer Institute under the grants RO1CA10640Z (DML and AMC) and UO1CA111275(AMC).

REFERENCES

1. Balkwill, F. and Mantovani, A., Inflammation and cancer: Back to Virchow?, *Lancet*, 357, 539, 2001.
2. Balkwill, F. and Coussens, L.M., Cancer: An inflammatory link, *Nature*, 431, 405, 2004.
3. Mintz, P.J. et al., Fingerprinting the circulating repertoire of antibodies from cancer patients, *Nat. Biotechnol.*, 21, 57, 2003.
4. Nilsson, B.O. et al., Autoantibodies to prostasomes as new markers for prostate cancer, *Ups. J. Med. Sci.*, 106, 43, 2001.
5. Yan, F. et al., Protein microarrays using liquid phase fractionation of cell lysates, *Proteomics*, 3, 1228, 2003.
6. Stockert, E. et al., A survey of the humoral immune response of cancer patients to a panel of human tumor antigens, *J. Exp. Med.*, 187, 1349, 1998.

7. Sreekumar, A. et al., Humoral immune response to alpha-methylacyl-CoA racemase and prostate cancer, *J. Natl. Cancer Inst.*, 96, 834, 2004.
8. Soussi, T., p53 Antibodies in the sera of patients with various types of cancer: A review, *Cancer Res.*, 60, 1777, 2000.
9. Brichory, F.M. et al., An immune response manifested by the common occurrence of annexins I and II autoantibodies and high circulating levels of IL-6 in lung cancer, *Proc. Natl. Acad. Sci. USA*, 98, 9824, 2001.
10. McNeel, D.G. et al., Antibody immunity to prostate cancer associated antigens can be detected in the serum of patients with prostate cancer, *J. Urol.*, 164, 1825, 2000.
11. Prasannan, L. et al., Identification of beta-tubulin isoforms as tumor antigens in neuroblastoma, *Clin. Cancer Res.*, 6, 3949, 2000.
12. Melhem, R. et al., Quantitative analysis of Op18 phosphorylation in childhood acute leukemia, *Leukemia*, 11, 1690, 1997.
13. de Wildt, R.M. et al., Antibody arrays for high-throughput screening of antibody-antigen interactions, *Nat. Biotechnol.*, 18, 989, 2000.
14. Zhu, H. et al., Analysis of yeast protein kinases using protein chips, *Nat. Genet.*, 26, 283, 2000.
15. Bussow, K. et al., A method for global protein expression and antibody screening on high-density filters of an arrayed cDNA library, *Nucleic Acids Res.*, 26, 5007, 1998.
16. Chong, B.E. et al., Differential screening and mass mapping of proteins from premalignant and cancer cell lines using nonporous reversed-phase HPLC coupled with mass spectrometric analysis, *Anal. Chem.*, 73, 1219, 2001.
17. Minamoto, T. et al., Distinct pattern of p53 phosphorylation in human tumors, *Oncogene*, 20, 3341, 2001.
18. Bouwman, K. et al., Microarrays of tumor cell derived proteins uncover a distinct pattern of prostate cancer serum immunoreactivity, *Proteomics*, 3, 2200, 2003.
19. Wang, X. et al., Autoantibody signatures in prostate cancer, *N. Engl. J. Med.*, 353, 1224, 2005.
20. Parmley, S.F. and Smith, G.P., Antibody-selectable filamentous fd phage vectors: Affinity purification of target genes, *Gene*, 73, 305, 1988.
21. Mullaney, B.P. and Pallavicini, M.G., Protein-protein interactions in hematology and phage display, *Exp. Hematol.*, 29, 1136, 2001.
22. Ernst, T. et al., Decrease and gain of gene expression are equally discriminatory markers for prostate carcinoma: A gene expression analysis on total and micro-dissected prostate tissue, *Am. J. Pathol.* 160, 2169, 2002.
23. Trepel, M., Arap, W., and Pasqualini, R., *In vivo* phage display and vascular heterogeneity: Implications for targeted medicine, *Curr. Opin. Chem. Biol.*, 6, 399, 2002.
24. Hoess, R.H., Protein design and phage display, *Chem. Rev.*, 101, 3205, 2001.
25. Smith, G.P., Filamentous fusion phage: Novel expression vectors that display cloned antigens on the virion surface, *Science*, 228, 1315, 1985.
26. Scott, J.K. and Smith, G.P., Searching for peptide ligands with an epitope library, *Science*, 249, 386, 1990.
27. Griffiths, A.D. et al., Human anti-self antibodies with high specificity from phage display libraries, *Embo. J.*, 12, 725, 1993.
28. Bonnycastle, L.L. et al., Probing the basis of antibody reactivity with a panel of constrained peptide libraries displayed by filamentous phage, *J. Mol. Biol.*, 258, 747, 1996.
29. Dybwad, A. et al., Identification of new B cell epitopes in the sera of rheumatoid arthritis patients using a random nanopeptide phage library, *Eur. J. Immunol.*, 23, 3189, 1993.

30. Sche, P.P. et al., Display cloning: Functional identification of natural product receptors using cDNA-phage display, *Chem. Biol.*, 6, 707, 1999.
31. Zozulya, S. et al., Mapping signal transduction pathways by phage display, *Nat. Biotechnol.*, 17, 1193, 1999.
32. Ransohoff, D.F., Rules of evidence for cancer molecular-marker discovery and validation, *Nat. Rev. Cancer*, 4, 309, 2004.
33. Fernandez-Madrid, F. et al., Autoantibodies to Annexin XI-A and other autoantigens in the diagnosis of breast cancer, *Cancer Res.*, 64, 5089, 2004.
34. Chatterjee, M. et al., Diagnostic markers of ovarian cancer by high-throughput antigen cloning and detection on arrays, *Cancer Res.*, 66, 1181, 2006.
35. Qiu, J. et al., Development of natural protein microarrays for diagnosing cancer based on an antibody response to tumor antigens, *J. Proteome Res.*, 3, 261, 2004.
36. Pepe, M.S. et al., Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker, *Am. J. Epidemiol.*, 159, 882, 2004.

17 DNA Interactions with Arrayed Proteins

Marina Snapyan and Vehary Sakanyan

CONTENTS

INTRODUCTION

Protein array technology is readily amenable to different binding assays with various types of molecules, including DNA. It has become a powerful tool for genome-scale screening of protein–DNA interactions and for routine measurements of DNA-binding parameters of wild-type and mutant proteins. The advantages and main challenges of this emerging technology are described in this chapter. Ongoing developments in the field of the investigation of DNA–protein interactions and their relevance in biomedical research are discussed.

Protein interactions play a central role in many biological processes and the interplay between proteins and DNA is the most fundamental of molecular interactions. Many regulatory proteins bind to specific nucleotide sequences and affect gene expression via modulation of the transcriptional machinery at all stages of RNA synthesis. DNA replication and maintenance is governed by DNA-binding proteins that associate with the origins of replication, centromeres, telomeres and other regulatory sites.[2] Another well-conserved process assuring the integrity of genetic information is the correction of mismatches generated during DNA synthesis and escaping proofreading. Mismatch repair proteins also participate in many DNA

transactions, such as genome rearrangements by site-specific recombination and the transposition, modification and restriction of DNA.[3,4] A large number of transcription factors control genesis, differentiation, apoptosis, and other vital processes involved in cellular homeostasis of eukaryotes. Therefore, the study of DNA-binding proteins and interactions between proteins and DNA has always been the focus of genetic, biochemical and structural investigations and has been indispensable for biotechnological and biomedical applications.

Numerous methods have been successfully used to study protein–DNA interactions, including fluorescence spectroscopy, nuclear magnetic resonance, mass spectroscopy, surface plasmon resonance (SPR), atomic force microscopy, etc. (see Section 3 of this book). Convenient ways to study simultaneously different protein samples include gel electrophoresis mobility shift assays (EMSA), Southwestern blotting and enzyme-linked immunosorbent assays (ELISA). However, these methods are limited in their analytical capacity as it can take months to assess binding between a large number of target proteins and DNA sequences. The constantly increasing number of sequenced genomes underlines the resurgence of functional and mechanistic studies of large numbers of proteins. Modern biology is at the beginning of this post-genomic era that requires high-throughput and multiplexed proteomic technologies to deduce various networks in living organisms and to use the acquired knowledge for diverse applications.

The recently developed protein array technology is perfectly situated to carry out large-scale screening of molecular interactions, including protein–nucleic acid interactions.[6] The use of a planar microarray format has several benefits over solution-based methods: (a) microspots provide greater sensitivity to detect signals; (b) the experimental control is better since multiple binding assays can be performed under the same conditions for all immobilized molecules; (c) the consumption of samples and reagents is noticeably lower; (d) the response readout is straightforward and rapid.

The scope of this chapter is to survey and categorize the feasibility, current state, applications, and recent advances of protein array technology for the analysis of protein–DNA interactions.

CLASSIFICATION PRINCIPLE OF DNA-BINDING PROTEINS

The diversity of DNA-binding proteins, involved directly or indirectly in numerous regulatory, metabolic and signaling pathways, reflects adaptation mechanisms that have evolved in organisms as a response to various environmental factors. DNA-binding proteins typically harbor two active structural motifs, one recognizes and binds nucleotide sequences while the other provides the assemblage of oligomeric molecules or the recruitment of cofactors and other protein partners. It is common to classify DNA-binding proteins according to the structure of DNA-binding domains and the mechanism of recognition of nucleotide sequences.[7]

A comprehensive analysis of 240 protein–DNA complexes allowed to divide DNA-binding proteins into eight different structural/functional groups (Figure 17.1), which contain at least 54 structural families.[8] The members of 28 families use α

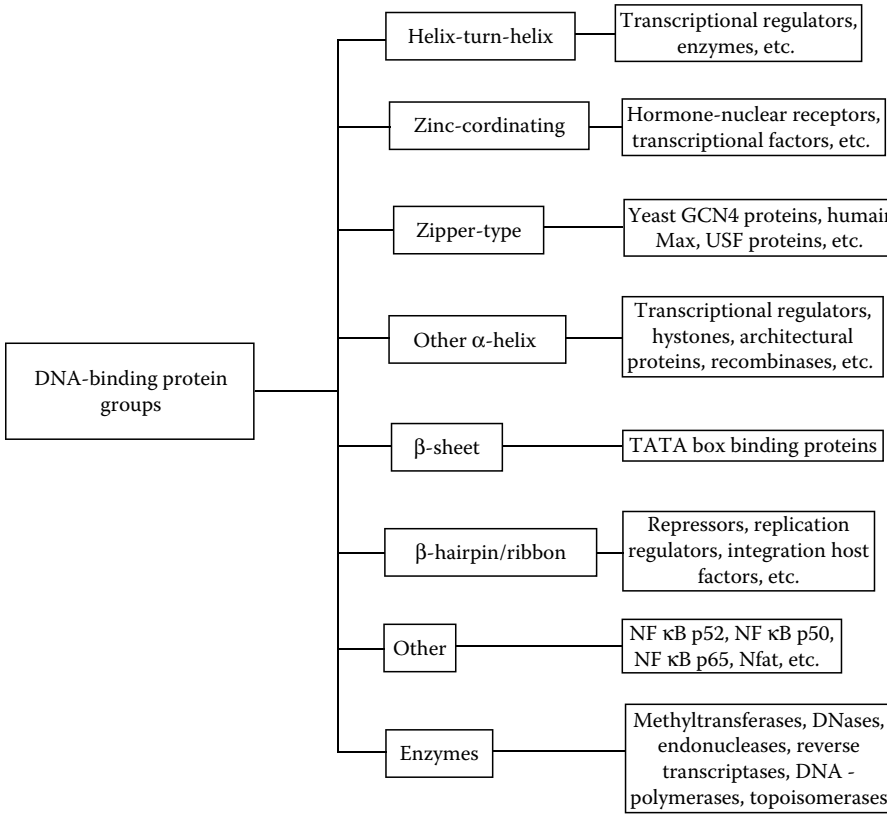


FIGURE 17.1 Structural/functional groups of DNA-binding proteins (from Luscombe, N. M. et al., *Genome Biol.*, 1, 2000).[8]

helices to bind the major groove in DNA and this interaction appears to be the main molecular recognition mechanism that has evolved in both prokaryotes and eukaryotes. The first group is composed of proteins using a helix-turn-helix (HTH) motif (including "winged" HTH) as a common recognition element of a specific nucleotide sequence. This major group of DNA-binding proteins includes most transcriptional regulators and some enzymes of prokaryotes and eukaryotes. The second group includes Zn-coordinating proteins, like transcriptional factors and hormone receptors, encoded mainly by eukaryotic genomes. A DNA-binding motif of these proteins is characterized by the tetrahedral coordination of one or two zinc ions by conserved cysteine and histidine residues. The third group consists of zipper-type proteins, exclusively from eukaryotes, which can be distinguished by the formation of homo- or hetero-dimers through leucine zipper and helix-loop-helix motifs. In such proteins, the DNA-binding segment is a direct extension of the leucine zipper or is separated by a loop from the dimerization region. A small fourth group, named "other α-helix proteins," contains members with very different structures and functions, such as

transcriptional regulators and replication initiators, histones, architectural proteins for DNA packaging or recombinases. All of them use α helices to bind DNA. The proteins of the fifth group use β-strand structures for DNA recognition. Typical representatives, TATA box-binding proteins, are basic components of a transcription initiation complex directed by RNA-polymerase II in eukaryotes. The members of the sixth group use shorter β-sheets or β-hairpin DNA-binding motifs and include proteins with very diverse functions, like transcriptional repressors, replication terminators, integration host factors etc. The proteins of the seventh group, named "Other," do not use well-defined structural motifs for DNA interactions, possess multidomain subunits and bind to DNA as dimers. The eighth group of DNA-binding proteins has been characterized on the basis of their function rather than protein structure. It includes the essential enzymes, like methyltransferases, restriction endonucleases, deoxyribonucleases, mismatch endonucleases, polymerases, reverse transcriptases, and topoisomerases, which affect DNA structure through catalytic reaction.

PREDICTION OF PROTEIN–DNA INTERACTIONS

The increasing number of high-resolution 3-D structures of proteins and protein–DNA complexes has generated a massive quantity of data stored in databases. Computational approaches provide highly useful information in the genome-scale prediction of DNA interaction regions in proteins and regulatory sites in DNA by aligning primary sequences with reference motifs.[9–12] However, sequence alignment approaches can generate false positive hits and provide only limited screening tools. From our own experience, we know that the transcriptional factor ArgR of a thermophilic bacterium, *Bacillus stearothermophilus*, which contains 4 leucines positioned at every eight residues of the α helix,[13] had been considered as a leucine-zipper regulatory protein in a databank. Sequence analysis of the same protein from other strains showed that leucine is substituted by other hydrophobic amino acids.[14] Only the resolved 3-D structure of the full-length protein completely excluded the presence of a zipper in ArgR.[15]

Structure-based bioinformatics estimations provide a deeper insight into protein–DNA interactions than the sequence-based methods. For example, one of these methods uses statistical potentials for amino acid interactions with a given nucleotide sequence.[16] For a protein–DNA complex, the total energy of the structure can be calculated as the sum of all the pairs of amino acid-base interactions. This total energy is then used as a scoring function to represent the fitness of protein sequences with respect to the structure of the complex. Threading protein sequence in the protein–DNA framework can reveal the specificity of the protein–DNA recognition. Sequence-structure threading may also be applied to protein–DNA complexes to predict DNA targets for regulatory proteins.[17]

DNA–protein complexes are dynamic structures that assemble, store, and transduce biological information for many physiological processes. The interruption or misregulation of protein–DNA interactions can lead to severe diseases. Therefore, the elucidation of when, why, how and which protein activates or represses gene expression is essential for understanding of the complexity of biological systems. Although sequence-based algorithms and structure-based informatics models are in

constant progress to develop fast and reliable predictions, these virtual approaches alone cannot be considered as proof of protein–DNA interaction-based functions and should be completed by experimental methods.

HIGH-THROUGHPUT PROBING DNA–PROTEIN INTERACTIONS

For a long time, several relatively simple methodologies, such as EMSA, ELISA, and the nitrocellulose filter-binding method, have been widely used to detect and characterize protein–DNA interactions. All these methods suffer from major drawbacks; they are relatively time-consuming and low-throughput.

In recent years, however, a number of high-throughput technologies have been developed to study protein–DNA interactions. Notably, the ChIP-chip method, a combination of chromatin immunoprecipitation (ChIP) and DNA microarray (chip), has enabled a genome-scale location of DNA-binding proteins.[18] In this technique, cells are treated with a cross-linking reagent to covalently link protein complexes *in situ* to DNA. The cross-linked chromatin is then isolated and fragmented and the protein–DNA complexes are precipitated with an antibody against the protein of interest. To identify the selected DNA fragments, the cross-links are reversed and the precipitated DNA is revealed by hybridization to a DNA microarray. Several thousands of interactions with more than 100 preselected transcription factors have been detected in the *Saccharomyces cerevisiae* genome with the ChIP-chip.[19–21] A similar approach uses the separation of bound protein–DNA complexes by EMSA instead of immunoprecipitation.[22]

The method has been improved and its unbiased version, sequence tag analysis of genomic enrichment (STAGE),[23] uses the advantage of the previously developed SAGE.[24] The genomic fragments, enriched by ChIP, are first amplified by PCR using biotinylated degenerate primers and digested by tetranucleotide recognition restriction endonuclease.[23] The biotinylated DNA fragments are captured with streptavidin beads and connected to linkers containing a digestion site for IIS type restriction enzyme, providing a release of 21-bp tags from DNA fragments. These DNA fragments are concatemerized by ligation, cloned and sequenced. STAGE has been used to identify chromosomal targets of the TATA-box binding protein TBP in the yeast genome. The method has also identified new targets for the transcriptional factor E2F4 in human cells.

The DNA adenine methylase identification (DamID-array) technique takes into consideration the fact that eukaryotic DNA can be subjected in cells to methylation only at chromosomal sites bound to a given transcriptional factor fused to adenine methyltransferase.[25,26] To identify such sites *in vivo*, the methylated regions are purified or selectively amplified from genomic DNA, labeled by fluorophore and then hybridized to a DNA microarray. However, in DamID-array, the fusion enzyme methylates adenine up to 2 kb from its binding site, which limits the mapping resolution.

In another method, single-stranded oligonucleotides are enzymatically converted *in vitro* into double-stranded DNA substrates to generate a double-stranded DNA (dsDNA) array able to bind transcriptional factors.[27] Such dsDNA targets have been used

to determine the binding specificity of zinc finger proteins[28] and adapted to detect single nucleotide polymorphism in human transcriptional factors NF-kB and OCT by SPR.[29]

Recently, the improved version of dsDNA array, named protein-binding microarray (PBM) has been proposed.[30] It is based on spotting large numbers of various intergenic dsDNA regions, carrying potential regulatory sites, which, if they bind to tagged transcriptional factors (individual reaction with each protein), can be detected with a fluorophore-conjugated antibody specific for the tag used. This method has detected binding sites specific for transcriptional factors Abf1, Rap1 and Mig1, including new targets located upstream of previously uncharacterized ORFs in the yeast genome. The identified *in vitro* binding sites correlate with *in vivo* sites detected by ChIP-chip. However, PBM appears to be particularly useful when enrichment of bound DNA fragments is not sufficient by ChIP. Another advantage of PBM is the absence of the cultivation of cells, in order to express the transcription factor of interest, as required by *in vivo* approaches described. Though PBM needs extensive bio-informatics support, its rapidity and relative simplicity are attractive for applying it to the location of protein-binding sites in other genomes.

The phage display approach also provides a high-throughput identification and characterization of protein–DNA interactions from large libraries. Active enzyme variants can be selected by conversion of the phage-linked substrate to product via the selection of the reacting phage particle by affinity chromatography.[31] In particular, DNA polymerase derivatives with improved catalytic properties, or even exhibiting RNA polymerase activity, have been selected by linking a DNA primer to the phage coat, which contained numerous variations of the target generated by directed evolution.[32, 33]

Although the high-throughput methods listed above have yielded very important data for understanding cellular regulation, they cannot realistically be applied to the study of protein functions in entire tissues or organisms. To assign functions on a broader scale, we must turn to miniaturized protein arrays that can test protein activities in a highly parallel format. While the DNA array-based methods allow regulatory protein-binding sites to be identified in the genome, protein arrays may have the greatest potential for providing direct information about protein functions and can also become a powerful tool for screening protein–DNA interactions, including functional assays of open reading frames (ORFs) of unknown function at a genome-scale level. Furthermore, the integration of information obtained from protein microarrays and other high-throughput methods may allow the construction of complete relational databases for metabolic and signal transduction pathways in organisms. Protein-array based assays are inherently scalable and easily adaptable to automation with low sample consumption. In fact, with the currently available technologies allowing high-throughput *in vivo* and *in vitro* protein production, sufficient material for printing a large number of protein arrays can be provided.[34]

RATIONAL FOR FABRICATION OF PROTEIN ARRAYS TO STUDY PROTEIN–DNA INTERACTIONS

The vast complexity of proteins, compared to nucleic acids, in terms of physicochemical property diversity, including post-translationally attached sugars, phosphates, and other active groups, requires a greater degree of sophistication in both

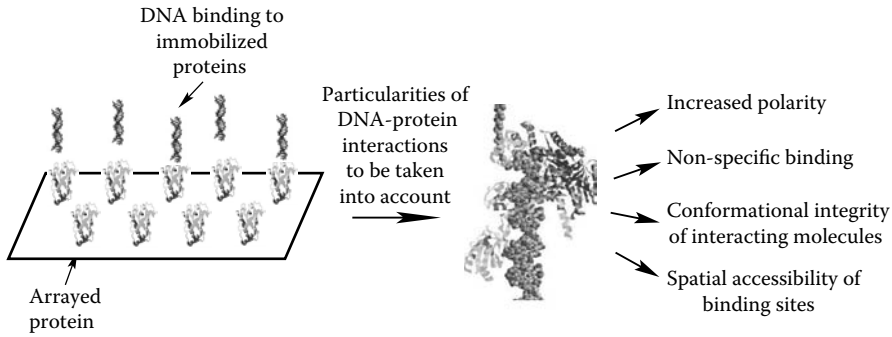


FIGURE 17.2 Major factors affecting interactions between immobilized proteins and labeled DNA probes in solution.

protein array design and data analysis. The fabrication of protein arrays designed specifically to study protein interactions with nucleic acids needs to surpass the same technical challenges as interactions with other types of molecules. The main issues are providing a functional state and binding specificity of proteins after immobilization, increasing the stability of arrays, improving the sensitivity of signal detection, and developing "low-noise" software for data analysis. In addition, the fabrication methods of protein arrays designed to explore protein binding to DNA have to take into consideration the specific features of these intermolecular interactions (Figure 17.2).

The energetics and mode of protein–DNA interactions differ from those of protein–protein and other intermolecular interactions.[35] The main differences concern the polarity and the charge of interactions, since protein–DNA interfaces comprise above average polar and positively charged amino acids than protein–protein interfaces. The most important interactions in protein–DNA complexes are the van der Waals contacts, H-bonds, and water-mediated contacts.[36] The latter play an important role in both the specificity and the affinity of protein–DNA interactions, acting as contact mediators and space-fillers.[37] Thus, the overall polar nature and the charge of protein–DNA interfaces are essential criteria in the choice of the immobilization support and the composition of a binding buffer.

DNA-binding transcriptional factors can recognize target sites in DNA by specific or nonspecific mechanisms.[38] Proteins with similar folds dock in similar ways.[39] However, essential differences can be observed between proteins, which possess structural motifs even if they align perfectly.[40] Direct and specific DNA recognition occurs between the amino acid side-chains and individual bases. Nonspecific readout occurs with the sugar-phosphate backbone of DNA and the protein appears to recognize general structural and conformational features of DNA. Amino acids that interact with the DNA backbone are well conserved and the contacts orient the protein in space, enabling contact with the DNA base edges.[36] The formation of DNA–protein complexes proceeds through changes in topology of both protein and DNA molecules, which possess enough flexibility to adopt distinct conformations.[41] The helical DNA structure is often distorted; the helix is unstaked and unwinged

when bound to the proteins. Conformational parameters, such as shift, slide, twist, rise, roll and tilt, are also modulated. The flexible side chains of a protein can be rearranged upon complex formation in order to achieve complementarity.[35]

Thus, both specific interactions and variations in protein and DNA structure and flexibility determine selective binding to a particular site or similar sites in DNA. Nonspecific protein–DNA contacts are important for the overall stability of protein–DNA complexes. However, these interactions can become a source of false positive signals when DNA–protein interactions are probed with protein arrays. Therefore, when preparing protein arrays and performing DNA interactions, particular care has to be taken to protect both the 3-D protein pattern and DNA integrity. The immobilization method should expose the attached proteins to the probed DNA. Good access of a protein interface to DNA can be achieved by the oriented immobilization of tagged molecules to the functionalized support (see section 2 of this book). Alternatively, using 3-D nitrocellulose or gel supports, which also ensure a better functional stability of immobilized molecules, can substantially increase a total protein surface, exposed for interactions with spotted protein samples. Modifications in the composition of blocking and binding buffers, changes in the binding and washing conditions (stringency, temperature and binding duration) can discriminate undesirable effects and improve the binding specificity thereby decreasing false positive hits. The uniformity of labeled DNA probes can be achieved by chemical synthesis (short dsDNA can be annealed from two ssDNAs) or by PCR (for longer dsDNA) using oligonucleotides usually marked at the 5' extremity by a fluorophore of interest. DNA probes can also be biotinylated for chemiluminescent detection or labeled by ^{32}P for radioactive detection. The length of the DNA should be taken into consideration for probing the binding specificity of regulatory sites.

FUNCTIONAL DISSECTION OF PROTEIN–DNA INTERACTIONS WITH PROTEIN ARRAYS

Pioneering work to screen DNA–protein interactions with protein macroarrays was performed by Ge.[42] Individual protein samples, including general and specialized transcriptional factors, were dot-blotted on a nitrocellulose membrane and probed with ^{32}P-labeled double- or single-stranded oligonucleotides, containing appropriate binding sites. High-intensity signals were detected from the spots of corresponding proteins reacted with these probes. In particular, a phosphorylated form of transcriptional activator PC4 bound a target dsDNA whereas a single amino acid substitution in this protein completely abolished this binding ability. This study also demonstrated the usefulness of protein array technology for simultaneous analysis of multiple interactions in parallel assays with various molecular probes.

Independently, we proposed another concept of detection of molecular interactions with protein arrays based on the use of near-infrared fluorescent dyes (IRDyes) and described in detail elsewhere.[34] Considering that the intrinsic fluorescence of proteins is high with visible range fluorophores, we used IRDyes with wavelengths 700 nm and 800 nm, which provide a very low critical threshold of the signal to be detected. Indeed, this innovation significantly increased the detection sensitivity and improved the performance of protein arrays fabricated on nitrocellulose membranes.

Protein arrays combined with IRDye fluorescence detection were first used to study transcriptional regulation in mesophilic and thermophilic bacteria.[43,44] In these studies, the ArgR paradigm was chosen, given that valuable structural and functional information is available for the *Escherichia coli* repressor.[45] Moreover, the greater stability of the thermostable ArgR of *B. stearothermophilus* with a resolved structure[15] was a good support to establish the optimal binding conditions with the arrayed samples. The ArgR protein consists of an N-terminal domain that contains a wHTH motif for binding to two adjacent 18-bp boxes in an operator sequence, and a C-terminal oligomerization domain that also binds L-arginine as co-repressor.

To characterize the protein DNA-binding specificity on microarrays, domain- and linker-replaced chimeras, comprising various regions of *E. coli* and *B. stearothermophilus* ArgR proteins, were constructed and spotted on nitrocellulose-coated slides.[43] Arrayed His-tagged proteins were probed with DNAs, carrying single or double Arg boxes, in the presence or absence of arginine (Figure 17.3). It was revealed that the DNA-binding affinity for the operator sequence in the presence of L-arginine depends on the source of the oligomerization domain. In contrast to *E. coli,* the *B. stearothermophilus* ArgR protein showed less arginine-dependent binding to a double-box operator. The differential binding response from arrayed proteins correlates well with the binding affinity of ArgR chimeras, as determined by EMSA and SPR. Extension of the microarray methodology to a wild-type and mutant ArgR of *Thermotoga neapolitana* indicated that arginine-independent recognition and/or binding to the *argRo* box elements is characteristic of the regulatory proteins of thermophiles.[44] Thus, multiplexed monitoring of DNA-binding specificity on protein microarrays contributed substantially to understanding the ArgR differential action on bacterial *arg* gene expression and to allocating the protein a position within an evolutionary pathway of transcription regulation. It is plausible that ArgR exhibits low repression and weak operator-binding specificity (interaction with a single *argRo* box in the absence of arginine) in ancestral thermophilic bacteria, whereas it strongly represses transcription of *arg* genes by an arginine-dependent binding mechanism to a double-box operator in mesophilic bacteria.[43,44] Such protein behavior supports the hypothesis that ArgR evolved from a global transcriptional regulator in ancestral bacteria to a highly specialized repressor in enterobacteria.[45]

In bacteria, the RNA polymerase alpha subunit (αRNAP) determines the promoter strength via recognition of a UP-element, an 18- to 20-bp AT-rich sequence located upstream of a –35 site.[46] Surprisingly, our attempt to detect bacterial αRNAP binding to the UP-element failed with EMSA. Moreover, we could not find a reliable indication in literature that EMSA was successfully used to study αRNAP interactions with target DNAs. It appears that the bound protein complex is dissociated during electrophoretic migration in gel. Therefore, it was attractive to develop the protein array method for identification of unstable protein–DNA complexes. In the "proof-of-concept" study, a strong fluorescent signal was detected from spots probed with a DNA fragment carrying a full-length UP-element of *B. stearothermophilus*.[47] Moreover, a clear reduction in the signal intensity was observed from spots when G>A substitutions or AT-deletion were introduced into the UP-element of the probed DNA. This difference was confirmed by cell-free protein synthesis when a reporter gene was transcribed from mutant or nonmutant promoters. In addition,

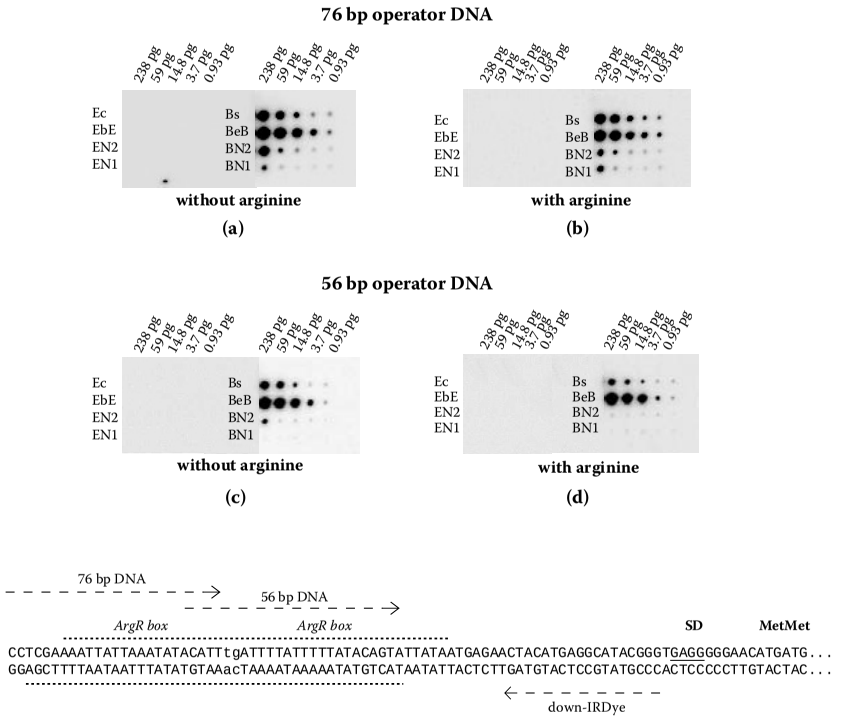


FIGURE 17.3 Detection of protein–DNA interactions with protein microarrays. Wild-type *E. coli* (Ec) and *B. stearothermophilus* (Bs) ArgRs and their derivatives were serially 4-fold diluted and spotted on a nitrocellulose membrane and probed to bind to a *B. stearothermophilus* P*argCO* DNA. Binding reactions were carried out with equal quantities of IRDye800-labeled 76-bp or 56-bp DNA in the presence or absence of L-arginine. Dashed arrows show the positions of oligonucleotide primers for generation of DNA probes. Dotted lines indicate the extent of the DNase footprints for the *B. stearothermophilus* ArgR on both DNA strands; spacer nucleotides between two Arg boxes are shown in small letters. ArgR proteins carrying the DNA-binding motif of the *E. coli* repressor are not able to recognize the used DNA sequence.

a real-time determination of αRNAP binding constants to DNA with SPR correlated well with the microarray data. Protein microarrays were also found to be rather sensitive for detecting the effect of amino acid substitution in *T. maritima* αRNAP on the recognition of a potential UP-element.[48]

It is noteworthy that it is possible to assess both protein–DNA and protein–protein interactions directly in spotted cell extracts, bypassing a long stage of protein purification (Figure 17.4).[47] Moreover, a gradual increase in the intensity of fluorescent

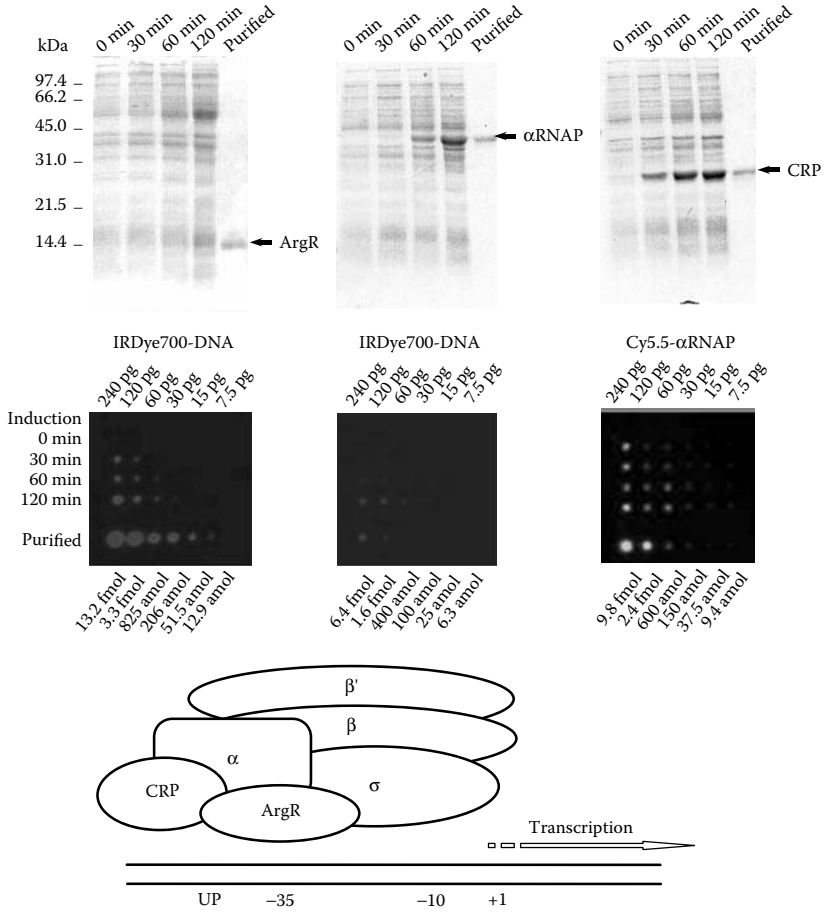


FIGURE 17.4 SDS-PAGE analysis and detection of protein–DNA and protein–protein interactions with arrayed cell extracts. Arrays were prepared with the cell extracts by a serial twofold dilution and with pure proteins by a serial fourfold dilution. Total protein in spotted cell extracts is shown in pg (top of the slides), the amount of spotted pure protein is shown in fmol and amol (bottom of the slides). Binding reactions were carried out with a 76-bp IRDye 700-labeled DNA of *B. stearothermophilus* P*argCo* region or Cy5-5-labeled RNA polymerase of *E. coli*. A possible RNA polymerase/transcriptional factor complex governing the P*argCo* region is shown below.

signals was observed from arrayed crude extracts as a function of the duration of IPTG induction of cells to express the cloned genes coding for ArgR, αRNAP or CRP. This is an indication of the specificity of binding to the probes used that has been confirmed with spots of purified protein in the same assay. In concept, this approach is reminiscent of reverse phase arrays for the evaluation of the expression level of proteins in total lysates, using antibody generated against the target protein.[49] However, the detection of DNA interactions of nonpurified proteins in cell extracts requires a real functional state of non-denatured molecules in cell extracts whereas only the accessibility of a corresponding epitope in the denatured protein is required to detect a signal in antigen-antibody interactions. The developed approach shows a general way for studying the complexity of protein–DNA and protein–protein interactions in relation with a bacterial transcriptional machinery.[47] An example of a possible interactome module, governing gene regulation from the *B. stearothermophilus* P*argCo* promoter-operator region, is shown in Figure 17.4.

Hence, by combining the advantages of immobilized proteins on a nitrocellulose membrane and IRDye-based detection, the protein arrays offer an alternative to the EMSA miniaturized tool for routine determination of the DNA-binding ability of proteins.

Next, a coupled transcription-translation system, which provides an enhanced yield and stability of target mRNAs,[50] was applied to fabricate macroarrays using proteins with unknown functions. The DNA fragments, amplified by PCR from the *T. maritima* genome and coding for putative proteins belonging to XylR, LacI, and GntR families of transcriptional factors, were used directly as templates for protein synthesis. The proteins, partially purified by heat treatment, were arrayed and probed with DNAs carrying well-characterized operator sequences from the *E. coli* genome.[51] Binding was detected from some spotted proteins and several interactions were confirmed by EMSA.

Kersten and coworkers have also shown a functional usefulness of protein arrays for studying the *E. coli* DNA-binding protein DnaA, which initiates bacterial replication with its cognate *oriC* composed of several DnaA boxes.[52] Arrays generated by spotting the wild-type and mutant DNA-binding domain proteins were probed with several Cy5-labeled DNA targets, representing high-affinity R4 and low-affinity R3 DnaA boxes. Ultraviolet cross-linking and mass spectrometry were then applied to localize a DNA-binding site in the cross-linked protein.

The power of protein microarrays to identify previously unrecognized regulatory DNA-binding proteins has been recently demonstrated in Snyder's laboratory.[53] Protein microarrays, covering almost the whole yeast proteome, were probed with single- or double-stranded genomic DNA labeled with Cy3. A total of more than 200 DNA-binding proteins were identified; however, only half of them were known or expected to bind DNA. Eight proteins from unrecognized DNA-binding candidates were subjected to ChIP-chip analysis, which revealed the DNA fragments immunoprecipitated *in vivo*. The Arg5,6 mitochondrial enzyme, which is autocleaved into N-terminal *N*-acetyl-gamma-glutamyl phosphate reductase and C-terminal acetylglutamate kinase involved in arginine biosynthesis,[54] was able to bind several mitochondrial and nuclear DNA regions. Moreover, deletion of the *arg5,6* gene altered transcription levels of both mitochondrial and nuclear target genes, further indicating the role of this biosynthesis-specific enzyme in the regulation of genes implicated in transcriptional and post-transcriptional processes.

Altogether, these data demonstrate the practicability of protein microarrays to analyze simultaneously protein–DNA interactions of purified or nonpurified proteins in macro- and micro-formats.

QUANTIFICATION OF DNA-BINDING CONSTANTS WITH PROTEIN ARRAYS

Binding between two molecules when one is immobilized as a minispot on a solid phase and the other is in a solution appears to be similar to that when both compounds are in the solution phase. However, interactions between various proteins and a single DNA probe on arrays, where the reaction conditions are common for all the spotted proteins, cannot reflect the same binding kinetics in the solution-phase reaction, where the conditions are optimized for each interacting couple. Microarrays provide the perfect opportunity to control simultaneously the experimental conditions for many parallel reactions. Indeed, using a series of known concentrations of interacting partners or buffer compounds, or including the co-factors or inhibitors in reactions can facilitate the quantitative characterization of protein–DNA interactions.

For quantitative measurements of protein–DNA interactions, several factors have to be taken into account to exclude those effects arising from variations in methodological approaches and biological samples. The use of positive and negative binding controls, such as nonbinding proteins or "empty" spots on supports, allows the normalization of results. Local variations in minispots, related to sampling deviation using contacting pins or unequal diffusion of the spotted proteins, cause serious fluctuations in the binding signal, thereby biasing biologically significant information. Currently, there are good quality software packages for image quantification designed to correct these variations and to assist in the high-quality, reproducible measurement of signal intensities.[55] Almost all of them adjust signal distribution to a comparable range, performing the background correction or using spot-quality assessment and trimming.

In an early study, a computational comparison of the signal intensity from eight wild-type and domain- and linker-replaced chimeras of ArgR, bound to two different lengths of DNA probes in the presence and absence of arginine, established their possible order according to their binding affinity (see Figure 17.3).[43] A similar approach was used to assess the effect of mutations in a nucleotide sequence specifically recognized by the human serum response factor.[56] A 16-fold decrease in binding to a mutant DNA compared to a nonmutant DNA target was detected, which was confirmed by solution phase approaches.

The concentration dependence of binding properties was quantified by comparing low- and high-affinity Dna boxes with respect to the DNA-binding domain of DnaA.[52] To control the amount of immobilized His-tagged proteins, the arrayed proteins were additionally tested with an anti-His-tag antibody, followed by detection with a Cy3-labeled secondary antibody. Spot intensity was analyzed with GenePix-Pro 4.0 software and the spot intensity background-subtracted values were used to calculate average intensity values. The signal intensity increased with rising specific DNA probe concentration and only at the highest DNA concentration was unspecific signal detected from the negative control.

Boutell and coworkers developed a functional microarray composed of a p53 wild-type and 49 mutant proteins for a deeper quantitative analysis of protein–DNA interactions, including calculation of both affinity (K_d) and relative maximum binding (B_{max}) values.[52] Microarrays were fabricated by spotting proteins fused to His-tag and a portion of the *E. coli* biotin carboxyl carrier protein onto streptavidin-derivatized, phosphocellulose membrane or neutravidin-derivatized, dextran-coated slides. His-tag was used to evaluate the amount of spotted proteins with the corresponding antibody. Binding assays were performed with ^{33}P- or Cy3-labeled DNA, carrying the GADD45 promoter element, at varying concentrations. The data were normalized against a calibration curve and backgrounds were subtracted. Four replicate values for each arrayed protein at each DNA concentration were fitted to simple hyperbolic concentration-response curves $R = B_{max}/((K_d/L) + 1)$, where R is the response in relative counts and L is the DNA concentration in nM. Thus, replicate values for all mutants were plotted and analyzed by nonlinear regression statistical analysis, enabling calculation of both K_d and B_{max} for wild-type and mutant proteins. This quantitative analysis allowed the functional classification of mutants into groups according to DNA-binding criteria: a group with wild-type affinity (K_d values near to 7), one with reduced stability (low B_{max}) and a group with complete loss of activity. The proteins with mutations outside the DNA-binding domain generally had near wild-type activity, whereas truncated mutants or oligomerization-deficient proteins showed total loss of binding. It is worth noting that a difference was observed when data from proteins spotted onto phosphocellulose or dextran surfaces were compared. In fact, the p53 mutant proteins with impaired DNA binding on phosphocellulose showed no DNA binding on the dextran surface, whereas mutants with an activity similar to that of wild-type p53 showed strong binding ability. This suggests that the phosphocellulose support might stabilize labile mutant proteins thereby enabling affinity measurements to be made. Thus, the choice of an adequate immobilization surface for the fabrication of protein arrays can provide better quantitative measurement conditions.

Although the use of protein microarrays in the quantitative analysis of protein–DNA interactions is not yet widely exploited, the studies mentioned above demonstrate their feasibility for obtaining accurate binding data.

CLINICAL VALUE OF MONITORING PROTEIN–DNA INTERACTIONS WITH PROTEIN ARRAYS

The potential of array technology to screen simultaneously hundreds and thousands of variations of molecular interactions in small amounts of the clinical patterns transforms it into a versatile platform for biomedical applications.[58] A recent advance in this direction, the analysis of a variety of p53 mutants,[57] is promising in terms of the development of similar arrays for other individual proteins and their mutants or families of proteins of clinical and biotechnological significance. The majority of mutations included in the panel of the p53 microarray are located within the DNA-binding domain and result in an autosomal dominant disorder, such as soft-tissue sarcoma, leukemia, osteosarcoma, breast or brain tumors, and adrenocortical carcinoma.

These arrays can be used to detect the effect of mutations and changes in post-translational modifications of proteins on DNA-binding ability.

The inactivation of the p53 oncoprotein is detected in 50% of human cancers and more than 17,000 somatic and germline sequence mutations have been described.[59] Some p53 mutant proteins conserve the transcription activation capacity and affect a set of genes that is different from those controlled by the wild-type protein, which specifically recognizes consensus sequences in DNA.[60] However, the lack of common features, such as the presence of sequence-specific motifs in DNA regions recognized by such mutants, posed a major difficulty in the elucidation of the factors determining the mutant protein interactions with DNA targets. The absence of sequence similarity between cognate binding sites of p53 mutants suggested that the general mode, determining DNA binding by mutant proteins, occurs by nonspecific recognition of structural or conformational features of DNA, rather than by a sequence-specific mechanism.

To investigate whether DNA topology is a relevant parameter for the binding of mutant proteins, Göhler and coworkers evaluated the ability of p53 mutant proteins to bind to distinct conformational DNA forms.[61] Different DNA sequences, with or without p53-binding consensus and exhibiting stem-loop or linear conformations, were designed and used as ^{32}P-labeled probes in protein microarray and conventional DNA-binding assays. The experiments revealed that many mutant proteins bind preferentially and with high affinity to nonlinear DNA, and the binding affinity is strongly dependent on favorable secondary structures of DNA (Figure 17.5). In addition, it had previously been shown that binding mutant proteins to DNA leads to their metabolic stability and constitutive accumulation, thereby compromising the functions of the wild-type p53.[62] Assuming that constitutive binding of mutant proteins to nonbinding DNA structures might promote increased stability, the authors performed

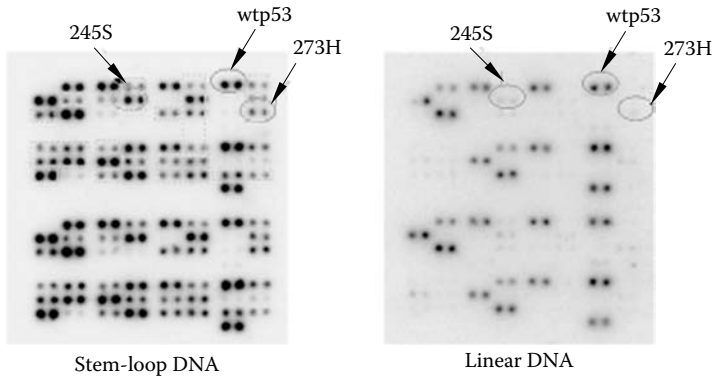


FIGURE 17.5 Protein microarray assay to detect the binding specificity of mutant p53 proteins to stem-loop and linear DNA probes. (From Göhler, T. et al., *Nucleic Acids Res.*, 33(3), 1096, 2005.)[61] The mutant proteins 245S and 273H bind exclusively to stem-loop DNA_{spec} and do not bind to linear DNA_{spec}; wtp53 corresponds to the wild-type p53 protein.

in vitro ubiquitination assays using again the 53 protein arrays.[61] The DNA-dependent protection of mutant proteins from ubiquitination was observed in the presence of the stem-loop DNA probes, whereas no effect was observed in the presence of linear DNAs. Thus, in contrast to the previous view affirming that the p53 mutants bind to DNA in nonspecific fashion, a strong selectivity and a requirement for a stereospecific DNA conformation for binding has been proven. Moreover, the constitutive binding of p53 mutants to secondary structures of DNA might be relevant to the protection of these proteins from degradation. Since a large class of tumor-linked genes is assigned to the DNA-binding proteins, such a multiplexed dissection of the binding properties of tumor-suppressors and other regulatory proteins will be useful for rapid, sensitive, and scalable studies of cancer with similar arrays.

Another example of a possible clinical application of protein arrays to assess DNA interactions is related to autoimmune diseases. In fact, autoantibodies against double-stranded DNA (anti-dsDNA) are often present at higher concentration in systemic lupus erythrematosus (SLE). These autoantibodies are considered pathogenic and important contributors to renal damage in this disease.[63] To date, the patterns of protein and DNA antigens have been used to distinguish various autoimmune disorders. However, each antigen–antibody reaction is measured with a separate assay and the results from various assays are not comparable. Arrayed antigens were used to detect reactive antibodies in the sera of patients in a single assay which, among other observations, confirmed the prevalence of anti-dsDNA autoantibodies in SLE patients.[64] More representative "glomerular proteome arrays" detected high levels of nephrophilic IgG and IgM anti-dsDNA/chromatin antibodies in the sera of lupus mice.[65] Moreover, a distinct IgM cluster, which is highly reactive to DNA and is associated with the disease activity, has been identified in the sera of SLE patients. Recently, the parallel assay with microarrays, but this time with a large number of immobilized antibodies that were partially purified from the sera of SLE patients and healthy people, has been applied to bind proteins and dsDNA in solution. The preliminary results are encouraging in terms of the efficiency of this multiplexed approach to distinguish SLE from other disorders in immobilized biological fluids of patients.[66]

Though an assessment of their diagnostic accuracy must still be carried out, these results underline the great potential of antigen and antibody arrays in the diagnosis of autoimmune diseases.

CONCLUSIONS AND FUTURE DIRECTIONS

Regulatory, signaling and metabolic networks govern the vital processes in organisms. An understanding of these complex interrelationships depends on the elucidation of protein functions, including protein–DNA interactions, which requires an integration of the knowledge acquired by different high-throughput methods. The data accumulated show the feasibility of protein arrays to monitor DNA-binding parameters and moreover, underline their advantages over other methods as a rapid and sensitive tool for the parallel analysis of numerous proteins. Large-scale formats of protein microarrays provide a unique possibility to screen various functions, including the DNA-binding ability of putative proteins deduced from sequenced genomes.

Recently commercialized microarrays, representing largely yeast and partly human proteomes, will accelerate basic and applied research in this attractive field of modern biology. Furthermore, cell-free synthesized proteins appear to be useful for the dissection of toxic or "nonclonable" proteins and their interactions with DNA and other molecular partners. An, as yet, unexploited area of protein arrays is probing small chemical compounds with a goal to comparing and selecting potential inhibitors or activators of protein–DNA interactions. Tailoring different clinical diagnostic and therapeutic strategies is perfectly envisaged by microarray-based binding assays of protein–DNA interactions.

REFERENCES

1. Ptashne, M. and Gann, A., Transcriptional activation by recruitment, *Nature*, 386, 569, 1997.
2. Kelly, T.J. and Brown, G. W., Regulation of chromosome replication, *Annu. Rev. Biochem*., 69, 829, 2000.
3. Surtees, J.A., Argueso, J.L., and Alani, E., Mismatch repair proteins: Key regulators of genetic recombination, *Cytogenet. Genome Res*., 107, 146, 2004.
4. Yang, W., Structure and function of mismatch repair proteins, *Mutat. Res*., 460, 245, 2000.
5. Nagashima, R. et al., Transcriptional factors in the cochlea within the inner ear, *J. Pharmacol. Sci.*, 99, 301, 2005.
6. Ekins, R. and Chu, F., *Protein Arrays, Biochips and Proteomics: The Next Phase of Genomic Discovery*, Albala, J.S. and Humphery-Smith, I., Eds., Marcel Dekker, Inc., New York, 2003, p. 81.
7. Harrison, S.C., A structural taxonomy of DNA-binding domains, *Nature*, 353, 715, 1991.
8. Luscombe, N.M. et al., An overview of the structures of protein-DNA complexes, *Genome Biol.*, 1, REVIEWS001, 2000.
9. Bajic, V.B. et al., Promoter prediction analysis on the whole human genome, *Nat. Biotechnol*., 22, 1467, 2004.
10. Xuan, Z. et al., Genome-wide promoter extraction and analysis in human, mouse, and rat, *Genome Biol.*, 6, R72, 2005.
11. Enright, A.J., Van Dongen, S., and Ouzounis, C.A., An efficient algorithm for large-scale detection of protein families, *Nucleic Acids Res.*, 30, 1575, 2002.
12. Li, L., Stoeckert, C.J., Jr., and Roos, D.S., OrthoMCL: Identification of ortholog groups for eukaryotic genomes, *Genome Res.*, 13, 2178, 2003.
13. Dion, M. et al., The highly thermostable arginine repressor of *Bacillus stearothermophilus*: Gene cloning and repressor-operator interactions, *Mol. Microbiol.*, 25, 385, 1997.
14. Karaivanova, I.M. et al., Mutational analysis of the thermostable arginine repressor from Bacillus stearothermophilus: Dissecting residues involved in DNA binding properties, *J. Mol. Biol.*, 291, 843, 1999.
15. Ni, J. et al., Structure of the arginine repressor from *Bacillus stearothermophilus*, *Nat. Struct. Biol*., 6, 427, 1999.
16. Jones, D.T., Progress in protein structure prediction, *Curr. Opin. Struct. Biol.*, 7, 377, 1997.
17. Sarai, A. and Kono, H., Protein-DNA recognition patterns and predictions, *Annu. Rev. Biophys. Biomol. Struct*., 34, 379, 2005.
18. Phimister, B., Getting hip to the chip, *Nat. Genet*., 18, 195, 1998.

19. Ren, B. et al., Genome-wide location and function of DNA binding proteins, *Science*, 290, 2306, 2000.
20. Iyer, V.R. et al., Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF, *Nature*, 409, 533, 2001.
21. Lee, T.I. et al., Transcriptional regulatory networks in *Saccharomyces cerevisiae*, *Science*, 298, 799, 2002.
22. Jiang, H. et al., Human catechol-O-methyltransferase down-regulation by estradiol, *Neuropharmacology*, 45, 1011, 2003.
23. Kim, J. et al., Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment, *Nat. Methods*, 2, 47, 2005.
24. Velculescu, V.E. et al., Serial analysis of gene expression, *Science*, 270, 484, 1995.
25. van Steensel, B. and Henikoff, S., Identification of *in vivo* DNA targets of chromatin proteins using tethered dam methyltransferase, *Nat. Biotechnol*., 18, 424, 2000.
26. van Steensel, B., Delrow, J., and Henikoff, S., Chromatin profiling using targeted DNA adenine methyltransferase, *Nat. Genet.*, 27, 304, 2001.
27. Bulyk, M.L. et al., Quantifying DNA-protein interactions by double-stranded DNA arrays, *Nat. Biotechnol*., 17, 573, 1999.
28. Bulyk, M.L. et al., Exploring the DNA-binding specificities of zinc fingers with DNA microarrays, *Proc. Natl. Acad. Sci. USA*, 98, 7158, 2001.
29. Linnell, J. et al., Quantitative high-throughput analysis of transcription factor binding specificities, *Nucleic Acids Res.*, 32, e44, 2004.
30. Mukherjee, S. et al., Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays, *Nat. Genet.*, 36, 1331, 2004.
31. Aharoni, A., Griffiths, A.D., and Tawfik, D.S., High-throughput screens and selections of enzyme-encoding genes, *Curr. Opin. Chem. Biol.*, 9, 210, 2005.
32. Xia, G. et al., Directed evolution of novel polymerase activities: Mutation of a DNA polymerase into an efficient RNA polymerase, *Proc. Natl. Acad. Sci. USA*, 99, 6597, 2002.
33. Fa, M. et al., Expanding the substrate repertoire of a DNA polymerase by directed evolution, *J. Am. Chem. Soc.*, 126, 1748, 2004.
34. Sakanyan, V., High-throughput and multiplexed protein array technology: Protein-DNA and protein-protein interactions, *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci*., 815, 77, 2005.
35. Jones, S., van et al., Protein-DNA interactions: A structural analysis, *J. Mol. Biol*., 287, 877, 1999.
36. Luscombe, N.M. and Thornton, J.M., Protein-DNA interactions: Amino acid conservation and the effects of mutations on binding specificity, *J. Mol. Biol*., 320, 991, 2002.
37. Schwabe, J.W., The role of water in protein-DNA interactions, *Curr. Opin. Struct. Biol*., 7, 126, 1997.
38. Pabo, C.O. and Nekludova, L., Geometric analysis and comparison of protein-DNA interfaces: Why is there no simple code for recognition?, *J. Mol. Biol.*, 301, 597, 2000.
39. Garvie, C.W. and Wolberger, C., Recognition of specific DNA sequences, *Mol. Cell*., 8, 937, 2001.
40. Siggers, T.W., Silkov, A., and Honig, B., Structural alignment of protein-DNA interfaces: Insights into the determinants of binding specificity, *J. Mol. Biol.*, 345, 1027, 2005.
41. Bosch, D., Campillo, M., and Pardo, L., Binding of proteins to the minor groove of DNA: What are the structural and energetic determinants for kinking a basepair step?, *J. Comput. Chem.*, 24, 682, 2003.

42. Ge, H., UPA, a universal protein array system for quantitative detection of protein-protein, protein-DNA, protein-RNA and protein-ligand interactions, *Nucleic Acids Res*., 28, e3, 2000.
43. Ghochikyan, A. et al., Arginine operator binding by heterologous and chimeric ArgR repressors from Escherichia coli and *Bacillus stearothermophilus*, *J. Bacteriol*., 184, 6602, 2002.
44. Morin, A. et al., Hyperthermophilic *Thermotoga* arginine repressor binding to full-length cognate and heterologous arginine operators and to half-site targets, *J. Mol. Biol*., 332, 537, 2003.
45. Maas, W.K., The arginine repressor of *Escherichia coli*, *Microbiol. Rev.*, 58, 631, 1994.
46. Gourse, R.L., Ross, W., and Gaal, T., Ups and downs in bacterial transcription initiation: The role of the alpha subunit of RNA polymerase in promoter recognition, *Mol. Microbiol*., 37, 687, 2000.
47. Snapyan, M. et al., Dissecting DNA-protein and protein-protein interactions involved in bacterial transcriptional regulation by a sensitive protein array method combining a near-infrared fluorescence detection, *Proteomics*, 3, 647, 2003.
48. Braun, F. et al., Similarity and divergence between the RNA Polymerase Alpha Subunits from *Thermotoga maritima* and mesophilic *Escherichia coli* bacteria. *Gene* 380, 120, 2006.
49. Espina, V. et al., Protein microarrays: Molecular profiling technologies for clinical specimens, *Proteomics*, 3, 2091, 2003.
50. Lebon, G. et al., unpublished data.
51. Sakanyan, V. et al., Protein Arrays, Methods for Their Preparation and Methods for the Detection of Intermolecular Interactions, International Patent WO 03/012451, 2001.
52. Kersten, B. et al., Protein microarray technology and ultraviolet crosslinking combined with mass spectrometry for the analysis of protein-DNA interactions, *Anal. Biochem*., 331, 303, 2004.
53. Hall, D.A. et al., Regulation of gene expression by a metabolic enzyme, *Science*, 306, 482, 2004.
54. Abadjieva, A. et al., A new yeast metabolon involving at least the two first enzymes of arginine biosynthesis: Acetylglutamate synthase activity requires complex formation with acetylglutamate kinase, *J. Biol. Chem*., 276, 42869, 2001.
55. Quackenbush, J., Microarray data normalization and transformation, *Nat. Genet.*, 32 (Suppl.), 496, 2002.
56. Huet, A. et al., Mechanism of binding of serum response factor to serum response element, *FEBS J*., 272, 3105, 2005.
57. Boutell, J.M. et al., Functional protein microarrays for parallel characterisation of p53 mutants, *Proteomics*, 4, 1950, 2004.
58. Ng, J.H. and Ilag, L.L., Biomedical applications of protein chips, *J. Cell. Mol. Med*., 6, 329, 2002.
59. Soussi, T. and Lozano, G., p53 mutation heterogeneity in cancer, *Biochem. Biophys. Res. Commun*., 331, 834, 2005.
60. Sigal, A. and Rotter, V., Oncogenic mutations of the p53 tumor suppressor: The demons of the guardian of the genome, *Cancer Res*., 60, 6788, 2000.
61. Gohler, T. et al., Mutant p53 proteins bind DNA in a DNA structure-selective mode, *Nucleic Acids Res*., 33, 1087, 2005.
62. El-Hizawi, S. et al., Induction of gene amplification as a gain-of-function phenotype of mutant p53 proteins, *Cancer Res*., 62, 3264, 2002.

63. Limaye, N. and Mohan, C., Pathogenicity of anti-DNA and anti-glomerular antibodies: Weighing the evidence, *Drug Discovery Today: Disease Models*, 1, 395, 2004.
64. Feng, Y. et al., Parallel detection of autoantibodies with microarrays in rheumatoid diseases, *Clin. Chem*., 50, 416, 2004.
65. Li, Q.Z. et al., Identification of autoantibody clusters that best predict lupus disease activity using glomerular proteome arrays, *J. Clin. Invest*., 115, 3428, 2005.
66. Lapin, S. et al., unpublished data, 2006.

18 G Protein–Coupled Receptor Microarrays for Drug Discovery

John Salon, Michael Johnson, Brian Rasnow, Gloria Biddlecome, Yulong Hong, Brian Webb, Ye Fang, and Joydeep Lahiri

CONTENTS

INTRODUCTION

G protein-coupled receptors (GPCRs) represent an extremely important class of drug targets. Approximately 50% of currently marketed pharmaceuticals elicit their actions via specific GPCRs and account for more than $23 billion dollars in yearly sales.[1–3] Individual members of the GPCR superfamily have been linked to a broad spectrum of diseases and are currently the focus of a variety of therapeutic initiatives encompassing pain, asthma, inflammation, and a host of assorted metabolic and feeding disorders, as well as a historically well-entrenched involvement in psychiatric and nonpsychiatric disorders of the CNS and PNS systems.[4,5]

The scope of these pathophysiological roles is in part grounded on the receptor's cellular localization and tissue distribution. Their selective expression throughout the body's organs enables their modulation of very specific yet highly varied tissue-centric physiologies, while their common localization in the cell's plasma membrane makes them directly accessible to both their endogenous transmitter ligands and exogenously applied therapeutic molecules. This distribution and accessibility are important determinants of their "drugability."[6]

An understanding of the structure-function basis of the receptor's signal transduction apparatus can also help explain the molecular action of therapeutic compounds for this target class. Excluding olfactory receptors, some 400 unique GPCR subtypes exist.[3–6] While the extent of sequence similarity varies, all members of this superfamily are considered to share a highly conserved tertiary structure. This structure is characterized by seven membrane-spanning stretches of relatively hydrophobic amino acids, which results in the presentation of an extracellular N-terminus along with three major extracellular loops and an intracellular C-terminus along with three major intracellular loops.[4,5] This serpentine topology presents the unifying molecular architecture through which each specific ligand-receptor system accesses and modulates a limited ensemble of intracellular effector systems.

This similarity in receptor structure underscores the potential for promiscuous interactions with any given pharmacophore. Such promiscuity, which can be especially problematic between closely related receptor subtypes, is an underlying cause of off-target side effects. Accordingly, the degree to which a drug candidate binds to its intended target is of paramount concern during a medicinal chemistry campaign. As the atomic coordinates of GPCR structures are not yet available, the chemical design of selective drugs must currently be driven by empirically testing candidate molecules against both the primary therapeutic target and any cross-reacting targets of concern. Since it is not always possible to predict which collateral targets will be most problematic, assay formats that can be configured to support the broadest possible counter screening paradigms are highly desirable. Array-configured assays,[3,7–9] when enabled with pharmacologically active receptor proteins, can provide such a parallel profiling process in a highly efficient drug discovery application.

GPCR MICROARRAYS FOR DRUG DISCOVERY AND PROFILING

Microarrays have long been used to examine the distribution pattern and regulation of gene expression.[10,11] However, as conventionally enabled to probe nucleic acid interactions, they do not address more fundamental drug-discovery questions regarding the interaction of a drug molecule with its more typical protein target. It is this initial biomolecular interaction that defines the inherent efficacy of a drug candidate and is therefore a key parameter to track during its development. To monitor these early events, arrays must be made to be compatible with assays of the protein target itself.

Currently, a diverse collection of pharmacological assays are used for GPCR drug discovery campaigns.[12–15] These assays are usually characterized as either binding or functional in nature and individually may monitor events ranging from the initial biomolecular interaction of ligand and receptor, to the activation of proximal

ancillary G protein complexes and their associated second messenger systems, and finally to the regulation of various integrated downstream cellular phenomena. While functional assays provide an important pharmacological dimension to the study of ligand efficacy, they can be cumbersome to enable, orchestrate, and execute in a pharmacologically rigorous way to unequivocally quantify drug-target affinity. Alternatively, binding assays can more simply provide a direct measurement of drug-target affinity in a manner that that is less susceptible to many of the signal amplification artifacts common to highly engineered cell based systems. Beyond these pharmacological distinctions, binding assays can provide a fundamental practical advantage. They employ standardized preparations of receptor-rich cell membrane fragments, which are simple to prepare in bulk, easy to quality control, and can be stably stored over long periods for use at a moments notice.

Accordingly, we will discuss here two types of membrane based GPCR pharmacology assays (ligand-binding and GTP-binding), which are configured in a microarray format to enable comparative studies of drug-target interaction across a broadly inclusive panel of receptor systems.

METHODS

The fabrication of GPCR microarrays presents fundamentally different problems from conventional DNA and soluble protein microarrays. The principal protein component of the GPCR system (the receptor proper) is integrally embedded in the cell membrane's lipid bilayer and requires this anisotropic environment to retain a correctly folded and fluid conformation required for proper pharmacological function.[16–18] As such, the production of GPCR microarrays must involve the co-immobilization of both the GPCR protein target and the host membrane in which it is embedded. This roughly planar proteo-lipid sheet needs to be offset from the microarray's physical surface to avoid any interference with access to or mobility of the receptor's extra- and intramembrane domains. The presence of the receptor's ancillary heterotrimeric G protein complex must also be preserved since it is an important determinant of a GPCR's affinity for certain types of ligands.[19] These prerequisites must be fulfilled while maintaining a stable association of the proteo-lipid system with the supporting array surface. Covalent methods for capturing either the protein proper or the surrounding lipid are considered undesirable as they may adversely impact the system's inherent molecular freedom and pharmacological fidelity.[18]

Surface Chemistry

Despite the demonstration of supported lipid bilayers almost two decades ago,[20] the molecular details underlying the interaction of membranes and surfaces remain poorly understood. It is believed to involve a delicate interplay of hydrophobic, electrostatic and surface hydration forces, which are in turn defined by the composition of the membrane and the physicochemical properties of the supporting surface.[17] While it is chemically feasible to construct a supported lipid bilayer entirely by covalent immobilization,[21] the resulting structure would lack the long range fluidity and steric freedom required for accurate function of its protein components. A more appropriate

surface chemistry for bio-membranes is one that can retain the proteo-lipid ensemble in a fashion that is both mechanically robust and sterically fluid.

With these contradictory requirements in mind, a variety of surfaces have been investigated. Surfaces that are deformable or penetrable such as those presenting amphiphilic tethers, polymer cushions or meshlike structures seem conceptually most appropriate. Of those tested, amine-presenting surfaces (e.g., γ-aminopropyl-silane [GAPS]-derivatized surfaces) possess the best combination of physical and functional characteristics.[7] For example, the mechanical stability of supported membranes on GAPS has been shown to be robust as evidenced by the ability of the supported membranes to resist spontaneous desorption when repetitively drawn through a buffer-air interface. This retention was observed for both gel and fluid phase lipids. Lipid bilayers immobilized on GAPS surfaces also exhibit a high degree of lateral fluidity as demonstrated by the rapid kinetics of dye-doped lipid movement seen in fluorescence recovery after photobleaching (FRAP) experiments. Finally and most importantly, the ability of GAPS-coated surfaces to preserve the conformational freedom of embedded proteins has been investigated by characterizing the pharmacological properties of complex receptor-membrane systems such as GPCRs and their associated G protein complex after capture.[9] Preservation of the receptor's native conformation can directly be assessed by measuring its ability to bind with its cognate ligand. For these purposes it has been found that flat GAPS surfaces prepared from smooth glass are ideal. The ability of the receptor to change conformation in response to ligand activation can be assessed by monitoring its ability to interact with peripheral membrane-associated G proteins via measurements of GTP binding. For these purposes we have found that porous GAPS surfaces prepared from fritted glass are preferable, which likely provide for simultaneous access to both the receptor's ligand binding pocket and the G protein activation complex.[8] We believe that the combination of both flat and porous substrates for supported membranes will offer fundamentally new opportunities for understanding and evaluating signaling events at the cell membrane. Accordingly we will focus our methods description on GPCR microarrays produced on such GAPS-coated surfaces.

Protocol: Surface Preparation

GAPS-coated glass slides (GAPS II) are commercially available from Corning Inc. (Corning, NY). Before use, GAPS II slides need to be evaluated for their suitability for GPCR microarrays. The evaluation is primarily based on the contact angle of 2 μl water droplets. Generally, GAPS surfaces with a water contact angle between 25 and 40° are used.

Printing Methods

Membrane microarrays can be fabricated in two fundamentally different ways. The first approach involves the direct deposition of membranes onto micropatterned substrates consisting of membrane-binding and non-membrane-binding regions.[22–27] However, extending this approach to the fabrication of arrays containing microspots of different composition is challenging because of registration issues. The second approach, which we employ, uses direct printing of biological membrane suspensions

onto a homogenous support surface.[7–9] The printing technology *per se* borrows from methods and instrumentation used to prepare DNA microarrays and can be categorized into contact and noncontact methods.

Contact methods using either quill- or solid-pin styluses have proven very suitable for fabricating GPCR microarrays.[28,29] For large-scale fabrication, quill-pins (e.g., Cartesian Technologies, http://www.cartesiantech.com) are preferable as they can significantly reduce consumption of membrane sample. For example, a single load of a quill-pin with an aqueous suspension of membrane fragments can yield several hundred microspots, which equates to sample usage of ~0.5 nl (typically 0.001 to 0.005 fmol of receptor binding sites) or less per data point.

Noncontact technologies, such as piezo and ink-jet printing may also be employed. Excellent results can be obtained with the Sciclone inL10 (http://www.caliperls.com) printer. It is suitable for printing both slides and 96-well microplates and incorporates MEMS-based microflow meters and temperature sensors to individually record the amount of fluid aspirated and dispensed in each channel. This printer is compatible with both porous and flat substrates. Although very well suited for high speed printing in microplates, the minimum volume that can be accurately dispensed is 10 nl, which is ~20 times greater than what is used for quill pin printing.

Whichever method is employed, it is important to understand the potential issues associated with their application to a fragile membrane suspension. For example, protein denaturation or shear-induced stripping effects associated with thermal ink-jet printing may adversely affect either the gross structure of the GPCR itself or the integrity of the GPCR-G protein complex. For either method, the homogeneity of the membrane suspensions used is critical to reproducible printing during the fabrication run and is an important contributor to both the quality and CVs of the signals ultimately observed in the assay.

Protocol: GPCR Microarray Fabrication

A typical GPCR microarray fabrication procedure will include the following steps. This protocol can be used to produce slides or microplates.

1. If necessary, GPCR membrane preparations may be reformulated (obtained from commercial vendor) with a buffered solution (50 m*M* Tris-HCl, pH 7.4, 10 m*M* $MgCl_2$, 10% glycerol, 10% sucrose, 0.1% bovine serum albumin) to a final concentration of 2.0 mg/ml total membrane protein.
2. Transfer 7 µl of each reformulated GPCR membrane preparation to a well of a low-volume 384-well microplate (Corning Life Sciences, Acton, MA).
3. Load a CMP3 quill pin (TeleChem, Atlanta, GA) by dipping into the GCPR membrane suspension.
4. The printing run is primed by preprinting 20 to 50 microspots on a spare slide.
5. The desired number of microspots is continuously printed on GAPS slides using a Cartesian PixSys 5500C arrayer (Cartesian Technologies, Irvine, CA). Approximately 0.5 nl of each membrane preparation (1 ng membrane protein) is deposited in each microspot.

6. Clean the pin and repeat steps 3 to 5.
7. Transfer the printed arrays into a humidity chamber (relative humidity of ~85%) and incubate for 1 hour
8. Store the arrays in a desiccator at 4°C until used.

GPCR Membranes

GPCR microarrays studied to date have employed membrane fragments derived from cultured cells engineered to heterologously express specific receptor targets. In general, conventional mammalian cell hosts (e.g., HEK293, CHO-K1) and expression systems (e.g., SV40 or CMV promoter driven) have been employed, but any membrane system expressing good levels of receptor and associated G protein complex should be suitable for microarray fabrication. The molecular engineering of these cellular systems and subsequent physical isolation of membrane fractions follows standard molecular pharmacology practices. Increasingly, such reagents are being offered commercially; Amersham Biosciences (www4.amershambiosciences.com), PerkinElmer Life Sciences (www.perkinelmer.com), Euroscreen (www.euroscreen.be), Upstate (www.upstate.com), and Sigma (www.sigma-aldrich.com).

In general, best assay results are obtained when these preparations display receptor expression levels of at least 1 pmol receptor/mg-membrane protein,[3,29] which is very modest for typically engineered membrane reagents. It should be kept in mind that the affinity of a receptor for certain types of ligands (especially agonists) is influenced by the receptor coupling to its ancillary G protein complex. Hence, stoichiometry issues arising in cases of receptor reserve may come into play as receptor expression level increases. In such cases and when the option is available, it is prudent to employ the lowest level of receptor expression that provides good assay signal.

The physical and molecular homogeneity of the membrane suspension is another factor that contributes to the consistency of array printing and assay performance.[3,29] The physical state of membranes in suspension will determine both the uniformity of the membrane layer and the density of available receptor binding sites that can be consistently deposited in the microspots. In general this is determined during the cell homogenization and membrane fractionation procedure but where necessary may be subsequently addressed through reformulation. Reformulation may be carried out by recovering the membrane fragments through ultracentrifugation and resuspension in buffer at the desired working concentration.[9] Reformulation buffers typically contain bovine serum albumin (BSA) and sugars such as sucrose or trehalose to further enhance the stability of GPCR arrays. By including Cy3-labeled BSA with the GPCR membrane preparations in the printing ink, we have found that BSA molecules effectively form a packed layer(s) surrounding the printed GPCR membrane microspots.[9] The use of proteins to stabilize supported membranes has also been reported by Cremer and coworkers, who showed that supported lipids presenting biotin resist desorption following the binding of streptavidin.[30] Disaccharide sugars are also effective at improving the integrity of GPCR microarrays.[9] The hypothesis is that the organization of water at a lipid/membrane interface is crucial to the structure and functionality of biomembranes; disaccharides are known to replace water molecules associated with lipid headgroups and therefore effectively stabilize

membranes exposed to drastic environmental changes.[31,32] GPCR microarrays fabricated from membrane preparations resuspended in buffers containing BSA and sucrose (or trehalose) are functionally stable for at least one month at 4°C.

Enabling multiplexed GPCR microarray assays requires a generic buffer formulation that is compatible with each of the printed receptors. A survey of assay conditions typically employed for both ligand- and GTP-binding suggest this is attainable, and our experience successfully enabling assays for a number of divergent GPCRs supports this conclusion.

Ligand-Binding Assay

Fluorescence detection affords an opportunity to avoid the costs of working with radioisotopes and provides the advantage of employing multispectral signal detection through the use of various fluorophores and multichannel laser-based microarray scanners. A variety of commercial instruments are available that support the measurement of fluorescence from microarrays, including but not limited to the Axon GenPix 4000B for slide based arrays or the Tecan LS400, which can accommodate both slide and plate-based formats. Data capture and analysis software is typically packaged along with the instrumentation and is generally useful for most low throughput applications.

Fluorescently labeled GCPR ligands can be obtained from a variety of sources; Molecular Probes (www.probes.com), Phoenix Pharmaceuticals (www.phoenixpeptide.com), Amersham Biosciences (www.amershambiosciences.com), PerkinElmer (www.perkinelmer.com), and Sigma Chemical (www.sigmaaldrich.com). These labeled ligands may encompass any molecules known to specifically bind to the receptor, but will typically fall into the categories of positive agonists and/or neutral antagonists. The pharmacological activity and fidelity of fluorescent ligands should be validated during a normal course of assay development. To maintain the largest possible assay window for ligand displacement assays the labeled ligand should be used at a concentration at or near its K_d. To minimize problems with nonspecific binding interactions the fluorescent ligand used should have a binding affinity in the nanomolar range and high specificity for the receptor(s) of interest.

Binding levels may be assessed in a variety of ways, including measurement of absolute signal intensity, relative signal intensity (vs. neighboring or control spots) or ratiometric signal intensity (in cases where different dyes are employed for different receptors or controls). Given the large dynamic detection range of fluorescence imagers (typically 2 to 3 logs), it is not necessary that each of the ligand-receptor systems included in the array generate similar signal intensities. The spatial encoding inherent in the microarray allows data from each receptor-ligand system to be deconvoluted informatically and archived individually prior to data analysis.

Array-based ligand binding assays may be configured in one of two ways; as a simplex assay employing one labeled ligand at a time, or as a multiplex assay employing a cocktail of labeled ligands. In either case several issues should be kept in mind when enabling array based ligand binding assays: (a) the possible loss of receptor binding caused by introduction of a bulky fluorophore into a receptor's cognate ligand; (b) the extent of cross-receptor reactivity inherent in the labeled-ligand;

(c) the stringency of assay conditions required to produce pharmacologically relevant binding for each receptor-ligand system; (d) the ultimate need to arrive at a generic set of assay conditions that support binding signals of suitable strength and fidelity simultaneously for all of the receptors present in the multiplexed assay. To help address these issues, we have adopted a few guidelines for the selection and design of labeled ligands and receptor systems to be included in arrays. We typically employ a K_d cutoff of < 10 n*M* and a specific binding signal of >80%, as determined by simplex microspot assays, for any labeled ligand to be used with an array. At assay CVs of ~10 to 15% these criteria usually produce a Z' of >0.4, which is marginal for a typical screen. For peptide ligands, we prefer to label the minimum recognition sequence (when known) that contains at least one reactive residue not critical for receptor interaction (e.g., *N*-terminal amino, lysine epsilon amino, or cysteine thiol). This preference is based on our observation that the level of nonspecific binding to the GAPS surface generally increased with peptide length. For nonpeptide ligands, we employ well-established labeling chemistries as dictated by both the reactive linker moieties and the basic pharmacophore requirements of the molecule. The best linker-dye combination for each ligand is determined empirically. While the spatial encoding of the microarrays removes an absolute need for spectrally distinct labeled ligands, use of spectrally distinct fluorophores can eliminate the problem of spurious signal crossover to pharmacologically unrelated receptor spots.

Protocol: Prompt Fluorescence Ligand Binding Assay

Customization and/or optimization of the assay protocol may be required. A typical microarray assay procedure will include the following steps:

1. Printed arrays are "rehydrated" at room temp in a closed chamber (~100% relative humidity) for 30 minutes.
2. Fluorescent ligands are dissolved in binding buffer at or near their K_d (as reported by vendor or as determined by standard saturation binding and Scatchard analysis) just prior to use and kept on ice.
3. A generic binding buffer suitable for most receptor ligand systems consists of 50 m*M* HEPES (pH 7.5), 5 m*M* $MgCl_2$, 1 m*M* $CaCl_2$, 1:40 Perkin Elmer blocking solution A, and 0.05% BSA (w/w).
4. The ligand containing binding solution is applied to cover the grid and allowed to incubate until binding equilibrium has occurred, typically 60 minutes at room temperature. For manual application to slide based arrays this usually requires ~10 µl to cover the grid.
5. After binding equilibrium has been attained, the grid is washed with a stream of water, air dried at room temp and scanned. Plates may be stored in the dark for several weeks with little or no degradation in signal intensity.
6. Binding signals may be detected using an Axon GenePix4000B fluorescent scanner, with quantification of individual spot intensities being made with aid of accompanying GenePix software.

7. Resulting numeric intensity values may be analyzed using standard statistical software (GraphPad Software) to correlate with relative levels of specific binding. For dose response studies IC_{50} values are extracted by standard regression analysis and converted to K_i values by Cheng-Prusoff correction using the K_d values for the respective fluorescent ligands.

Figure 18.1 shows fluorescence images of microarrays comprised of the following GPCR systems: apelin (APJR), bradykinin receptor subtype 2 (BK2R), urotensin (UR2R), melanocortin (MC5R), β-adrenergic receptor subtype 1 (β1R), galanin receptor subtype 2 (Gal2R), motilin (MOTR), neurotensin receptor subtype 1 (NTS1), muscarinic receptor subtype 1 (M1R) and δ2-opioid (OP1R). The fluorescent ligands used consisted of [corresponding GPCR]: bodipy-tetramethylrhodamine (BT) labeled apelin (1–13) (BT-apelin) [APJR], BT-HOE 140 [BK2R], BT-urotensin (BT-urot) [UR2R], BT-α-melanocyte stimulating hormone (BT-NDP-α-MSH) [MC5R], BT-CGP12177 [β1R], BT-motilin (1–16) (BT-Mot) [MOTR], Cy5-naltrexone (Cy5-Nal)

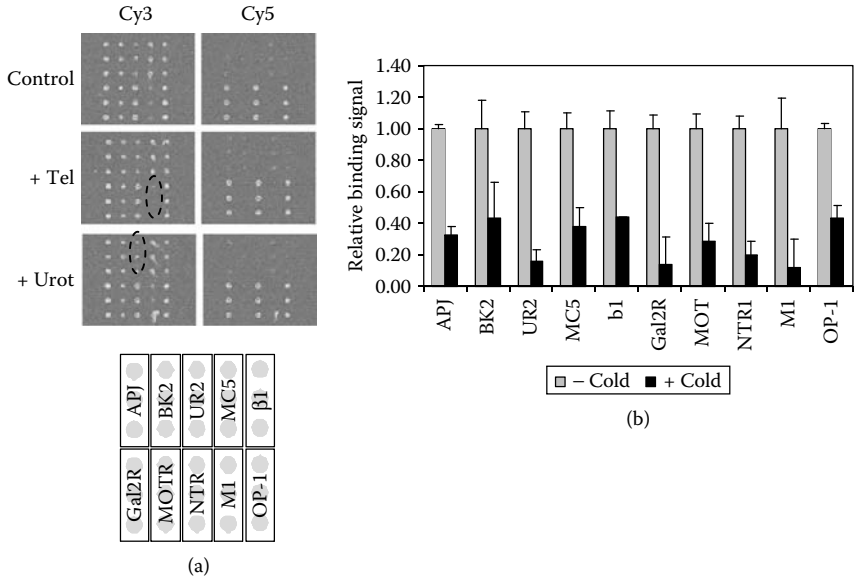


FIGURE 18.1 (A) Fluorescence images (in false color) of a microarray consisting of 10 GPCRs. Each GPCR was printed in triplicate in a column and are positioned as indicated in the legend. Fluorescent signals are generated by incubating the microarray with a cocktail containing: BT-apelin (0.8 n*M*), BT-HOE 140 (0.20 n*M*), BT-urot (0.4 n*M*), BT-NDP-αMSH (0.25 n*M*), BT-CGP12177 (0.25 n*M*), BT-Mot (1.5 n*M*), cy5-Nal (5.0 n*M*), cy5-NT (1.5 n*M*), cy5-gal (1.0 n*M*), and cy5-Tel (0.60 n*M*). The BT- and cy3-labeled ligands were both recorded in the cy3 channel of fluorescence scanner. Receptor specific attenuation of signals is seen upon competition with excess telenzepine (5 μ*M*) or urotensin (5 μ*M*) as indicated. (B) Histogram quantifying the specificity of binding to the GPCRs. Each ligand "in excess" was present at a concentration of 5 μ*M* in the cocktail of labeled ligands. (Hong, Y., Webb, B.L., Sadashiva, P., Ferrie, A., Peng, J., Lai, F., Lahivi, J., Biddlecome, G., Rasnow, B., Johnson, M., Min, H., Fang, Ye, and salon, J., *j. Biomolec Screening,* 6, 2006, 11:435–438. With permission.)

[OP1R], Cy5-neurotensin (2–13) (Cy5-NT) [NTR], Cy5-galanin (Cy5-gal) [Gal2R], and Cy3B-telenzepine (Cy3B-Tel) [M1R]. In preliminary experiments, the selectivity and affinity of each of these labeled ligands was individually validated against their cognate receptors in simplex mode (data not shown). Multiplexed binding assays were then performed using a cocktail of the ten fluorescent ligands. This ligand cocktail generated positive binding signals for all receptors included on the microarray (Figure 18.1A). While the degree of specific binding observed for each receptor was consistent, the absolute signal intensity between different receptor systems did vary. This is most likely attributable to the use of membrane preparations expressing different levels of receptor (B_{max} for these receptors ranged from 0.88 to 24.6 pmol/mg membrane protein) and the use of different working concentrations of fluorescent-ligands which themselves had different inherent fluorescent quantum yields. Since our subsequent analysis is based on the proportion of binding that is displaceable by an excess of unlabeled competing ligand, these variances in signal maxima do not compromise data interpretation providing the window falls within the dynamic range of the scanner. For example, when excess telenzepine (a specific M1R antagonist) was added to this cocktail, only the M1R signal was attenuated (Figure 18.1A). Similarly, the presence of an excess of urotensin resulted specifically in decreased fluorescent signal to UR2R (Figure 18.1A). Histograms quantify the ability to attenuate between 60% and 90% of specific binding for each of the receptors in the array (Figure 18.1B).

The methodology can accurately measure ligand affinity when operating in multiplex mode. For maximum precision, dissociation constants for each labeled ligand against its cognate receptor should first be determined by Scatchard analysis in simplex mode and then used to design the multiplex displacement protocol. Figure 18.2A shows fluorescence images of a microarray consisting of the APJR, UR2R, and Gal2R receptors, co-incubated with a cocktail of their respective fluorescent ligands. The addition of increasing amounts of unlabeled antagonist for each receptor in the microarray results in specific displacement of that receptor's cognate fluorescent ligand (Figure 18.2A). The potencies of the reference compounds tested, as determined by K_i values (Cheng-Prusoff corrected; Figure 18.2B), are in reasonable agreement with values determined by radio-ligand displacement studies (data not shown). Competition curves for compounds at unrelated receptors showed no demonstrable dose-dependant displacement behavior, further supporting the ability to simultaneously quantify compound potency and specificity at each of the multiplexed targets in the panel.

GTP-Binding Assay

Conventional methods for measuring receptor-mediated activation of G proteins require the use of comparatively large amounts of receptor membrane suspensions and employ the radioactive GTP analog [^{35}S]-GTPγS to report the "activated state" of the receptor–G_α protein complex. The development of a solid-state fluorescence-based microarray configured version of such an activation assay is a natural complement to the similarly configured ligand displacement assay we have described, and where labeled GPCR ligands cannot be secured can provide an efficient and generic means of profiling pharmacological activity.

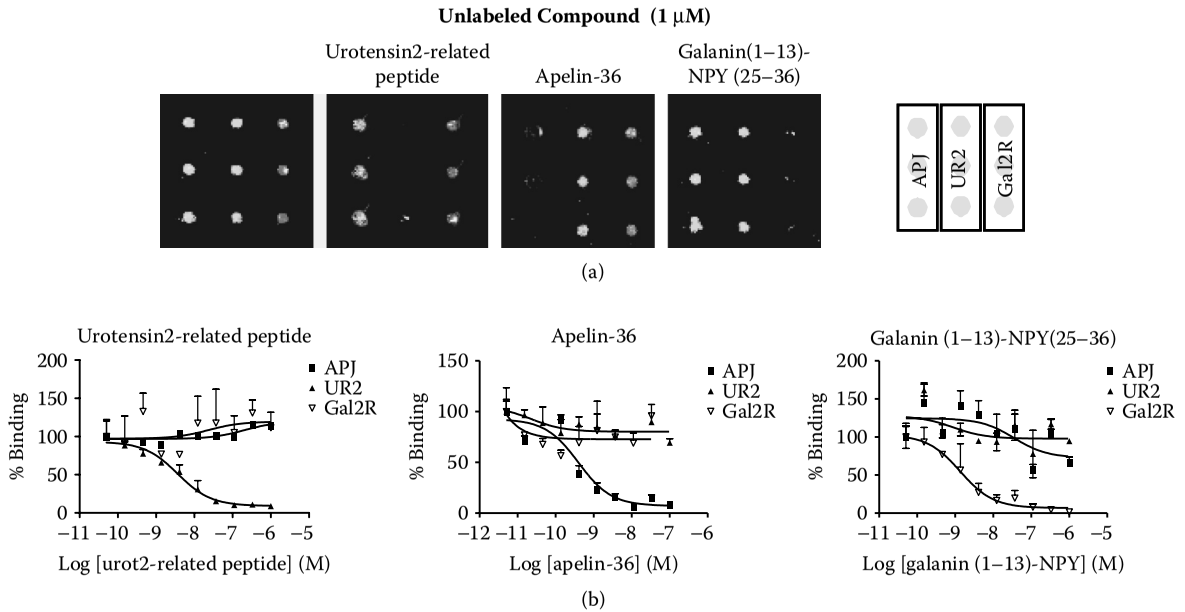


FIGURE 18.2 Evaluation of compound potency and specificity using multiplexed GPCR microarray assays. (A) Fluorescence images of a GPCR microarray consisting of the APJR (*left*), UR2R (*center*), and Gal2R receptors (*right*), treated with a labeled ligand cocktail containing BT-apelin (0.8 n*M*), BT-urot (0.2 n*M*), and cy5-gal (0.4 n*M*), and one of three unlabeled compounds at 1 μ*M* as indicated (urotensin2-related peptide, apelin-36, and galanin [1-13]-neuropeptide Y [25-36]). (B) Displacement curves showing specific inhibition of cognate receptors by increasing concentrations of unlabeled compounds. The estimated K_i values (Cheng-Prusoff corrected) were 2.0 n*M* for urotensin2-related peptide binding to UR2R, 0.22 n*M* for apelin-36 binding to APJR, and 0.75 n*M* for galanin (1-13)-neuropeptide Y (25-36) binding to Gal2R. These values are in reasonable agreement with values previously determined by radio-ligand displacement assays. Competition curves for compounds at unrelated receptors showed no demonstrable dose-dependant displacement.

Two key properties of the commercially available nonhydrolyzable GTP analog, europium labeled GTP (Eu-GTP), make it a probe of choice for enabling such a microarray assay.[33] First, the long fluorescence decay of Eu allows for delayed signal detection and eliminates the problem of background fluorescence arising from other nonpharmacologically relevant components of the assay. Second, the large Stoke's shift of Eu minimizes fluorescent cross-talk, leading to a high signal-to-noise ratio. At the present time, commercially available readers used for detecting homogenous Eu-GTP assays lack the spatial resolution to read microarray assays. Thus, an in-house imaging system capable of detecting the time-gated fluorescent signal of Eu-GTP with sufficient spatial resolution for microarray applications has been constructed. In this device, a moderate-power, CW argon laser emitting at 351 nm is employed to pump the Eu-chelate, and conventional band-pass and UV-reject optical filters used to transmit the europium fluorescence to an intensified CCD detector. In order to take advantage of the long decay lifetime of the europium fluorophore, the pump light is chopped at ~300 Hz to establish a time base, and the intensified CCD is time-gated to integrate the fluorescence beginning roughly 100 μs after the pump pulse. In this manner, any short-lived autofluorescence from the biological array is avoided. UV-grade optics are used to image onto a 512 × 512 pixel CCD, offering a resolution of ~10 μm.

Protocol: Time-Resolved Fluorescence GTP-Binding Assay

Customization and/or optimization of the assay protocol may be required. A typical microarray assay procedure will include the following steps:

1. GPCR microarrays for Eu-GTP binding assays are fabricated the same way as for binding assays except that porous γ-aminopropylsilane treated (GAPS) slides are used instead of smooth GAPS slides.
2. Printed arrays are "rehydrated" at room temperature in a closed chamber (~100% relative humidity) for 30 minutes.
3. Eu-GTP assays are performed by adding 100 μl of assay solution onto each microarray grid. The assay solutions consist of 50 m*M* HEPES, 100 m*M* NaCl, 5 m*M* $MgCl_2$, 3 μ*M* GDP, 100 μg/ml Saponin, 10 n*M* Eu-GTP with or without receptor specific agonists and antagonists.
4. After 1 hour incubation at room temperature, assay solutions are aspirated; the slides washed with GTP wash buffer (Perkin Elmer), and dried with a stream of air.
5. The GPCR microarray is then imaged using an in-house CCD-based imaging system (as described earlier). A microchannel plate within the intensified camera is employed to time-gate the detection window such that fluorescence emission is detected only when the pump light pulse is off, thereby achieving time-resolved fluorescence detection.
6. The signal intensities of each spot are quantified and confirmed using various software packages, some commercially available (e.g., Molecular Devices GenePix Pro, www.moleculardevices.com), and others developed in house (Corning, Inc.). In one version of the Corning software, the user defines the number of spots expected in an array together with an encompassing area of

interest on a particular image. The software subdivides the area into array columns and each column is binned horizontally (summed in intensity) to produce a line profile with peaks representing each spot within the column. The software then produces an intensity value for each spot based upon a peak find algorithm on this line profile. Other software packages can be written to threshold the array image to locate the array spots and then quantify the spot intensities based upon either mean or median intensity levels within each region. A software package developed by Corning (Grid Grinder) that utilizes fast spot locator algorithms and a user-definable algorithm for intensity quantification is available publicly (http://gridgrinder.sourceforge.net).

7. Resulting numeric intensity values may be analyzed using standard statistical software (e.g., GraphPad Software) to correlate with relative levels specific binding. For dose response studies IC_{50} values are extracted by standard regression analysis and converted to K_i values by Cheng-Prusoff correction using the K_d values for the respective fluorescent ligands.

Figure 18.3 shows results of Eu-GTP binding performed on GPCR microarrays fabricated from the neurotensin receptor 1 (NTSR1), the cholinergic receptor muscarinic

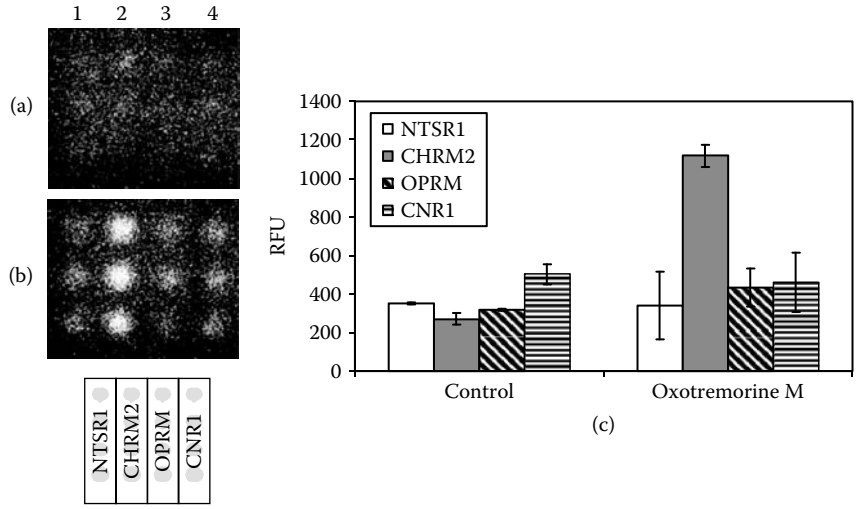


FIGURE 18.3 Images based on agonist induced europium fluorescence that demonstrates the functional activation of GPCR microarrays. The microarrays (from *left* to *right*) consist of the NTSR1 (**1**), CHRM2 (**2**), OPRM (**3**) and CNR1 (**4**) receptors, printed in triplicate. (A) Fluorescence image of the microarray exposed to buffer (HEPES (50 m*M*, pH 7.4) containing GDP (3 m*M*), $MgCl_2$ (5 m*M*), NaCl (100 m*M*), saponin (0.1 mg/ml), and Eu-GTP (10 n*M*). Image of the microarray exposed to either buffer alone (A; control) or buffer containing oxotremorine M (10 μ*M*) (B). The histogram on the right quantifies changes in Europium fluorescence for the receptors upon exposure to oxotremorine M. (Hong, Y., Webb, B.L., Sadashiva, P., Ferrie, A., Peng, J., Lai, F., Lahivi, J., Biddlecome, G., Rasnow, B., Johnson, M., Min, H., Fang, Ye, and salon, J., *J. Biomolec Screening,* 6, 2006, 11:435–438. With permission.)

2 (CHRM2), the opioid receptor mu (OPRM), and the cannabinoid receptor 1 (CNR1). These receptors couple through the G_α proteins $G_{\alpha i}$ or $G_{\alpha q}$ and are well suited for GTP-binding assays. The GPCR microarrays were incubated for 1 h in buffer containing GDP (3 μ*M*) and Eu-GTP (10 n*M*), with or without an agonist. Excess GDP shifts the GDP-GTP equilibrium at the G_α subunit and helps reduce basal fluorescence. Figure 18.3A shows background fluorescence of receptor spots in the absence of agonist. Figure 18.3B shows an image of the microarray incubated with oxotremorine, an agonist for CHRM2. A comparison of Figure 18.3A and 18.3B shows a ~3.5-fold increase in the fluorescence signal for CHRM2, demonstrating the selective activation of the receptor. The selective activation of the NTSR1, OPRM and CNR1 receptors by neurotensin, DAMGO, and anandamide, respectively, which are specific agonists for these receptors, has also been demonstrated (data not shown).

The method accurately reports the inhibitory action of antagonists. A comparison of Figure 18.4A and 18.4B shows the functional activation of all GPCRs by a cocktail of receptor agonists. These agonist induced signals can be selectively inhibited by receptor specific antagonists as shown in Figure 18.4C, where the microarray was incubated with the agonist cocktail and atropine (a known muscarinic antagonist) which results in a specific decrease in fluorescence for CHRM2 alone. The three- to sixfold signal window observed for agonist activation is comparable or better than what is typically observed with conventional radiometric GTP-binding assays and improves the accuracy of dose response estimates of binding and inhibition constants. Through titration experiments, we estimate that oxotremorine M stimulates CHRM2 activation with an EC_{50} of ~53 n*M* (Figure 18.4D). Atropine attenuates this activation with an IC_{50} of ~12 n*M* (Figure 18.4E), both of which are potencies in general agreement with the literature.[34,35]

Assay Characterization and Validation

As with any assay, a certain amount of characterization should be carried out to establish confidence that the tool is reporting pharmacologically relevant results. Some general considerations apply;

It is advisable to validate and compare the pharmacological fidelity of each microarrayed GPCR membrane preparation in both simplex and multiplex mode. Examination of array-configured assay performance must comply with standard theories of GPCR pharmacology.

Confidence must be established that the activity of the principal components (i.e., GPCR, heterotrimeric G protein, displaceable fluorescent-ligand, and the labeled-GTP analog) is as expected. In many cases the fluorescent ligands or membrane reagents to be used may derive from other types of assays (e.g., GPCR Fluorescent Polarization kits) and already be documented as having appropriate "activity." In other instances, such reagents may be engineered in-house. In all cases, it is advisable to characterize their activity via standard pharmacological studies such as saturation binding experiments with Scatchard analysis or dose dependant activation with Shild analysis.

While 60-minute incubations typically prove sufficient for ligand–receptor and G protein–receptor interactions to reach equilibrium, it may be desirable to optimize

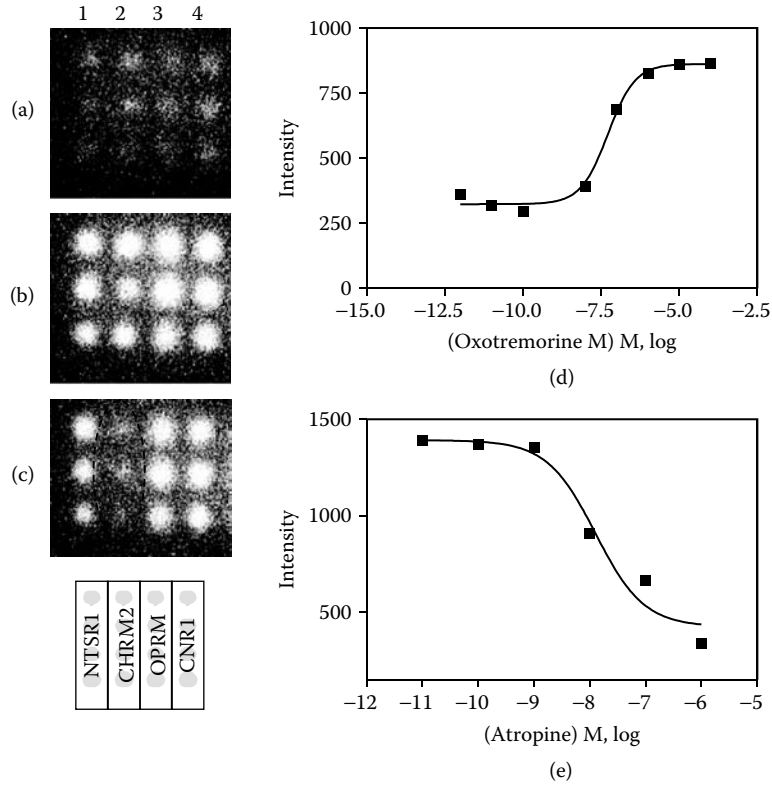


FIGURE 18.4 Fluorescence images of GPCR microarrays demonstrating the screening of antagonists using a cocktail of agonists. (A) Image of a microarray consisting of NTSR1 (**1**), CHRM2 (**2**), OPRM (**3**), and CNR1 (**4**) exposed to a solution containing buffer (50 m*M* HEPES (pH 7.4), GDP (10 m*M*), $MgCl_2$ (5 m*M*), and saponin (0.1 mg/ml)). (B) Image of the microarray exposed to the cocktail of agonists. The solution of agonists in buffer contained neurotensin (1 μ*M*), oxotremorine M (10 μ*M*), DAMGO (10 μ*M*), and anandamide (10 μ*M*), which are cognate agonists to NTSR1, CHRM2, OPRM, and CNR1, respectively. (C) Image of the microarray exposed to the agonist cocktail and atropine (10 μ*M*); the selective inhibition of fluorescence for the CHRM2 receptor is observed (circled). (D and E) Estimation of EC_{50} and IC_{50} using GPCR microarrays. Microarrays of NTSR1, CHRM2, OPRM, and CNR1 were exposed to solutions containing different amounts of oxotremorine M, the cognate agonist for CHRM2. (D) Plot of the increase in fluorescence (at the CHRM2 receptor) as a function of the concentration of oxotremorine M. EC_{50} ~ 53 n*M*. (E) Plot of the decrease in fluorescence for CHRM2 with increasing concentrations of atropine, at a fixed concentration of oxotremorine M (10 m*M*); IC_{50} ~ 12 n*M*. (Hong, Y. et al., Functional GPCR microarrays, *J. Am. Chem. Soc*., 127, 15350, 2005. With permission.)

assay protocols for each receptor system tested. Simple time course studies monitoring signal output will suffice to confirm equilibrium conditions exist and should show one-phase association and dissociation curves. On-rates will typically be dependant on the concentration of the reactants (e.g., labeled ligand, labeled-GTP, G protein alpha subunit, GPCR) while observed off-rates will typically be independent of the reactants.[9]

Once equilibrium conditions have been established, the rank order of potency for several reference compounds for each receptor system should be determined. Where possible and to facilitate rank order placement, commercially available reference compounds should be chosen that have adequate separation in potencies (i.e., 5- to 10-fold). Ten point displacement curves using three-fold concentration dilutions usually suffice to extract a useable IC_{50} value. Once converted to K_i values, the resulting profile should agree with rank orders generated by bench-mark methods such as radio-ligand displacement. Hill slopes should be approximately 1.0, in general agreement with single site binding behavior.

FUTURE DIRECTIONS AND CONCLUSIONS

Future enhancements of the methodology are likely to address three key aspects: (a) Developing microarray configured assays in a microplate format. While the results discussed herein have been based on slide based arrays, we have begun to enable the method in 96-well microplates with the aim of transitioning the technology to automated drug discovery processes. (b) Developing alternative detection methods, in particular radiometric ligand binding and GTP cycle activation assays. A large number of ^{3}H- and ^{125}I-labeled displaceable GPCR agonists and antagonists as well as ^{35}S-γ-GTP analogs are available and are compatible with phosphorimagers such as the Typhoon 9410 (http://www.amershambiosciences.com) that have the resolution and sensitivity required for microarray analysis. (c) Developing applications for target families beyond GPCRs. Membrane arrays are uniquely well suited to probing cell surface phenomena such as receptor dimerization, antibody-receptor clustering during an inflammatory response, and multivalent pathogen–host cell recognition during infection. Systematic variation and control of receptor protein density and composition without compromising physiological membrane fluidity is relatively straightforward using membrane microarrays and impossible or very difficult using other biochemical or cell based systems.

The results described here demonstrate the ability to enable microarray configured multiplexed GPCR binding assays that can be used to simultaneously screen and characterize compound collections against a panel of receptor targets. The pharmacological fidelity of the assay is comparable to conventional methods but improves upon conventional methods by enabling parallel target queries in a miniaturized format. These features can effectively accelerate the elucidation of structure activity relationships for compounds against both the therapeutic target proper and a spectrum of collateral targets of interest. The microarray approach not only increases information content on biological and chemical axes, but does so in a way that minimizes turn around time and reagent consumption. These attributes can most effectively support medicinal chemistry efforts to increase a drug lead's therapeutic efficacy through enhanced target affinity and selectivity. Taken to its logical endpoint, the biological (target) and chemical (compound) throughput afforded by the microarray system has the potential to enable a true chemical-genomics approach to drug discovery.

REFERENCES

1. Drews, J., Drug discovery: A historical perspective, *Science,* 287, 1960, 2000.
2. Ma, P. and Zemmel, R., Value of novelty, *Nat. Rev. Drug Discov*., 1, 571, 2002.
3. Fang, Y., Lahiri, J., and Picard, L., G protein-coupled receptor microarrays for drug discovery, *Drug Discov. Today*, *8*, 755, 2003.
4. Rockman, H.A., Koch, W.J., and Lefkowitz, R.J., *Seven-transmembrane-spanning receptors and heart function, Nature,* 415, 206, 2002.
5. Schöneberg, T., Schulz, A., and Gudermann, T., The structural basis of G-protein-coupled receptor function and dysfunction in human diseases, *Rev. Physiol. Biochem. Pharmacol.,* 144, 143, 2002.
6. Hopkins, A.L. and Groom, C.R., The druggable genome, *Nature Rev. Drug Discov.,* 1, 727, 2002.
7. Fang, Y., Frutos, A.G., and Lahiri, J., Membrane protein microarrays, *J. Am. Chem. Soc.,* 124*,* 2394, 2002.
8. Hong, Y. et al., Functional GPCR microarrays, *J. Am. Chem. Soc*., 127, 15350, 2005.
9. Fang, Y. et al., Air-stable G protein-coupled receptor microarrays and ligand binding characteristics, *Anal. Chem.,* 78, 149, 2006.
10. Schena, M. et al., Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science,* 270, 467, 1995.
11. Levy, S.E., Microarray analysis in drug discovery: An uplifting view of depression, STKE, pe46, 2003.
12. Milligan, G., High-content assays for ligand regulation of G-protein-coupled receptors, *Drug Discov. Today,* 8, 579, 2003.
13. Williams, C., cAMP Detection methods in HTS: selecting the best from the rest, *Nat. Rev. Drug Discov.,*3, 125, 2004.
14. Hu, C.-D. and Kerppola, T.K., Simultaneous visualization of multiple protein interactions in living cells using multicolor fluorescence complementation analysis, *Nat. Biotechnol.,* 21, 539, 2003.
15. Leifert, W.R. et al., G-protein-coupled rectpros in drug discovery: Nanosizing using cell-free technologies and molecular biology approaches, *J. Biomol. Screening*, 10, 765, 2005.
16. Karlsson, O.P. and Lofas, S., Flow-mediated on-surface reconstitution of G-protein coupled receptors for applications in surface plasmon resonance biosensors, *Anal. Biochem.*, 300, 132, 2002.
17. Sackmann, E., Supported membranes: Scientific and practical applications. *Science,* 271, 43, 1996.
18. Fang, Y., Frutos, A.G., and Lahiri, J., G protein-coupled receptor microarrays. *ChemBiochem,* 3, 987, 2002.
19. Haga, T. and Berstein, G., *G Protein-Coupled Receptors*, CRC Press, Boca Raton, FL, 1999.
20. Tamm, L.K. and McConnell, H.M., Supported phospholipid bilayers, *Biophys. J.,* 47, 105, 1985.
21. Ross, E.E. et al., Formation of self-assembled, air-stable lipid bilayer membranes on solid supports, *Langmuir*, 17, 2305, 2001.
22. Hovis, J.S. and Boxer, S.G., Patterning and composition arrays of supported lipid bilayers by microcontact printing, *Langmuir*, 17, 3400, 2001.
23. Cremer, P.S. and Yang, T.J., Creating spatially addressed arrays of planar supported fluid phospholipid membranes, *J. Am. Chem. Soc*., 121, 8130, 1999.

24. Morigaki, K. et al., Patterning solid-supported lipid bilayer membranes by lithographic polymerization of a diacetylene lipid, *Ang. Chem. Inter. Ed.,* 40, 172, 2001.
25. Orth, R.N. et al., Creating biological membranes on the micron scale: Forming patterned lipid bilayers using a polymer lift-off technique, *Biophys. J.*, 85, 2066, 2003.
26. Hovis, J.S. and Boxer, S.G., Patterning barriers to lateral diffusion in supported lipid bilayer membranes by blotting and stamping, *Langmuir,* 16, 894, 2000.
27. Kung, L.A. et al., Patterning hybrid surfaces of proteins and supported lipid bilayers, *Langmuir,* 16, 6773, 2000.
28. Fang, Y., Ferrie, A.M., and Lai, F., Production of protein microarrays using robotic pin printing technologies, in *The Proteomics Protocols Handbook,* Walker, J.M., Ed., Humana Press Inc., Totowa, NJ, 2005, chap. 61.
29. Fang, Y., et al., Fabrication and application of G protein-coupled receptor microarrays, in *Methods in Molecular Biology 264: Protein Arrays,* Fung, E., Ed., *Humana Press*, Totowa, NJ, 2004, chap. 20.
30. Holden, M.A. et al., Creating fluid and air-stable solid supported lipid bilayers, *J. Am. Chem. Soc.,* 126, 6512, 2004.
31. Bardos-Nagy, I. et al., Effect of trehalose on the nonnond associative interactions between small unilamellar vesicles and human serum albumin and on the aging process, *Langmuir*, 19, 146, 2003.
32. Luzardo, M.C. et al., Effect of trehalose and sucrose on the hydration and dipole potential of lipid bilayers, *Biophys. J.*, *78*, 2452, 2000.
33. Frang, H. et al., Nonradioactive GTP binding assay to monitor activation of G protein-coupled receptors, *Assay Drug Dev. Technol.,*1, 275, 2003.
34. DeLapp, N.W. et al., Determination of [35S]guanosine-59-*O*-(3-thio)triphosphate binding mediated by cholinergic muscarinic receptors in membranes from Chinese hamster ovary cells and rat striatum using an anti-G protein scintillation proximity assay, *J. Pharmacol. Exp. Ther.,*289, 946–955, 1998.
35. Daeffler, L. et al., Inverse agonist activity of pirenzepine at M2 muscarinic acetylcholine receptors, *Br. J. Pharmacol.,*126, 1246–1252, 1999.
36. Hong, Y., Webb, B.L., Sadashiva, P., Ferrie, A., Peng, J., Lai, F., Lahivi, J., Biddlecome, G., Rasnow, B., Johnson, M., Min, H., Fang, Ye, and salon, J., *J. Biomolec Screening,* 6, 2006, 11:435–438.

19 Kinase Substrate Identification Using Yeast Protein Microarrays

Geeta Devgan and Michael Snyder

CONTENTS

INTRODUCTION

Protein phosphorylation is one of the most abundant posttranslational modifications affecting cellular function in both lower and higher life forms.[1] Reversible protein phosphorylation is an essential mechanism for regulating basic functions such as DNA replication, cell cycle control, gene transcription, protein translation, and energy metabolism. Such control is achieved by protein kinases and protein phosphatases. All protein kinases catalyze the transfer of the γ-phosphate group of ATP to the hydroxyl groups of serine, threonine, or tyrosine residues in protein substrates, with the exception of histidine kinases (which phosphorylate histidine residues).

The significance of protein phosphorylation in eukaryotic signaling pathways is illustrated by the fact that protein kinase domains are found in about 2% of eukaryotic proteins including those of yeast, flies and humans.[2] Moreover, approximately 30% of cellular proteins contain covalently bound phosphate, and abnormal levels of protein phosphorylation are a cause or consequence of major diseases such as cancer, diabetes, and rheumatoid arthritis.[1] For example, the first discovered proto-oncogene v-Src encodes an aberrantly regulated tyrosine kinase.[3] Phosphorylation not only activates or deactivates a protein target, but can also alter the rate at which a protein is degraded, its ability to translocate from one subcellular compartment to another, and its capacity to bind with other proteins. Therefore, it is the spatial and temporal

control of kinase activity that achieves signal integration within cells by allowing different proteins to work synergistically or antagonistically.

Since the discovery of enzymatic modification by phosphorylation, kinases have been the focus of experimental studies for over 50 years. Pivotal to the identification and study of protein kinases and their functions has been the model organism, *Saccharomyces cerevisiae*. The budding yeast kinome contains 122 protein kinases that take part in many cell functions including DNA replication, cell cycle control, gene transcription, protein translation, energy metabolism, signal transduction, environmental responses, and differentiation.[4,5] Despite the widespread occurrence of phosphorylation in a range of diverse functions, less than 200 proteins are known to be directly phosphorylated among the 6000 proteins of the yeast proteome.[1,5–7] If a third of the proteome is estimated to be phosphorylated, then a significant amount of kinase substrates in yeast has yet to be revealed. Therefore, current efforts are geared toward defining and understanding the key substrates of protein kinases.

Before the sequencing of the yeast genome, the analysis of mutant phenotypes facilitated both the identification of novel protein kinase genes and the characterization of pathways in which these kinases function. For example, the fundamentals of MAP kinase signaling have largely been elucidated in budding yeast and have provided valuable information for their mammalian orthologs. Screens for mating-defective mutants (STE mutants) enabled the resolution of the kinase cascade that controls mating, beginning with the upstream pheromone receptor to the downstream transcription factor Ste12p.[8] Yeast-based studies have also demonstrated how distinct MAPK cascades can share several protein components, yet avoid inappropriate cross-talk. Multiple MAPK phosphatases, scaffolding proteins, and subcellular compartmentalization are among the regulatory mechanisms yeast employs to maintain pathway specificity.[9] Individual laboratories studying various mutant phenotypes have contributed significantly to the kinase field by deciphering key steps of kinase signaling pathways.

Although significant information has been garnered with respect to protein kinases, the same does not hold true for the substrates of protein kinases. Identification of the full range of protein kinase substrates has been the slow step in this area of research. Traditional methods of substrate identification involve hypothesizing which proteins in a particular pathway would be expected to be phosphorylated by a defined kinase. Following the purification of the suspected kinase-substrate pairs, testing for phosphorylation is achieved using *in vivo* labeling or *in vitro* solution assays. Such experiments are time consuming as purification and characterization of a protein may take a number of years. Moreover, in cases where the identified protein kinases have no ascribed function, the molecular targets are harder to predict. Accordingly, using traditional methods, the identification of the major substrates of every protein kinase would be a massive undertaking that would take several decades complete. Thus, more powerful methods are necessary to accomplish this feat.

In the post-genome era, the availability of the complete nucleotide sequences from a number of eukaryotic organisms has made it possible to understand gene function on a global scale. High-throughput genomic experiments have been made possible with the development of DNA microarrays. DNA chip technology has been extremely valuable in profiling gene expression patterns, mapping novel transcripts,

determining sequence mutations and deletions, and identifying transcription factor binding sites, all on a genome-wide scale.[10] However, genomic analysis cannot provide functional or biochemical characterization for gene products. Therefore, chip technology has been developed for proteomic analysis in order to understand the biochemical activities of encoded proteins in the genome.

Functional protein microarrays provide a platform to screen tens to thousands of proteins and have been used to detect protein–protein, protein–lipid, protein–DNA and protein–small molecule interactions and to perform enzymatic assays. In the case of yeast, a "proteome array" was first described in which the majority of the yeast proteome is expressed and deposited on a surface in an addressable format.[11] Thus far, the yeast proteome array has been used to screen for (a) calmodulin binding partners, (b) phospholipids interactions, (c) novel DNA binding activities, (d) small molecule inhibitors and enhancers of rapamycin, and (e) antibody specificity.[11–14] In addition, a microwell-type protein microarray fabricated from a silicone elastomer has been used to screen the activity of 119 kinases from *S. cerevisiae* on 17 substrates.[5] Each kinase was incubated in a microwell with a specific substrate and ^{33}P-γ-ATP. In addition to identifying known phosphorylation events, the kinase assays demonstrated that 27 kinases were found to be capable of phosphorylating a tyrosine substrate, polyGlu-Tyr. This was a striking discovery as yeast protein kinases are generally thought to phosphorylate only serine or threonine residues. Both novel and known kinase activities were observed, thereby demonstrating the utility of protein chip technology in kinase studies. Therefore, it is now possible to globally identify kinase substrates by incubating addressable proteome arrays with each of the yeast kinases and labeled ATP and identifying those proteins that are phosphorylated. As described below, this approach was used to identify *in vitro* substrates for 87 yeast protein kinases. This chapter seeks to explain (a) the methodology used to optimize kinase assays on proteome arrays, (b) the analysis of kinase assays on proteome arrays and, finally, (c) the future directions that can be addressed using such technology.

DEVELOPMENT OF KINASE ASSAY ON YEAST PROTEOME ARRAY

The initial step in the development of the yeast proteome array was to construct a comprehensive expression library that consists of all proteins encoded by the genome of *S. cerevisiae*. Due to the simplicity of the yeast genome architecture, the process of identifying open reading frames (ORFs) is relatively clear-cut. The consideration that complicates this process is to build an expression library in a vector that allows for both high throughput and pure protein production. An expression library of *S. cerevisiae* ORFs was first created with such aspects taken into account. A total of 5800 ORFs were cloned with N-terminal glutathione-S-transferase polyhistidine fusion tags (GST::His_6) and over-expressed under a Gal-inducible promoter.[11] Following high throughput expression and purification in yeast, each protein was printed in duplicate onto glass slides using a standard robotic microspotter.

A customized version of the proteome array, the ProtoArray (manufactured by Invitrogen), was created on a surface-modified microscope slide and used to study

kinase activity.[15] The ProtoArray contains approximately 4400 proteins purified from the GST::His_6 yeast library that consistently express proteins of correct size as verified by western blot analysis. In order to ensure for proper kinase assay conditions, a number of controls were also added to the proteome array. The autophosphorylating kinases Pka2, Pkc-, α and calmodulin-dependent kinase Cmk1 were added at various locations to serve as both positive controls and landmarks for the identification of phosphorylation signals on the array. Common kinase substrates, such as myelin basic protein (MBP), histone H1, casein, polyGlu-Tyr, and a carboxy terminal domain (CTD) peptide containing three copies of the acidic CTD of RNA polymerase II were also included to exhibit the addition of kinase activity on the array. With an optimized array in hand, the next step was to purify individual yeast protein kinases that are suitable for presentation on an array format.

Before being added to the yeast proteome array, yeast protein kinases must be optimized for activity and purity. Production of kinases in native cells optimizes the activity so that proper post-translational modifications may occur. Therefore, in most cases, yeast protein kinases were expressed and purified from the GST::His_6 library. A total of 82 unique kinases were tested; two cyclin-dependent kinases, Pho85 (in complex with Pcl1, Pcl2, Pcl9, and Pho80) and Cdc28 (in complex with Cln2 and Clb5), were also analyzed. The Pho85 kinases were purified from insect cells and the remaining 81 kinases were purified from yeast.[16] In short, yeast protein kinases were expressed in 50 to 500 ml volume and lysed by bead-beating in lysis buffer in the presence of phosphatase and protease inhibitors.[15] Kinases were eluted into kinase buffer and tested for purity by immunoblot analysis using GST antibodies. The activity of each kinase preparation was then tested by solution assays with common substrates such as MBP, histone H1 and casein, each in the presence of ^{33}P-γ-ATP. The solution assays were analyzed by gel electrophoresis followed by exposure to X-ray film. Activity of each kinase was assessed by the amount of ^{33}P-γ-ATP incorporated by the common substrates. Also, any contaminating kinase activity was visualized by the gel assay. Once the purity and activity of each protein kinase was determined, the concentration of kinase needed to phosphorylate immobilized proteins on the surfaces of glass slides was optimized.

The ideal signal to noise ratio was determined for each kinase using test protein arrays containing approximately 300 yeast proteins and common kinase substrates. A dilution series (typically 1:1, 1:2, 1:5, 1:10, and 1:20) of each kinase was made in a total volume of 200 µl kinase buffer. For example, 20 µl of an eluted kinase preparation was diluted into 180 µl of kinase buffer with a consistent amount of ^{33}P-γ-ATP and overlaid onto a test slide. The signal to noise ratio was determined for every kinase on a test array and the same concentration was used on the proteome array.

For every kinase, two yeast proteome arrays were probed in the presence of ^{33}P-γ-ATP using the optimized conditions. For each experiment, two additional arrays were incubated in the absence of kinase to serve as an autophosphorylation reference. Additional negative controls were obtained by incubating the arrays with kinases containing inactivating mutations in their catalytic domains with four kinases (Rim15, Dbf2, Hsl1 and Rad53) and the arrays exhibited signals identical to those

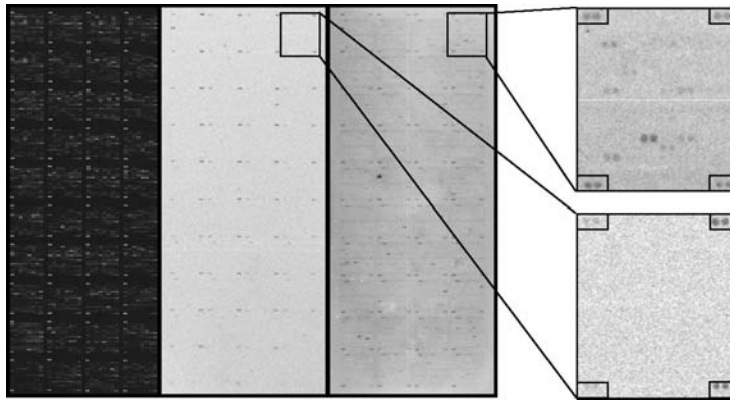


FIGURE 19.1 Large-scale identification of substrates for a yeast protein kinase. **(A)** 4400 different budding yeast proteins tagged with GST::His_6 were over expressed and purified by affinity chromatography and spotted in duplicate on a surface-modified glass slide. The amount of bound protein was detected with a fluorescent antibody to GST. **(B, C)** Two proteome arrays were incubated with a yeast kinase in the presence of ^{33}P-γ-ATP. In addition, two proteome arrays were probed in the absence of kinase to identify proteins that autophosphorylate. Radioactive phosphorylated proteins were detected as pairs of dark spots by autoradiography. Commercial kinases were spotted at many defined locations, displayed in the four boxed corners of the two magnified views on the right; these served as landmarks for the identification of phosphorylation signals. (From Ptacek, J. et al., Global analysis of protein phosphorylation in yeast, *Nature*, 438, 679, 2005. With permission.)

obtained in the absence of protein kinase. Following the kinase assay, each proteome array was exposed to X-ray film. The optimal exposure time was selected for each kinase and compared to the corresponding autophosphorylation slides. Substrate proteins that displayed reproducible signals higher than those of neighboring spots in at least three of the four spots were identified and then compared to the autophosphorylation control. Only those spots that were phosphorylated in the presence of active kinase relative to the control were scored as positive substrates. Protocol 1 outlines the kinase assay experiment and Figure 19.1 depicts a kinase assay on a proteome.

Protocol 1: Kinase Assay on Protein Microarray[15]

1. Express and purify active kinase–elute into kinase buffer: 100 m*M* Tris pH 8.0, 100 m*M* NaCl, 10 m*M* $MgCl_2$, 1 m*M* DTT, 0.5 mg/ml BSA, 0.1% Triton X-100.
2. Verify purity of kinase by immunoblot analysis and/or Coomassie staining.
3. Verify activity of kinase by solution assay: Incubate 2 µl of eluted kinase with 25 ng casein, 25 ng histone H1 and 25 ng MBP and 1 n*M* ^{33}P-γ-ATP in a total volume of 10 µl. Allow kinase reaction to proceed for 1 hour

at 30°C. Add sample loading buffer to stop kinase reaction and analyze by gel electrophoresis.

4. Assess optimal kinase concentration to be used on proteome arrays by testing a dilution series on test arrays containing 100 to 300 immobilized substrate proteins and common kinase substrates. Follow same steps as shown below (5, 7–10) for kinase incubation on the array.
5. Block two proteome arrays for every kinase and two additional slides for negative controls in Superblock (Pierce) with 0.1% Triton X-100 at 4°C.
6. Prepare optimal kinase concentration determined in step 4 by diluting eluted kinase into kinase buffer (total volume = 200 µl) that contains 33.3 n*M* ^{33}P-γ-ATP (Amersham, Piscataway, NJ, USA).
7. Overlay each proteome array with 200 µl of kinase + ^{33}P-γ-ATP, cover with a coverslip and place in a humidified chamber at 30°C for 1 hour.
8. Remove kinase and unbound radionucleotides by two washes with 0.5% SDS in 10 m*M* Tris 7.4 and one wash in ddH_2O.
9. Dry proteome arrays by spinning at 1500 rpm and then expose to X-ray film or phosphoimager. Determine optimum exposure for each kinase (typically 3 exposures were taken for each kinase assayed: 1, 3, and 7 day).
10. Scan the X-ray film at 1800 dpi and analyze phosphorylated spots by Genepix and an algorithm (Ptacek et al., 2005) specifically designed to detect positive signals.

ANALYSIS OF KINASE ASSAY RESULTS

A total of 4192 phosphorylation events involving 1325 proteins were identified from the 87 yeast protein kinase assays on proteome arrays. Each kinase phosphorylated 1 to 256 substrates on the array, with an average of 47 substrates per kinase. Furthermore, most substrates (73%) were recognized by fewer than three kinases. Therefore, proteome arrays are sensitive enough to reveal a unique substrate recognition profile for each kinase. This work represents the first large-scale functional assay to determine all the phosphorylation events in an organism, the "phosphorylome."

The phosphorylome revealed potentially untold biological information and novel regulatory interactions for each kinase tested. Using global localization data, the phosphorylome data was systematically analyzed for kinase and substrates that reside in the same cellular compartment.[17] A third of these interactions (1384) occur between kinases and substrates that are in the same localization category, representing a significant enrichment ($p < 10^{-99}$) over the proteome as a whole. Additionally, the phosphorylome was filtered for kinase substrate pairs that occur in the same functional category. Based on functional data from the Munich Information for Protein Sequences (MIPS) database, 18.4% (768) of the interactions occur between kinases and substrates in the same functional category, also a significant enrichment ($p < 10^{-99}$) over the proteome as a whole.[18] These results were validated *in vivo* for fifteen cases by generating mutants where the kinase has been deleted from yeast strains containing the endogenous candidate substrate protein tagged with a TAP tag and looking for a mobility shift by gel electrophoresis or loss of phosphorylation using phospho-specific antibodies.[19] In addition to finding likely kinase-substrate

pairs, nearly every kinase phosphorylated substrates with functions other than those previously known for such kinases. Accordingly, kinases with known functions may have novel roles and, moreover, uncharacterized kinases may be assigned potentially new functions. Thus, the data generated in this study represents a leap forward in the field of kinase proteomics that will lead to the investigation of new pathways for the different protein kinases.

The investigation of novel pathways have begun with the examination substrate specificity among related kinases. For example, three related kinases (Tpk1, Tpk2, and Tpk3) were analyzed on a proteome-wide level by incubating each kinase on proteome arrays prepared and probed at the same time. The yeast protein kinase A homologs Tpk1 and Tpk3 are 84% identical in amino acid sequence and each is 67% and 76% identical to Tpk2, respectively. Yeast strains lacking all three are inviable whereas those containing any one of the three are viable, indicating that each Tpk is genetically redundant for cell growth. However, each kinase has different roles in pseudohyphal growth and thereby has distinct biochemical functions. To determine if the Tpk kinases are functionally redundant biochemically, their substrate profiles were compared with one another. Only six substrates were recognized by all three kinases and the majority of the *in vitro* substrates (87.7%) were recognized by only one of the Tpks. Thus, the amino acid differences of the Tpks have a significant effect on substrate recognition.

In order to fully understand the scope of the phosphorylome, this comprehensive phosphorylation map was integrated with similar global networks in yeast such as transcription factor binding and protein interaction data.[18–26] Through the integration of such data sets, numerous regulatory networks have been identified that were not otherwise apparent. Distinct modules have been identified to explain how kinase activity can achieve signal integration within cells as displayed in Figure 19.2. Eight particular modules were observed from this fully integrated global network involving kinase-substrate pairs referred to as "kinates:" (1) interacting kinates, (2) scaffolds, (3) kinase cascades, (4) transcription-factor-regulated kinates, (5) kinate regulon, (6, 7) feedback loops, and (8) heterosubstrate regulation. Many of these networks were of high statistical significance and will stimulate the further characterization of kinase pathways, acting as a starting point for the identification of new kinase mechanisms.

FUTURE DIRECTIONS

Protein kinases act as master regulators of cells and because of their key role as potential oncoproteins, they constitute one of the most important classes of drug targets. For example, the development of kinase inhibitors, including Gleevec and Herceptin, are among the most promising anticancer therapies. Due to the highly conserved nature of kinase signaling pathways from fungi to humans, the comprehensive identification of the yeast phosphorylome will shed light on the kinase circuitry of all eukaryotes.[27] The utility of proteome arrays in kinase research is a significant advancement for the signal transduction community. Researchers at pharmaceutical companies will utilize this data to determine equivalent kinase interactions

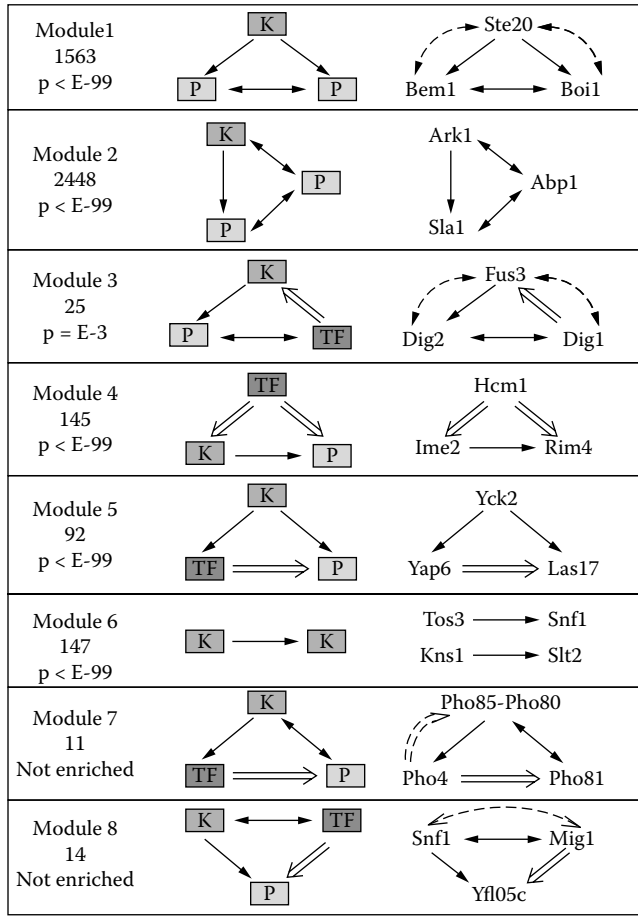


FIGURE 19.2 Integration of phosphorylome with other data sets reveal common regulatory modules. Shown are protein–protein interactions (←→), kinase phosphorylations (→), and transcription factor (TF) regulation (→). K, kinase; P, protein. Modules are numbered from 1 to 8, listed below each is the number of occurrences and the statistical significance of such events. (From Ptacek, J. et al., Global analysis of protein phosphorylation in yeast, *Nature*, 438, 679, 2005. With permission.)

in humans and to further test and develop drugs that can affect such pathways. Furthermore, the development of kinase assays on proteome arrays is poised to aid in the identification of kinase inhibitors. For example, the dose-dependent addition of an ATP-competitive inhibitor, H89, was shown to inhibit the activity of protein kinase A on a group of proteins printed on an array.[28] Similarly, small molecule inhibitors can be used to test the inhibition of both yeast and human kinases on arrays printed with both yeast and human protein substrates. The ability to perform kinase reactions on protein arrays will greatly enhance kinase research in the post-genome era.

CONCLUSION

As described above, protein microarrays have proved valuable for providing a platform to elucidate kinase function on a global scale in *Saccharomyces cerevisiae*. There are many advantages to using the proteome chip to study protein phosphorylation. The ability to rapidly screen the majority of the yeast proteome in an unbiased and high-throughput manner constitutes a fundamental shift in the way kinase-substrate relationships have been previously identified. Typical approaches of identifying interactions between substrates and protein kinases may take upwards of months to years to complete. Now, an entire proteome can be surveyed by a single kinase to come up with a list of candidate interactions. By combining that list with other data sets such as cellular localization, functional categorization, transcription factor binding, and protein interactions, networks that are likely to occur *in vivo* can be determined and follow-up experiments can be undertaken. Another key advantage is the use of minimal amounts of reagents; only 200 μL of an active kinase preparation, typically of nanomolar quantity, is needed to scan the entire proteome array. Finally, the proteome arrays are sensitive enough to detect the biochemical differences between related kinases based on their substrate profiles. In total, the proteomic approach will provide a more powerful and definitive method to elucidate how kinases mobilize diverse regulatory strategies within all living organisms.

REFERENCES

1. Cohen, P., The regulation of protein function by multisite phosphorylation — a 25 year update, *Trends Biochem. Sci.*, 25, 596, 2000.
2. Rubin, G.M. et al., Comparative genomics of the eukaryotes, *Science*, 287, 2204, 2000.
3. Levinson, A.D. et al., Evidence that the transforming gene of avian sarcoma virus encodes a protein kinase associated with a phosphoprotein, *Cell*, 15, 561, 1978.
4. Hunter, T. and Plowman, G. D., The protein kinases of budding yeast: six score and more, *Trends Biochem. Sci.*, 22, 18, 1997.
5. Zhu, H. et al., Analysis of yeast protein kinases using protein chips, *Nat. Genet.*, 26, 283, 2000.
6. Chervitz, S.A. et al., Comparison of the complete protein sets of worm and yeast: Orthology and divergence, *Science*, 282, 2022, 1998.
7. Ficarro, S.B. et al., Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*, *Nat. Biotechnol.*, 20, 301, 2002.
8. Gustin, M.C. et al., MAP kinase pathways in the yeast *Saccharomyces cerevisiae*, *Microbiol. Mol. Biol. Rev.*, 62, 1264, 1998.
9. Schwartz, M.A. and Madhani, H.D., Principles of MAP kinase signaling specificity in *Saccharomyces cerevisiae*, *Annu. Rev. Genet.*, 38, 725, 2004.
10. Bertone, P., Gerstein, M., and Snyder, M., Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery, *Chromosome Res.*, 13, 259, 2005.
11. Zhu, H. et al., Global analysis of protein activities using proteome chips, *Science*, 293, 2101, 2001.
12. Hall, D.A. et al., Regulation of gene expression by a metabolic enzyme, *Science*, 306, 482, 2004.

13. Huang, J. et al., Finding new components of the target of rapamycin (TOR) signaling network through chemical genetics and proteome chips, *Proc. Natl. Acad. Sci. USA*, 101, 16594, 2004.
14. Michaud, G.A. et al., Analyzing antibody specificity with whole proteome microarrays, *Nat. Biotechnol.*, 21, 1509, 2003.
15. Ptacek, J. et al., Global analysis of protein phosphorylation in yeast, *Nature*, 438, 679, 2005.
16. Moffat, J. and Andrews, B., Late-G1 cyclin-CDK activity is essential for control of cell morphogenesis in budding yeast, *Nat. Cell Biol.*, 6, 59, 2004.
17. Huh, W.K. et al., Global analysis of protein localization in budding yeast, *Nature*, 425, 686, 2003.
18. Mewes, H.W. et al., MIPS: A database for protein sequences, homology data and yeast genome information, *Nucleic Acids Res.*, 25, 28, 1997.
19. Ghaemmaghami, S. et al., Global analysis of protein expression in yeast, *Nature*, 425, 737, 2003.
20. Ito, T. et al., A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc. Natl. Acad. Sci. USA*, 98, 4569, 2001.
21. Horak, C.E. et al., Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*, *Genes Dev.*, 16, 3017, 2002.
22. Lee, T.I. et al., Transcriptional regulatory networks in *Saccharomyces cerevisiae*, *Science*, 298, 799, 2002.
23. Xenarios, I. et al., DIP: The database of interacting proteins, *Nucleic Acids Res.*, 28, 289, 2000.
24. Uetz, P. et al., A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*, *Nature*, 403, 623, 2000.
25. Gavin, A.C. et al., Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature*, 415, 141, 2002.
26. Bader, G.D. and Hogue, C.W., BIND — a data specification for storing and describing biomolecular interactions, molecular complexes and pathways, *Bioinformatics*, 16, 465, 2000.
27. Manning, G. et al., Evolution of protein kinase signaling from yeast to man, *Trends Biochem. Sci.*, 27, 514, 2002.
28. Merkel, J.S. et al., Functional protein microarrays: just how functional are they?, *Curr. Opin. Biotechnol.*, 16, 447, 2005.

Section 5

Bioinformatics & Data Analysis

20 Protein Microarray Image Analysis

Minzi Ruan

CONTENTS

INTRODUCTION

As an emerging technology similar to their DNA counterparts, protein microarrays have been increasingly used to study the function and expression of proteins. Based on their applications, there are several types of protein arrays currently used by research scientists, such as functional, semi-quantitative or quantitative, and reverse phase protein microarrays. While semi-quantitative or quantitative and reverse phase protein arrays are mainly used in target validation and clinical research, functional protein arrays are primarily used in biomarker and drug discovery in high throughput screens where automation is a must-have. A protein array image usually consists of one or multiple blocks of arrays. Each block contains grids of spots similar to what is shown in Figure 20.1, with the spot intensity representing the abundance of proteins detected by capturing reagents on the protein array. In functional protein array applications, multiple images representing different binding events may be generated from a single array. For the purpose of comparison, these images often need to be superimposed together to generate one composite image (Figure 20.2).

Compared to DNA microarrays, protein array technology presents additional challenges with image analysis, primarily due to both low signal-to-noise ratio and limited number of abundant protein signal spots to align grids. The variety of array

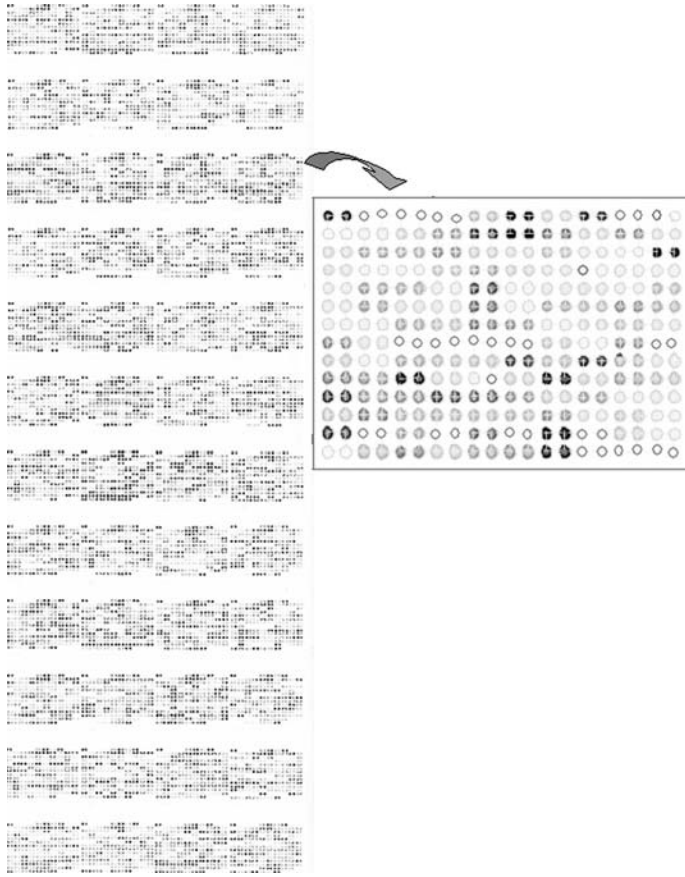


FIGURE 20.1 An image generated from an Invitrogen ProtoArray® array (human protein Microarray NC, v3.0) that contains 5000 proteins on a 1" × 3" nitrocellulose (nc)-coated glass slide. It consists of multiple blocks and each block has 12 rows and 4 columns of spot. A screen capture of one block array shows the image analysis results from MicroVigene automatic segmentation and grid algorithm.

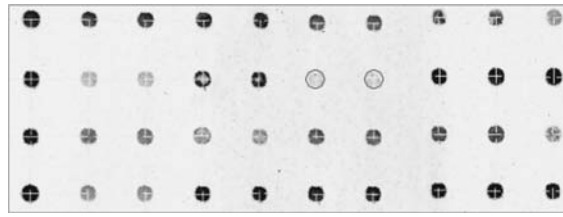


FIGURE 20.2 A small area of a dual-channel composite image generated from the Clontech™ Ab Microarray 500. The experimental sample was labeled with one fluorescent dye, cy3 (green), and the reference control with a different fluorescent dye, cy5 (red). Each colored outline of the spot signal area is the spot boundary found by MicroVigene segmentation algorithm.

formats, spot shapes, and intensity profiles makes it particularly challenging to extract spot signals correctly. In addition, the different substrates, printing mechanisms and protocols, staining/blocking processes, and broad applications result in varied kinds of complex images which make it extremely difficult to develop a silver bullet solution — one algorithm to be applied to all scenarios. Instead, we use object-oriented technology to develop an automated and integrated system that is robust, flexible, configurable, and extensible. Being extensible means that the system can be easily extended to provide customized solutions, support any future needs, and adapt along with this emerging field through proper plug-ins.

In this chapter, we will provide a brief introduction on each of the basic steps in protein array image analysis, with emphasis on potential difficulties and possible corresponding solutions.

IMAGE SEGMENTATION AND SPOT BOUNDARY REFINEMENT

The first and most essential step in image analysis is the feature extraction of the protein microarray image, to measure the signal intensity for each spot. A typical 2-D digital image may contain more than a million pixels and each pixel has a gray value or z-value representing pixel intensity. The signal intensity of any given spot is the collective or statistic measurement of pixel intensities or gray values within the spot. Most images produced by scanners have bright spots on a dark background (i.e., the gray levels in spot pixels are higher than those in background). For easier visualization, those images are often inverted to show dark spots on bright background as shown in Figure 20.3. This image inversion is for display only and does not affect the image analysis, including quantification. The collective or statistic measurements of pixel intensities or gray values in these spots represent the spot signal intensity. Tiff images are currently the most widely used file format for microarray image analysis. An 8-bit format image can contain up to 256 gray levels and a 16-bit image can contain up to 65,536 gray levels. Obviously, the 16-bit format provides much higher dynamic range than the 8-bit format and is well suited for high-quality image analysis.

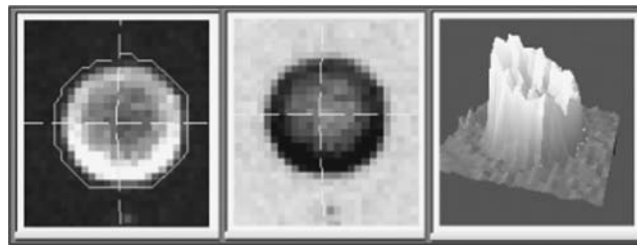


FIGURE 20.3 Digital image and pixel intensity of a sample spot illustrated in 2-D images and 3-D intensity profile. The left image has a bright spot with a dark background. The inverted image (middle) shows a dark spot with a light background. The 3-D image at the right is the spot intensity profile. The z-values are the pixel gray levels of the digital image.

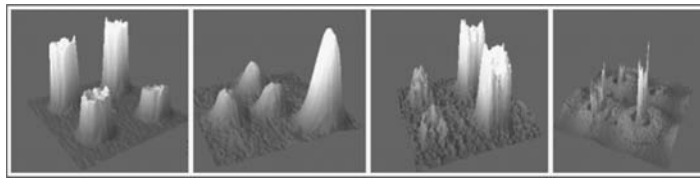


FIGURE 20.4 Different types of spot intensity profiles: (a) cylinder shape, (b) Gaussian bell shape, (c) rectangular prism, and (d) visible amount of halo area as protein bonding also taking place at the spots edges.

Commonly seen basic types of spot intensity profiles include cylinder, Gaussian bell shape, rectangular prism, and cylinder with halo surrounding (Figure 20.4). For example, images from Invitrogen ProtoArray®, Clontech Antibody 500, and Whatman Serum Biomarker Chip all have cylinder shapes, while the nucleic acid-programmable protein array NAPPA from Harvard Institute of Proteomics produces spot images with halo-surrounding shapes. Depending on the type of applications, the signal areas may vary even for the same type of intensity profile. For some applications the mean value of all pixels within a spot is used to represent the spot intensity, while in other applications researchers may be more interested in the total expression measured as volume.

To identify spots, we first need to segment the image to separate signal pixels from background pixels, which is one of the most fundamental and critical tasks in microarray analysis. In order to ensure accurate quantification, proper thresholding algorithms are needed to segment images. In general, there are two types of thresholding methods: global threshold and local threshold. In global threshold, a single threshold is applied to an entire image, while in local threshold, different thresholds are applied to different regions of an image. For example, one of the most widely used global thresholding algorithms developed by Otsu et al.[1] is histogram-based in which the optimal threshold is chosen by maximizing the between-class variance with an exhaustive search. However, global threshold is rarely used in microarray image analyses due to the wide background variations. A popular local thresholding algorithm is Niblack's local mean and standard derivation method[2] in which the local threshold at location (x, y) is determined by equation 20.1.

$$t(x, y) = m(x, y) + k*s(x, y) \tag{20.1}$$

where m(x, y) is the local average, s(x, y) is the local standard derivation, and k is the adjustable constant. The size of the neighborhood is usually set at around twice the average spot size so that it is small enough to preserve the local background variation yet large enough to suppress noise. Each pixel has its own threshold based on its neighborhood background as calculated from equation 1. Thus a higher background will result in a greater threshold and *vice versa*. One potential problem with the local mean and standard derivation methods is that blooming spots may have a strong effect on their neighbor spots particularly when they are close to each

other. In order to eliminate any effect by the neighboring spots, thresholding needs to be adaptive through iteration to exclude any spot pixels in neighborhood background calculation. Another common problem with the local threshold is that the threshold applied may vary quite a bit from pixel to pixel within the same spot. In this case, a regional threshold may be applied to each spot to refine spot boundary. More precisely, each spot instead of each pixel has its own threshold, which can be determined using Otus's thresholding method, minimum error,[3] maximum entropy,[4] or fuzzy logical algorithm[5] from the histogram of a rectangle region around each spot.

BACKGROUND SUBTRACTION

Once the spots are identified, background correction is needed to estimate the true amount of protein expressed. This step is especially critical when measuring small changes in the analyte with high and uneven background. Typically, the background levels in protein microarrays change throughout a slide (Figure 20.5). These changes appear as a variety of forms as shown in Figure 20.6, in which some of the background appears in the blank areas instead of in the spot signal areas or others where spots are cut through. When spots are printed very close to each other and the analytes are captured with high abundance, the spot sizes grow and penetrate into surrounding areas as shown in Figure 20.6D. As a result, the background values surrounding these spots are usually higher than they really should be.

A variety of methods have been used for the background correction. These methods include **global, local, regional,** and **morphological opening** background correction, etc. In **global background** correction, an averaged background from a portion of the image is chosen, usually just outside of the spot array near the image boundary, and applied to all spots in the entire image. Use of a global threshold is only good for a uniform background and thus not suitable for most of microarray image analyses. In **local background** correction, the background intensity is calculated locally in a small region near the spot boundary such as shown in Figure 20.7.

The major weakness of the **local background** correction is that it is sensitive to background noise and contamination. In the **regional background** correction, a

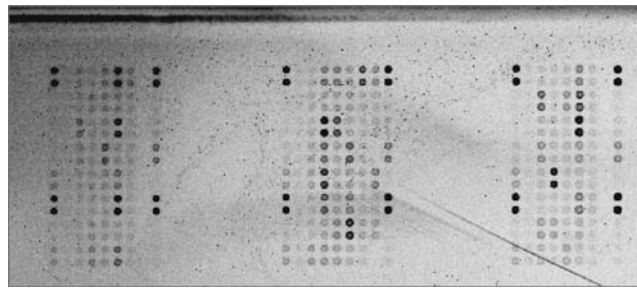


FIGURE 20.5 A typical protein array image that has some spikes, contamination and uneven background. The spot sizes vary and some spots are not lightened up.

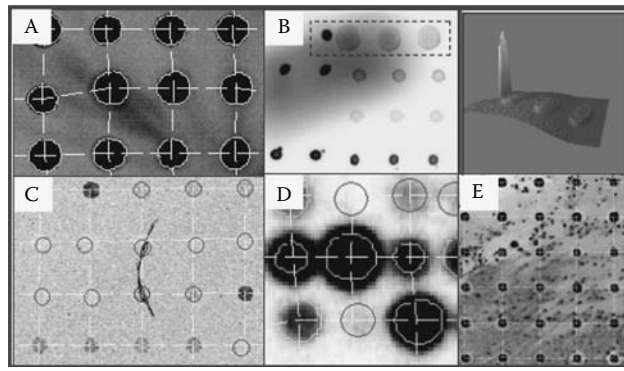


FIGURE 20.6 (A) and (B) An uneven regional background image; (C) background that cuts through the spots; (D) blooming spots; and E) dusty background.

rectangle region around each spot is defined either by the user or by the software such that the size of the net background area is similar to or slightly larger than the signal area. Since the pixels just outside of the boundary can be strongly influenced by the spot signal, a couple of the pixel layers can be defined as the buffer zone between spot pixels and the background. All pixels in spots, buffer zones, and contamination are then excluded from background pixels and the remaining area is called the net background area. The mean, median, mode or certain percentile values of the background pixel intensity histogram can then be used as the spot background. In order to account for spot bleeding effect as described above, a bit less than 50th percentile instead of the median can be used for the background correction. We've found that in many high density images, when 25th percentiles is used, the average

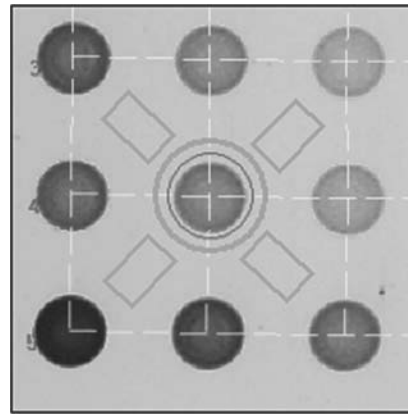


FIGURE 20.7 Different methods for local background correction. The average background value can be calculated either between red and orange circles or within green rectangles. The fuzzy area between green and red circles is called the buffer zone that belongs to neither the spot nor the background.

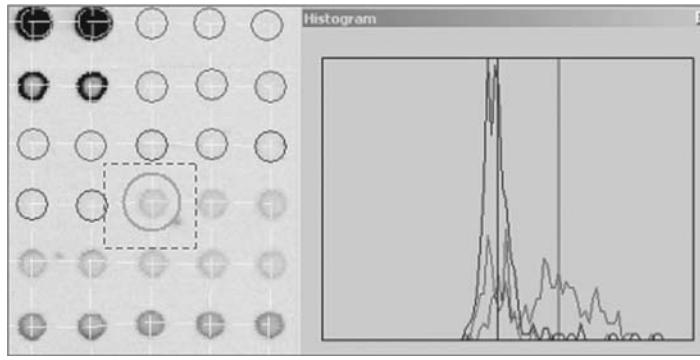


FIGURE 20.8 A histogram that is from a rectangle shown in red dash line around a spot. The size of the rectangle is at least twice its spot diameter. The blue histogram is from the background, the green one from the buffer zoom, and the red one from the spot signal. The red and blue vertical lines are spots mean and 25th percentile background value, respectively. The contamination has little effect on its background value.

of all spot background is close to the global background determined from the region just outside of the spot array near image boundary. As shown in Figure 20.8, this method is very robust for eliminating the background contamination effects.

Using a low percentile value for regional background subtraction allows users to significantly reduce the number of negative intensity spots. To completely eliminate the negative intensity spots, an option is also available in MicroVigene™ to set the minimum intensity as the percent of its background standard derivation as shown in Figure 20.9.

The fourth option for background correction is the **morphological opening**[6] in which the entire protein array image is smoothed by applying a nonlinear local minimum filter (an *erosion*) followed by a nonlinear local maximum filter (a *dilatation*). This method also results in lower background estimates than those from using mean, median, or mode methods.

Two-dimensional curve fitting is the fifth option to correct background. In this method, all spots and buffer zones are first masked out or excluded from the background pixels and the averaged background values near each or several spots can be calculated. A **2-D curve fitting** is then performed with these averaged background values. This is particularly useful to correct the blooming effects as shown in Figure 20.10.

FIGURE 20.9 A method to specify the minimum intensity in background standard deviation. For instance, 0.5 means the minimum intensity will be half its background standard deviation. Any negative value indicates no minimum intensity.

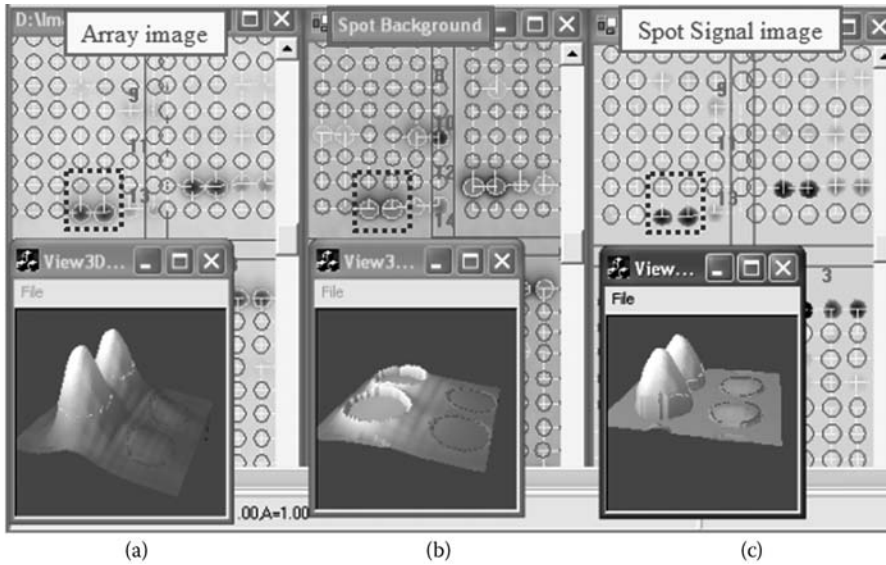


FIGURE 20.10 2-D background curve fitting method used to correct the background that is elevated by neighbor blooming spots. (a) Image without background correction; (b) Background image; (c) Image after background correction.

The last option is not to use the background pixels of the spots but rather use the **blank** or **negative control** spots for background subtraction. The background of each spot is determined as the average background of several nearest blank or negative control spots. This method is suitable for small spot-to-spot spacing arrays and when the pixels surrounding the spot can't be used for background subtraction.

CONTAMINATION REMOVAL

Contamination in the final array image can be introduced in many steps of a protein microarray experiment, including staining, blocking, binding, washing, and even florescence scanning when there is dust in the air. The contamination effects introduced from both staining and washing can usually be minimized using an algorithm with proper background correction as shown in Figure 20.6. On the other hand, artifacts resulting from dust, dye, or other impurities have to be handled with special care. They are usually very bright and small in size and thus can affect the signal reading significantly if not removed properly. Dust effects in the background can be easily detected based on intensity and size. However, the detection of dust effects in a spot signal area can be tricky depending on the location of the dust within a spot and the spot intensity profile (ex. Gaussian bell or hallow cylinder shape).

A simple histogram-based approach can be used to detect the dust effect within a spot area. A bell shape histogram of spot pixel intensity, which is nearly symmetric, can be expected if no dust is present. On the other hand, a very skewed histogram indicates that there is a strong dust effect within the spot (Figure 20.11). A simple

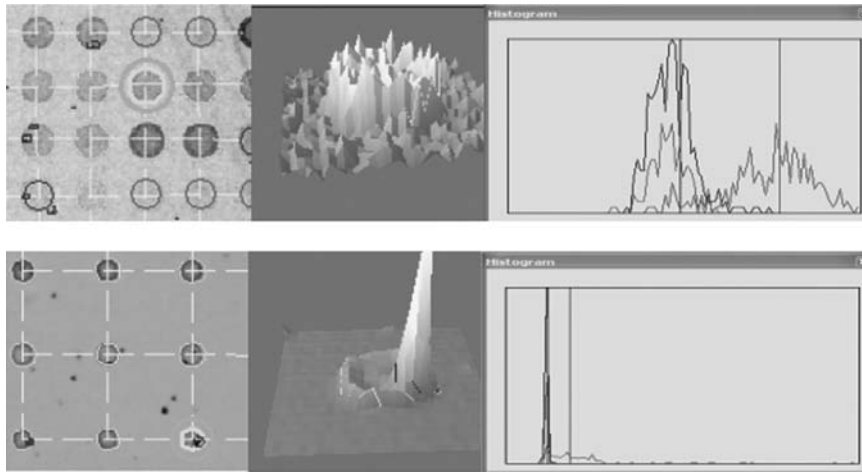


FIGURE 20.11 Histogram patterns. The top set of pictures is from a spot without dust that shows symmetric pixel intensity distribution. The bottom set is from a spot contains a dust. The spot histogram shows some very strong pixels in the red line distribution.

test can be performed on each spot to identify possible dust effects. The degree of dust effect within a spot can be determined by the degree of its histogram skew. If there is a dust effect, all pixels above certain threshold would be classified as potential dust pixels. To be qualified as true "dust," they have to meet the maximum area and minimum mean criteria defined by users, as shown in Figure 20.12.

How should they be corrected once identified? One approach is to simply exclude the contaminated region(s) from the spot. The problem with this simple exclusion is that the same dust can have different effects depending on its location on the spot, e.g., near the center or boundary. A better strategy is to subtract the dust intensity from the signal intensity if the signal can be back filled. A two-dimensional curve fit is needed in order to back fill pixel intensity correctly, as demonstrated in Figure 20.13.

When dust resides partially on a spot, subtracting the dust signal from the spot signal will result in over-subtraction because part of the dust is outside of the spot. In this case, the part of the dust overlapping with the spot needs to be determined before the subtraction.

FIGURE 20.12 Measurements used by MicroVigene algorithm to describe the dust effect. A threshold is used to test if the dust effect exists. For dust segmentation, the dust must have a diameter less than Max Diameter, and have intensity higher than the intensity threshold, which is calculated as the Min Mean multiplied by the image range (maximum intensity — minimum intensity of the image).

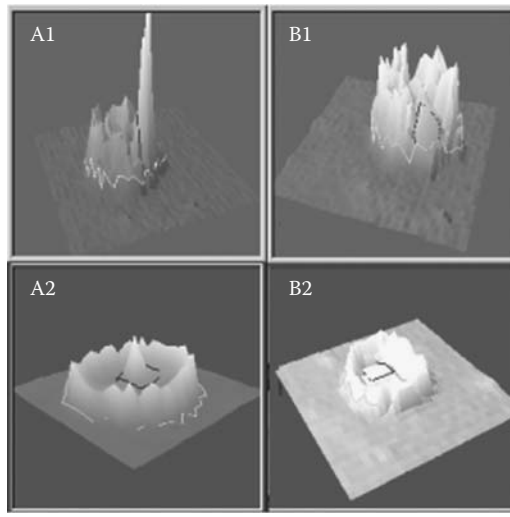


FIGURE 20.13 2-D curve fitting method of background subtraction with backfill that can be used to correct the dust on the spots. (A) 3-D view with dust on top of spot; (B) Spot image in 3-D view after dust be removed.

SPOT QUANTIFICATION AND NORMALIZATION

The purpose of image segmentation, background subtraction, and contamination removal is to quantify the signal that truly represents the signal intensity corresponding to the gene or protein deposited in a given spot. There are several options to quantify signals: **fixed circle**, **adaptive circle**, and **actual spot boundary**. In the **fixed circle** method, every spot in the image is fitted into a circle with a constant diameter. This is easy to implement but not applicable when the shape or size of spots is not uniform across the array. In the **adaptive circle** method, the circle diameter is estimated separately for each spot, which works well as long as all spots are circular. The **actual boundary** method needs to find the actual spot boundary and thus is harder to implement. However, it works well regardless of whether or not spots are circular or the same size.

Typically, signals are quantified by the spot volume (total pixel intensity above background in the spot), mean (volume/signal area), median, or mode. Which measurement best represents the true signal of a spot is largely case-dependent. In most cases, volume, mean, and median give very much the same representation, but mode generally results in larger variation and thus is not used as often as the others. Median is less prone to error caused by contamination, but tends to have large variation, especially when the ratio between its dynamic range and spot size is large.

If the **fixed circle** method is used, quantitation by either volume or mean yields the same result. If the **adaptive circle** or **actual spot boundary** method is used, quantitation by mean normally is more consistent than that by volume. On the other hand, quantitation by volume is more suitable for dealing with larger intensity variations and is also a bit more robust against saturation because the saturated spots are

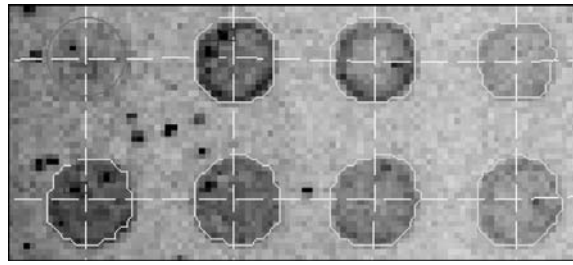


FIGURE 20.14 A protein array image that has some dust-like spikes in the spot pixel area. It can be difficult to determine whether the small spikes inside the spot signal area are true signal or dust effects.

generally bigger than the unsaturated ones. Quantitation using the **fixed circle** method usually gives more consistent measurements, but not as sensitive as the **actual spot boundary** approach. Even with the **fixed circle** method, determining the actual spot boundary is still important for two reasons. First, the circle should be centered at the mass center of a spot, and the actual boundary helps to find the mass center more accurately. Second, knowing where the actual boundary is enhances background subtraction. Another commonly used approach is to trim off a certain percentile of both the low and high end of the intensity distribution of each spot before calculating the mean to reduce or eliminate any outlier effects. This is particularly useful when there are many small "features" in an image and it is hard to determine if they are true signals or artifacts, even aided by the human eye (Figure 20.14).

Current expression protein arrays can be used to qualitatively screen analytes against hundreds to thousands of proteins on a single chip, or to measure the relative fold changes of a few of the same proteins in the experiments versus their corresponding references or controls. In order to obtain accurate quantitation, normalization of both the sample and control signals is essential before they can be compared to each other. The purpose of the normalization is to eliminate the labeling bias introduced by the dyes and variation among the protein arrays themselves. One method of normalization is the median- or mean-based global normalization in which a single normalization factor is applied to all protein spots on an array.[7,8] This global normalization is a linear transformation and is widely used because of its simplicity, but it does not take into account the dependence of intensity and location on dyes. The intensity-dependent variation in dye bias may introduce spurious variation in the collected data, and hence cannot be handled well by linear normalization.

There are a number of nonlinear transformations that take into account the intensity and spatial dependence on dye bias to normalize data.[9,10] Lowess normalization (locally weighted linear regression) is one of the most widely used nonlinear transformations. It merges two-colored data by applying a smoothing adjustment that removes the intensity variations.[11]

Both global and intensity/location-based normalization methods assume that most of the proteins are not differentially expressed between the two samples stained on different array slides, and that for the differentially expressed proteins, the direction of the change is symmetric between the two samples. Besides using all

proteins to normalize the data between two images, you can also use the internal control spots to perform the normalization if the controls are present and they are expected to have no changes between both images. Not all the control spots are suitable for normalization factor determination. For example, the landmark spots are only used to help with the grid placement or the control spots are only used to make sure certain assay is working.

Another method to calculate the protein abundance is to use the correlation slope of their log intensities. Since multiplying the intensity by a factor, which is the same as subtracting a constant in log scale, does not affect the slope, the background subtraction and data normalization become much less, if at all, important. In an ideal case, both the correlation slope and normalized ratio approach should give very similar results. However, in our own experience, the protein abundance is not represented by the correlation slope approach as well as by the normalized log ratio approach. Nevertheless, it is useful as a quality measurement.

OVERLAYING IMAGES

Dual-stained protein microarrays are the most common type of functional protein microarray. In such applications, the experimental and control samples are labeled with different dyes on a single slide and then scanned under different excitation wavelengths to form dual channel images, which eliminates possible slide-to-slide variation. Instead of analyzing two images separately, it is better to find the common grids and spots in the composite images. The dual channel image has to be normalized first between the two colors so that the two channels have similar contributions to identifying the grid and spot segmentations. The signal quantification can then be done in the same spot and background pixel areas with intensities from two channels, which limits the variation from using different signal or background areas but maintains the 16-bit dynamic range. Figure 20.15 shows the overlaid image with Cy3 and Cy5 channels on a single chip. The normalized two channel images are quantified based on the same segmentation.

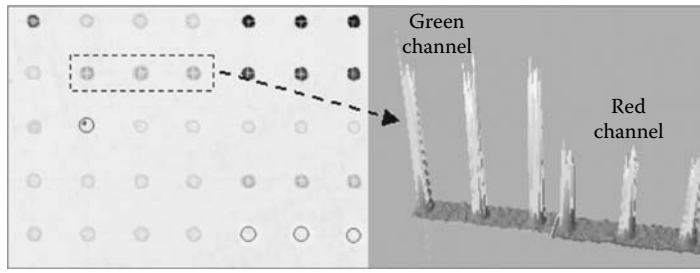


FIGURE 20.15 Signal segmentation that is performed in an overlaid image. With the same signal and background pixels areas, the spot quantification of spots on green and red channels is individually done in their original 16-bit images. The proteins are spotted in triplicate spots. The 2-D overplayed image is on the right side. On the left side, the 3-D intensity profile of triplicate spots are normalized with global normalization at the mean value of spots intensities.

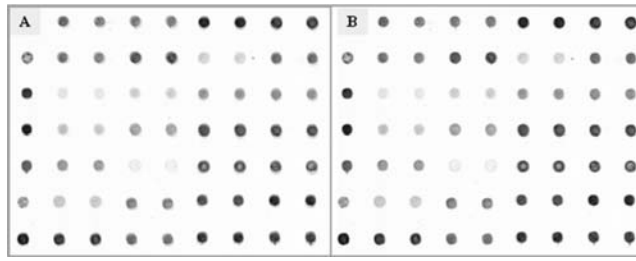


FIGURE 20.16 (A) A simple overlaid image from two different slides, in which corresponding spots are not overlapped with each other perfectly. (B) An overlaid image after performing image analysis on each individual array image, spots registration and warping, in which most of the spots can be overlapped with each other nicely.

In Figures 20.2 and 20.15, the two-channel images are scanned from the same dual-stained chip under two different wavelengths. The array location and orientation on the image are almost identical with very little deviation, and thus the two images can be directly overlaid on top of each other to form a composite image. On the other hand, overlaying two images from two different chips is much more difficult, especially when warping is involved. This is primarily because the global orientation of two arrays on two slides cannot be lined up perfectly with each other, due to the separate handling during the scanning process. Moreover, the spot mass centers may also be shifted from the regular location individually. Therefore, we need to first find all the spots from individual images, then use the spots as control points to register images using thin-plate spline warping,[12] and finally find the grids and spots again on the composite image before quantification (Figure 20.16).

FLEXIBLE GRID PLACEMENT AND ORANGE-PACKED ARRAY

In a more ideal protein microarray image (Figure 20.17), the spot boundaries are well defined, the spots line up straight with even spacing across the entire array, the background is uniform and low, and there is very little or no contamination. In this case, the auto determination of the slide position, block location, and grid lines can be simply accomplished by averaging all the pixel intensities horizontally (or vertically) to get a one-dimensional array. Each peak corresponds to the horizontal (or vertical) grid line as shown on the top of the right side in Figure 20.18. Given the knowledge of grid row/column and spacing, a simple global threshold can be used to identify the peaks and the center of each spot.

In most cases, commercially available slides have good quality control in printing spots in straight lines with even spot spacing, although a certain degree of imperfection can still be introduced during the experimental processes such as staining, chemical and biological processing protocol, and scanning process. On the other hand, many customized arrays and self-printed slides have uneven spot spacing, large spot shifting, and grid crock, which are often due to a bended pin or other

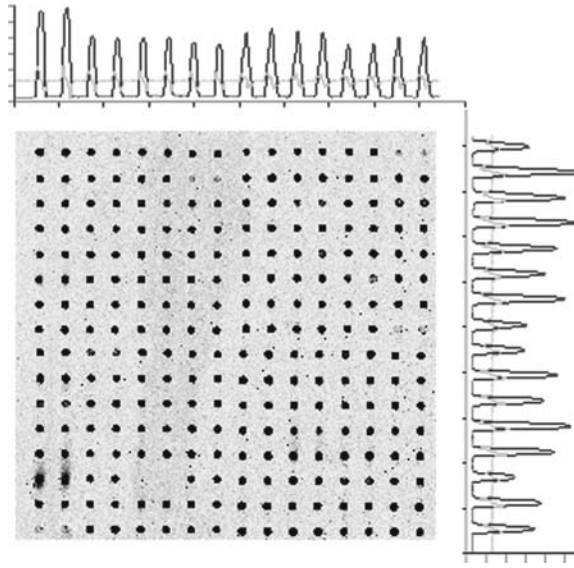


FIGURE 20.17 A clean and straight array image for which the grid can be easily defined. Both the vertical and horizontal histograms have very clear patterns.

array printing defects. In addition, the uneven penetration of the analyte solution into certain 3-D substrates (like nitrocellulose) can shift the spot mass center away from the original printed position. As a result, the spot spacing on one side of the slide can be narrower than that on the other side, particularly when a low quality scanner system is used. However, customized array and self-printed slides provide the flexibility needed for developing new protein array applications and lower the cost for projects that require the use of many slides. Thus it is critical for any robust image analysis software to be able to automatically handle those shifted spots and crocked arrays properly. A good approach to analyze shifted and crocked array images is to use a flexible grid to position spots. An automatic flexible grid placement can be accomplished through an iterative process; that is, grids are used to locate spots, and in turn the spots are used to refine the grids once identified, and these

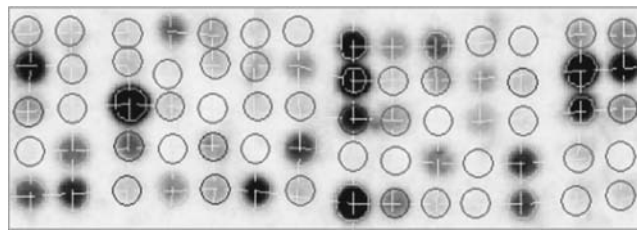


FIGURE 20.18 An algorithm for auto placement of flexible grid in which the array grid lines go through the actual spot location that are shifted and not in the straight lines. The spots were sorted into four categories: blooming (green), good (cyan), poor (purple), and noise (red).

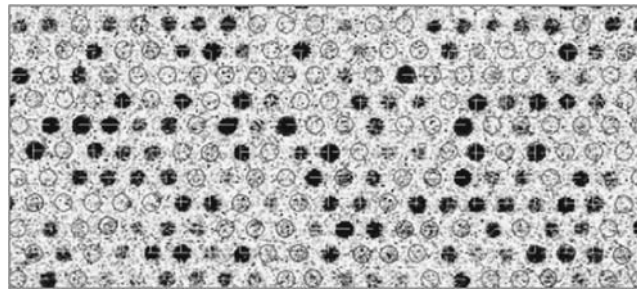


FIGURE 20.19 An Orange-packed image with detected spots and grid placement on top.

two processes should be repeated alternatively with the latest refined grid and spot positions, until updates are no longer required (Figure 20.18).

In order to maximize the print density, protein arrays are increasingly printed in a way that all odd rows shift half the grid space to either left or right side in relative to the adjacent even rows (Figure 20.19). This printing method is often referred to as "orange-pack." Determining the grid for orange pack is a bit more complicated than that for straight regular arrays, but is supported by more and more software packages.

QUALITY CONTROL WITH STATISTICAL ANALYSIS

Quality control or assurance is an important task for all protein array image analyses and needs to be applied to almost every step described above, such as grid placement, spot segmentation, background correction, contamination removal, signal quantification, and outlier flag. Since human eyes can only tell a maximum of 64 shades, it is difficult to visually inspect whether the spot segmentation is accurate or whether a feature is real or just an artifact. In our experience, a 3-D view tool is very valuable for checking the quality of image segmentation by making sure the grid placement and spot boundary are found correctly (Figure 20.20). To better explain how quality control works, a few technical terms need to be defined first as the following:

Solidity: Percentage ratio between the spot area and the minimum convex area. Imagining that a rubber band is placed around the spot, the area inside the rubber band is called the minimum convex area.

Circularity: Defined as (Perimeter^2)/(4πArea), which is 1 for circle and > 1 for any other shapes. Circularity measures the shape roundness as well as the boundary roughness.

Aspect Ratio: Maximum diameter divided by its corresponding minimum diameter. The maximum and minimum diameters are defined as the diameters along and perpendicular to the principal axis, respectively.

Uniformity: Measurement of spot coefficient of variance, which depends on the spot intensity profile. It can also be calculated as maximum intensity divided by its mean or maximum slope divided by its mean.

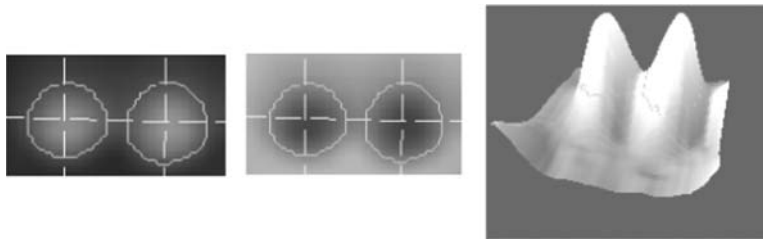


FIGURE 20.20 An image with two spots overlapped with each other. (A) The original image — bright spots on black background; (B) The inverted image — a bit easier to be visualized by eye but still hard to determine where the actual boundary should be; (C) The 3-D view — much easier to tell whether or not the spot boundaries are accurate.

Signal to Noise Ratio: Spot signal intensity divided by its background standard deviation.

Dustiness: Number of dust spots or percent of dust area in a signal area.

A quality index for each spot can be calculated based on a combination of intensity, size, and the measurements described above. The spots can be sorted into multiple categories based on their quality index as shown in Figure 20.18, with bad spots flagged out with yellow crosses. The quality index can be used as a weighting factor in refining the grid placement. Correct grid placement is very critical since a shift by any row or column will result in wrong protein ID and spot pairing. One way to check if there is grid misalignment is to use some positive controls as landmarks. If no landmarks exist and a similar number or more of spots at the block boundary are expressed as the ones in the block, we can calculate the average intensity of boundary spots and compare them with the average of the whole block. A warning or error message should be displayed if grid alignment validation fails.

A simple correlation plot is also valuable to view up- or down-regulated proteins from two channels or two samples slides. Figure 20.21 shows the correlation scatter-plot from a dual-channel composite image. This correlation of scatter-plot feature

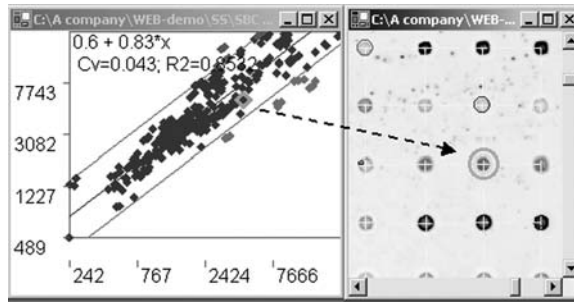


FIGURE 20.21 A correlation scatter plot of two channels. By clicking a point on the scatter plot, the corresponding spots in the images will be highlighted and vice versa.

can also be used to determine which algorithm gives more reliable data and to test reproducibility. If an image is analyzed multiple times with the same settings and same algorithms, the result should be 100 percent reproducible, which can be easily visualized in the correlation plot. All the spots should be perfectly lined up diagonally in the correlation scatter plot. If different algorithms or settings are used, some off-line points should be expected. Comparing off-line points resulting from two different algorithms help us tell which algorithm gives better quality.

DATA VISUALIZATION AND INTEGRATION OF DOWNSTREAM ANALYSIS

The importance of data visualization and integration of downstream analysis is probably not recognized as well as it should be. In our experience, it is extremely useful for the research scientists to be able to seamlessly integrate the downstream analysis tools with the image analysis package. In Figure 20.22, data analysis results for a dual-stained protein array are viewed with four original channels on two slide images. This customized protein array data analysis and visualization module provides users a single-screen view that contains all types of information for final quality examination and results confirmation. Users can quickly identify the changes and information on the screen and verify them with the original image. Incorrect spot

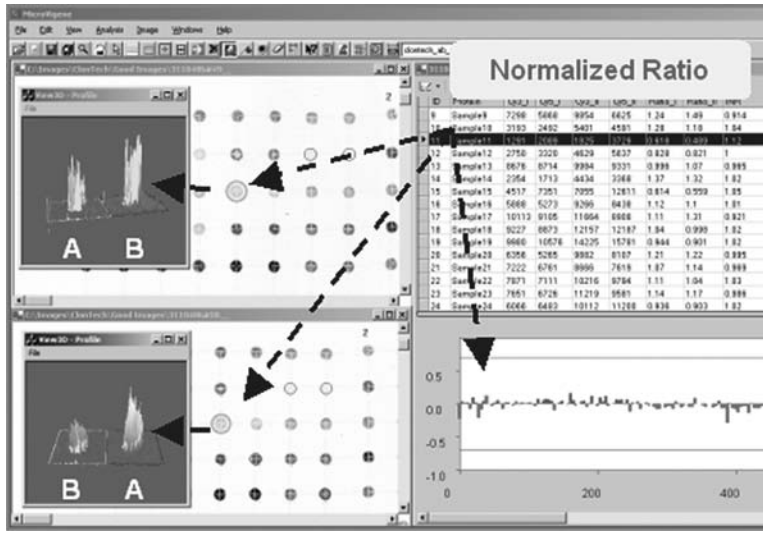


FIGURE 20.22 A screen shot of MicroVigene's data visualization and integration tools. The array application results (Clontech INR) can be easily verified with original images through one of the integrated visualization features in MicroVigene. The final normalized data list and bar graph are directly linked to the original images. If there is modification on segmentation or change flagging on the image window, the output data list and bar graph is updated immediately.

quantification can also be easily corrected by rerunning the image analysis process with different configuration settings or just use the manual tools for a quick edition.

For example, if an outlier spot is the main contributor to the detected significance, it can be flagged out by just clicking the spot on the image to exclude the spot from the quantification. The final normalized ratios as a result of such flag changes on the image will be automatically updated. This feature can be easily expanded or tailored to meet any other application-specific protein arrays via plug-in modules.

SUMMARY

To extract features correctly from millions of data points is an extremely complex process, especially in protein microarray analysis where noise, nonuniform background, and contamination almost always exist. Instead of trying to develop a single algorithm to fit all types of applications, it is better to use an object-oriented and flexible design to develop an automated system with multiple algorithms that is robust, configurable, and with an open architecture (e.g., support plug-ins, Automation, and/or .Net remoting) to meet many different types of protein microarray applications.

REFERENCES

1. Otsu, N., A threshold selection method from grey level histogram, *IEEE Trans. Syst. Man Cybern.*, 9, 377, 1979.
2. Niblack, W., *An Introduction to Digital Image Processing*, Prentice Hall, 1986, pp. 115–116.
3. Kittler, J. and Illingworth, J., Minimum error thresholding, *Patt. Recog.*, 19, 41, 1986.
4. Abutaleb, A.S., Automatic thresholding of gray-level pictures using two dimensional entropy, *Comp. Vis. Graph. Image Process.*, 47, 22, 1989.
5. Liang, K.H. and Mao, J.J., Image thresholding by minimizing the measures of fuzziness, *Patt. Recog.*, 28, 41, 1995.
6. Soille, P., *Morphological Image Analysis: Principles and Applications*, Springer, New York, 1999.
7. Zien, A. et al., Centralization: A new method for the normalization of gene expression data, *Bioinformatics*, 17, S323, 2001.
8. Quackenbush, J., Microarray data normalization and transformation, *Nat. Genet.*, 32, 496, 2002.
9. Yang, Y.H. et al., Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation, *Nucl. Acids Res.*, 30, e15, 2002.
10. Kepler, T.B., Crosby, L., and Morgan, K.T., Normalization and analysis of DNA microarray data by self-consistency and local regression, *Gen. Biol.*, 3, RESEARCH0037, 2002.
11. Cleveland, W.S., Robust locally weighted regression and smoothing scatterplots, *J. Am. Stat. Assoc.*, 74, 829, 1979.
12. Bookstein, F., Principal warps: Thin-plate splines and the decomposition of deformations, *IEEE Trans. Pattern Anal. Mach. Intell.*, 11, 567, 1989.

21 The Analysis of Protein Arrays

Brad Love

CONTENTS

INTRODUCTION

Protein microarrays provide several challenges for data analysis, in part because of their numerous applications. This chapter will focus on data analysis for biomarker discovery using single-channel (non-multiplexed) protein microarray data. However, while the discussion is focused on biomarker discovery, many of the algorithms can

be generalized to other protein microarray applications. For instance, we routinely use the CIP Value normalization algorithm for simple protein-protein and other molecular interactions.

The basic biomarker discovery workflow typically employs up to three steps, with analysis challenges associated with each:

1. Normalization of protein microarray data
2. Identification of differential markers
3. Prediction and classification

Rather than attempting an exhaustive treatment of all potential solutions, I will focus on a selection of approaches that have proven successful in my hands.

NORMALIZATION

The normalization of protein array data is an essential first step when the analysis includes two or more protein arrays. Variability of observed data on arrays can be classified as originating from two sources: interesting and obscuring. Interesting variability can be attributed to the difference in samples, whether it is a difference from patient to patient, between cancer and normal, or between experimental sample and negative control. Obscuring variability is due to sample prep, differences in printing, scanner settings, differences in labeling etc. Normalization attempts to correct for obscuring variability, variability that is not of interest, so that we can measure and attempt to understand the interesting variability — the reason for the experiment.[1]

Since normalization methods attempt to correct for this obscuring variability, all algorithms start out with an assumption about what data should look like if there were no obscuring variability present in the observed data. The algorithm is designed to correct the observed data to make this assumption true and hence remove the obscuring variability. It is vitally important to understand what assumptions are being made with the normalization technique that is used. We will next describe two methods for normalizing data from single dye (single-channel) protein array experiments. For both normalization algorithms we discuss the associated assumptions, advantages, and disadvantages.

CIP Value

CIP Values (Chebyshev's Inequality Probability Values) provide a method to identify protein hits (signals that are significantly different from the negative controls) on a single array. This method requires the presence of negative controls on the array. It makes no other assumptions about them other than that they are true biological, meaningful, negative controls for the experiment.

Ideally, if we understand how true negative controls behave, we can look at signals coming from a protein, and determine (with some measure of probability) whether it is (a) behaving like a negative control, and hence *not* a hit on the array, or is (b) behaving different than the negative controls and hence *is* a hit.

While there are many methods to calculate probabilities, most rely on assumptions about the distribution of negative controls. Normal distributions are typically assumed. However, if these assumptions are incorrect, the probabilities will be erroneous as well.

Chebyshev's inequality[2] makes no assumptions about the overall distribution of the negative controls, the result given as,

$$P\left(\frac{X-\mu}{\sigma} \geq k\right) \leq \frac{1}{k^2} \tag{21.1}$$

assuming that $k > 1$ (meaning that X is at least one standard deviation larger than the mean), $E(X) = \mu$ (μ is mean the random variable X) and $Var\ (X) = \sigma^2$ (σ^2 is the variance of the random variable X). Note that with a bit of mathematical manipulation it can be restated as,

$$P(X \geq k) \leq \frac{\sigma^2}{(k-\mu)^2} \tag{21.2}$$

which assumes that $k > \mu + \sigma$.

Chebyshev's inequality is therefore an upper bound on the true probability and requires no prior knowledge about how the negative controls behave in order to calculate p-values. Accordingly, these p-values are conservative.

Using these results we can build the following definition,

$$\text{CIP Value} = \begin{cases} \left(\dfrac{s_{neg}}{X - \bar{X}_{neg}}\right), & X \geq \bar{X}_{neg} + s_{neg}^2; \\ 1, & \text{otherwise.} \end{cases} \tag{21.3}$$

where X is the signal for a particular protein probe and define

$$\bar{X}_{neg} = \frac{1}{n_{neg}} \sum_{i=1}^{n_{neg}} X_{i,neg} \tag{21.4}$$

as the observed average of all of the negative controls (here $X_{i,neg}$ is the i^{th} the observed negative control) on the protein chip and

$$s_{neg}^2 = \frac{1}{n-1} \sum_{i=1}^{n_{neg}} (X_{i,neg} - \bar{X}_{neg})^2 \tag{21.5}$$

is the observed variance of the negative controls on the protein chip.

This is considered a normalization technique because for each protein signal on a chip the probability of not being a negative control is calculated based on the behavior of the negative controls on that chip. This effectively normalizes the values for each protein by the negative controls allowing for comparisons of CIP values between chips.

The Assumptions

The fundamental assumption is that on each array we have an unbiased and relatively large number of observations of true biological negative controls. These negative controls then act as a measurement of the obscuring variability on each protein array.

The most important implication of this assumption is that the arrays are designed for this before the experiment is done. Not only must proper negative controls be identified for the array/experiment, they must be be replicated many times across the entire array. We typically use several hundred negative controls on each of our arrays. Depending on the density of the chip, this can commit a substantial percentage of the array surface.

The Upside

The algorithm is fairly simple to apply and can easily be performed using commonly available software programs such as Microsoft Excel.

The algorithm's results are easy to interpret. The CIP Value is the probability you would be wrong if you said that the signal from the protein is behaving like a negative control. The smaller the value, the more likely the hit is "real."

There are almost no statistical assumptions made in this algorithm. For example, you don't need to assume that the negative controls have any particular statistical distribution. The method is therefore robust to statistical nitpicking about assumptions and how they affect the overall results.

Chebyshev's inequality gives a maximal probability, meaning that the true probability is less than what is actually calculated (to calculate the true probability we would have to know the true distribution of negative controls). Therefore, the hits identified are likely true positives.

CIP values are also a great tool when the experiment can be summed up in a single chip. An example of this is protein–protein interaction experiments on protein arrays. This type of experiment can typically be performed with single samples to observe how one protein interacts with potentially thousands of other proteins.[3] Similarly, this can be a useful application when profiling antibodies to look for a cross-reactivity.[4]

The Downside

This algorithm relies on using true biological negative controls. Properly defining and identifying a true negative control is not always a trivial task. For instance, are blank, non-printed spots on a protein array suitable for negative controls or does something have to be printed? Does printing just buffer constitute a real negative control because buffer does not have protein in it? If actual proteins are required, which ones and at what concentration(s)? Once a proper negative control is identified, it is essential that many replicates are used across the array to get accurate estimates of the average and variance of the negative controls.

This method can be susceptible to outliers within the negative controls. An increase in the standard deviation of the negative controls results in a squared increase in the CIP value. For example, a 10% increase in the standard deviation results in a 21% increase in the CIP value. It then becomes important to understand where outliers come from and what they are. Some outliers can arise from bad printing, mis-acquired spots, dust, scratches, bleedover of neighboring spots, bad local background, etc. These problems typically result in overestimating the true signal of the negative control, usually resulting in a signal that is orders of magnitude larger than most. This effect ripples through the algorithm by increasing the value of both the average negative control and standard deviation. This forces CIP values of 1 for protein signals that are no longer one standard deviation larger than average. This is not a failure of the algorithm per se. The algorithm requires a good estimation of the true average and variance of the negative controls and outliers that are not unbiased measures of the negative controls break this assumption. Buffering the algorithm to outliers would improve the robustness of this approach.

Although false positives are not an issue with this algorithm, false negatives can be. While CIP values proved a maximal p-value, the true p-value is likely less than that, and in some cases it could be significantly lower. That translates to a problem with false negatives, i.e., some proteins will not be considered hits when they really are. This is a general problem when considering false positives and false negatives. They typically work against each other; reducing one occurs at the expense of increasing the other.

QUANTILE NORMALIZATION

Quantile normalization first started to make an impact on the DNA array analysis community through several early publications,[5–7] though the basic idea derives from an unpublished manuscript 2 years before.[8] The major use of quantile normalization has been for analysis of Affymetrix DNA array data, where it outperformed other normalization techniques.[6] It is a component of the "robust multichip average" algorithm, which was commonly used to analyze Affymetrix data.

The basis for this approach is to set the distributions of the signals on each array identical for every array in a set of arrays. This can be measured by creating a quantile–quantile plot (like Figure 21.1), sometimes referred to as a Q–Q plot, for every pair of arrays in the analysis. If the distributions of the signals are the same, this plot should give a straight diagonal line. If all the chips have the same distribution, then all pairwise Q–Q plots would look like a diagonal line. The normalization procedure forces this assumption to be true.

Quantile Normalization Algorithm

1. Sort each array from smallest to largest.
2. Calculate the median of the smallest value across each array, calculate the median of the second smallest value across each array, etc. through to the largest value on each array.
3. Replace the smallest value on each array with the median (across arrays) smallest value, replace the second smallest value on each array with the median (across arrays) second smallest value, etc. through to the largest value on each array.

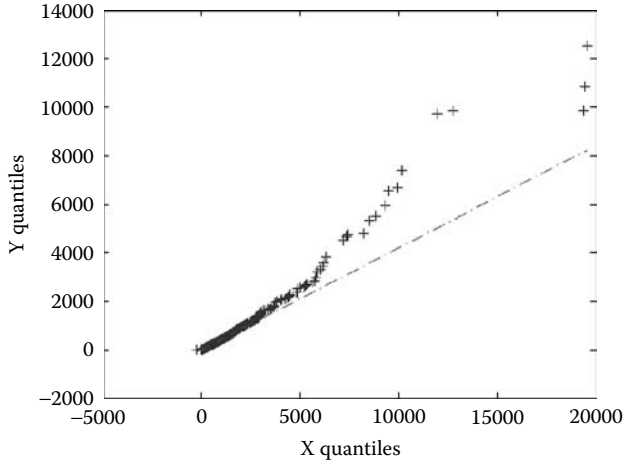


FIGURE 21.1 A Q–Q plot of two protein microarrays from the same experiment the same group. Note that the data on the right side of the plot do not fall on the dash-dotted line. The dash-dotted line represents the line in which the distributions are equal.

It is important to note that this algorithm does not give the same protein the same signal value, rather it gives the same ordered value on each array the same value. It is blind to what the actual protein is when reassigning the signal. It is also worth noting that this version is an implementation that is slightly different from other sources.[5–8] These references use the average across arrays, while here we suggest using the median.

In Figure 21.2, we can see the overall effect of the quantile normalization on individual arrays. Here, results from 127 arrays are plotted in a distribution graph. Clearly, the signal distribution varies from chip to chip. The heavy dashed line shows the resulting distribution of the quantile normalized data. Quantile normalization forces the signals from each array into this distribution.

The Assumptions

This algorithm assumes that the overall distribution of signals should be the same from chip to chip within an experiment. This, in turn, implies that any shifts observed from this assumption are explained by obscuring variability or bias.

The Upside

This method has been shown to have superior qualities compared with other normalization methods.[5] Specifically, it was shown that this method provides superior performance when considering bias as well as variance. In addition, it was judged to be the quickest, in a computational sense, to perform. Largely because of these features, quantile normalization has been heavily used for DNA array analysis.

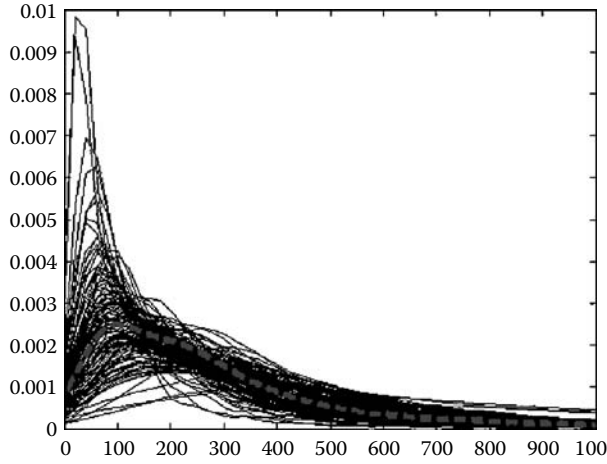


FIGURE 21.2 The heavy dashed line is the resulting quantile normalized distribution all of 127 chips worth of data; the other lines are the distributions of the individual 127 protein arrays. See color insert following page 236.

This normalization avoids preconceived notions of what the correct distribution looks like, and instead estimates the most likely distribution based on all of the observed data.

The Downside

This algorithm can be more difficult to apply to large data sets without appropriate software. Also, the assumption that the distribution needs to be the same from sample to sample may not be appropriate.

IDENTIFICATION OF MARKERS

Biomarker discovery with functional protein microarrays presents particular data analysis challenges. A typical experiment may involve tens or hundreds of disease or treated samples and matched controls. In many cases, such as profiling sera for autoantibodies, markers may exist individually at prevalences of only 20–30%. For this reason, approaches like t-test[10] or nonparametric equivalents such as the Mann-Whitney test[11] don't work well for this application. Biomarker data analysis can be divided into two phases. The first is identification of individual markers (a step that is sometimes skipped) and the second is combining these markers for prediction and classification. This section addresses two approaches for the identification of individual markers.

The "Bayesian prevalence" approach defines a method to estimate prevalence of a particular marker, calculate confidence intervals for the true prevalence, and provides an approach to assign statistical significance for a marker in two different populations.

M statistics[9] use rank ordered signals to compare groups. P-values are then assigned through a combinatorics approach.

For the purposes of this section, I intentionally disregard the measurement or calculation used to assess a hit (signal, signal-background, signal or background, CIP value, etc). In practice, the user needs to determine the most appropriate method for the given experiment.

Bayesian Prevalence

Although Bayesian theory tends to be fairly math heavy we will limit our discussion to the basic formulae and how they can be used for our purposes. While not exhaustive, this treatment should serve as a useful basic introduction to the application of Bayesian concepts to protein array experiments.

Suppose that we have n protein arrays, and for a particular protein we have x arrays where the protein is considered a hit. Then, using a Bayesian technique to estimate prevalence, given an uninformative prior, we can show that p (the prevalence) has a beta distribution (with parameters x + 1 and n − x + 1) given by,

$$f(p|x,n) = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} p^{(x+1)-1}(1-p)^{(n-x+1)-1} \tag{21.6}$$

where,

$$\Gamma(x) = x! = x*(x-1)*(x-2)\ldots*2*1 \tag{21.7}$$

We can use the beta distribution to calculate the mean of the distribution and show that a Bayesian estimate of prevalence of a marker is given by (note that the p has a hat on it to denote that it is an estimate),

$$\hat{p} = \frac{x+1}{n+2} \tag{21.8}$$

In fact, to estimate the prevalence all you need is equation (21.8). With this formula for prevalence estimates you will note that the range of answers is $\frac{1}{n+2}, \frac{2}{n+2}, \ldots, \frac{n}{n+2}$ and $\frac{n+1}{n+2}$ so you will never estimate prevalence to be either 0 or 1, though as the number of protein arrays increases the lower and upper estimates will get closer to 0 and 1.

Knowing the resulting probability distribution of estimator of prevalence allows us to calculate confidence intervals of the true prevalence for each marker. As can be seen in Figure 21.3, each probability distribution gives the likelihood of possible prevalences. To calculate a confidence interval, the middle area of the curve must be calculated, which can be very complicated without software.

Microsoft Excel can be used for this purpose. Say we have $n = 10$ protein arrays and observe $x = 7$ hits. We can then use (21.8) to estimate the prevalence as $\hat{p} = 66.67\%$. If we would like a 95% interval we need to determine what values give the middle 95%. To do this we must determine the 2.5 and the 97.5 percentiles (1 – .95 = .05).

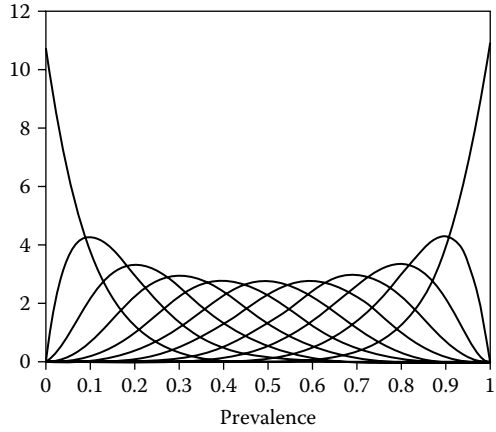


FIGURE 21.3 A plot of all the possible distribution for prevalence for a 10-chip experiment; note that there are 11 distributions. The distributions are observed left to right for 0 through 10 hits.

We then equally split this .05 to the extremes of the distribution, i.e., .025. Thus, the lower bound of the confidence interval is the 2.5 percentile and the upper bound of the confidence interval is the 97.5 percentile.

In Excel, the lower bound of the confidence interval can be calculated by using

$$= \text{BETAINV}(\alpha/2, x+1, n-x+1) \tag{21.9}$$

and the upper bound of the confidence interval is given by

$$= \text{BETAINV}(1-\alpha/2, x+1, n-x+1) \tag{21.10}$$

where your level of confidence is equal to $1 - \alpha$ (α denotes the acceptable probability of being wrong when saying the true prevalence is in the range of the confidence interval). Continuing our example from the previous paragraph where $n = 10$, the number of protein arrays and $x = 7$, the number of hits, to calculate a 95% confidence interval we get 39.03% to 89.07%. All of the 95% and 99% confidence intervals are calculated in Table 21.1, for this example to give an idea of the values for confidence intervals.

From Table 21.1 we can see that both the 95% and the 99% confidence intervals are very large. The 95% confidence interval (the small interval) covers more than 28%, while the 99% confidence interval covers a little over 38%. To reduce the width of these confidence intervals would require either reducing the level of confidence or increasing the number of protein arrays used in the experiment.

With estimates of prevalence and an ability to calculate confidence intervals, the last task is to calculate p-values for differences between states (e.g., before and after treatment, or normal vs. disease). The p-value corresponds to the area that is jointly under both probability distributions. Figure 21.4 shows an example where we have

TABLE 21.1
Confidence Intervals Calculated for an Experiment with 10 Protein Arrays, Calculating Both 95% and 99% Confidence Intervals for All Possible Numbers of Hits

Hits	$\hat{p}$	Lower Bound 95% Confidence	Upper Bound 95% Confidence	Lower Bound 99% Confidence	Upper Bound 99% Confidence
0	8.33%	0.23%	28.49%	0.05%	38.22%
1	16.67%	2.28%	41.28%	0.98%	50.86%
2	25%	6.02%	51.78%	3.33%	60.85%
3	33.33%	10.93%	60.97%	6.88%	69.33%
4	41.67%	16.75%	69.21%	11.45%	76.68%
5	50%	23.38%	76.62%	16.93%	83.07%
6	58.33%	30.79%	83.25%	23.32%	88.55%
7	66.67%	39.03%	89.07%	30.67%	93.12%
8	75%	48.22%	93.98%	39.15%	96.67%
9	83.33%	58.72%	97.72%	49.14%	99.02%
10	91.67%	71.51%	99.77%	61.78%	99.95%

a 10 vs. 10 protein array experiment. In one of the two groups we observe two hits, while in the other we have seven hits. The shaded area under both distributions identifies a p-value of .1262.

Calculating the p-value requires complicated mathematics: First find the value in which the two probability distribution are equal. Call this p^*. Then calculate the

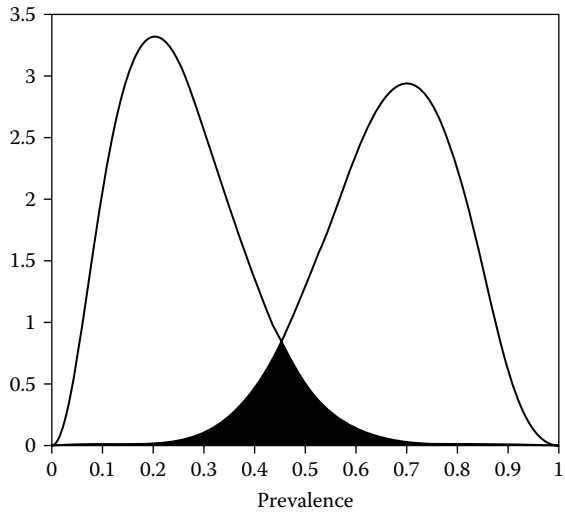


FIGURE 21.4 A plot of the probability distributions for two hits versus seven hits with 10 protein arrays. The p-value of the difference is the shaded area under the two plots, which is .1262.

area from 0 to p^* for the probability distribution for the larger number of hits, calculate the area under the curve from p^* to 1 for the smaller of the two hits, then add these two areas together to get the p-value.

Finding the exact point where both curves meet involves solving the following problem for p,

$$\frac{\Gamma(n_1+2)}{\Gamma(x_1+1)\Gamma(n_1-x_1+1)}p^{x_1}(1-p)^{n_1-x_1} = \frac{\Gamma(n_2+2)}{\Gamma(x_2+1)\Gamma(n_2-x_2+1)}p^{x_2}(1-p)^{n_2-x_2} \tag{21.11}$$

where n_1 is the number of protein arrays in group 1, x_1 is the number of hits for the protein of interest in group 1, n_2 is the number of protein arrays in group 2 and x_2 is the number of hits for the protein of interest in group 2.

The solution to this equation falls into two different classes: (1) when $n = n_1 = n_2$, i.e., when the number of protein arrays is equal in each group, or (2) when $n_1 \neq n_2$, i.e., when the number of protein arrays is not equal. We will look at only the simpler of the two cases (when the number of protein arrays is equal). The solution is given by,

$$p^* = \frac{\sqrt[x_1-x_2]{\binom{n}{x_2}/\binom{n}{x_1}}}{1+\sqrt[x_1-x_2]{\binom{n}{x_2}/\binom{n}{x_1}}} \tag{21.12}$$

where for ease of notation we will assume that $x_2 > x_1$ and recall that

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{21.13}$$

To do this in Microsoft Excel you would enter in the following into the cell,

$$\begin{aligned} &= \text{POWER}(COMBIN(n,x_2)/COMBIN(n,x_1),(1/(x_1-x_2)))/ \\ &(1+\text{POWER}(COMBIN(n,x_2)/COMBIN(n,x_1),(1/(x_1-x_2)))) \end{aligned} \tag{21.14}$$

Note that careful attention to parenthesis is required, but once equation (21.12) is entered, it can be cut and pasted as needed.

Now that we know the exact place where the two probability distributions cross we can calculate the area under the curve using calculus, i.e.,

$$\begin{aligned} \text{p}-\text{value} = &\int_0^{p^*} \frac{\Gamma(n+2)}{\Gamma(x_2+1)\Gamma(n-x_2+1)} t^{x_2}(1-t)^{n-x_2}dt \\ &+\int_{p^*}^1 \frac{\Gamma(n+2)}{\Gamma(x_1+1)\Gamma(n-x_1+1)} t^{x_1}(1-t)^{n-x_1}dt \end{aligned} \tag{21.15}$$

This calculation can be performed in Excel:

$$\begin{aligned} &= \text{BETADIST}(p^*, x_2+1, n-x_2+1) \\ &+(1-\text{BETADIST}(p^*, x_1+1, n-x_1+1)) \end{aligned} \tag{21.16}$$

Table 21.2 shows the calculation for all p-values for a 10 vs. 10 protein array experiment for all possible hit comparisons. As can be seen in the table, when using 95% confidence in a 10 vs. 10 protein array experiment, a marker will be considered statistically different between two groups if there more than seven hits between them (i.e., 0 vs. 7, 0 vs. 8, 0 vs. 9, 0 vs. 10, 1 vs. 8, 1 vs. 9, 1 vs. 10, 2 vs. 9, 2 vs. 10 and 3 vs. 10) or if there are 0 vs. 6 or 4 vs. 10, note from the table these are all of the pairs with p-values less than 0.5.

The Assumptions

The assumptions for this algorithm are fairly light. Here we assume that the data collected for a particular marker across all of the samples within each group are

TABLE 21.2
p-Values for a 10 vs. 10 Protein Array Experiment

	1	2	3	4	5	6	7	8	9	10
0	.6145	.3793	.2301	.1355	.0765	.0410	.0204	.0091	.0035	.0010
1		.7013	.4746	.3081	.1904	.1107	.0595	.0287	.0117	.0035
2			.7380	.5201	.3476	.2180	.1262	.0654	.0287	.0091
3				.7562	.5421	.3648	.2266	.1262	.0595	.0204
4					.7639	.5488	.3648	.2180	.1107	.0410
5						.7639	.5421	.3476	.1904	.0765
6							.7562	.5201	.3081	.1355
7								.7380	.4746	.2301
8									.7013	.3793
9										.6145

Note: If the number of hits is equal in each group, then the p-value is 1. The left side is the smaller of the two hits and across is the larger of the two hits.

unbiased and randomly selected from the population that is being measured. In other words, we assume that there is no sampling bias in the data collected for the population for which we are making inferences.

The Upside

Despite the rather complicated mathematics, everything can be done in Microsoft Excel. This makes a sophisticated task relatively easy and within the grasp of anyone who has access to this common software.

This method allows us to estimate prevalence and calculate corresponding confidence intervals even when we observe no hits (or all hits). On the surface this may not seem particularly valuable, but in reality, this represents an improvement on the more typical approach of estimating prevalence. That approach, often called the frequentist approach, is usually taught in basic statistics courses, along with associated calculations for confidence intervals and p-values. This method typically requires that the number of hits and non hits in each group be more than five. For low prevalence markers or experiments with low numbers of arrays this assumption can easily be violated.

The Downside

As odd as it sounds, statistics can be split into two opposing views of what exactly a probability is. That world is split into two groups: the Frequentists and the Bayesians. Perhaps not surprisingly, these groups sometimes have differing opinions about the other's techniques. Bayesians tend to be the minority among statisticians, and so their methods are not as widely published.[15]

Even though all of the techniques presented here can be done in Excel, careful attention to entering the commands is needed to make sure the results are correct.

M STATISTIC

M statistics are based on comparing the order of values between two groups, and determining the chance of that order happening randomly. M statistics will test for differences in prevalence no matter what the difference is.

M statistics are calculated by order. For example, a first order M statistic for group 1 counts the number of observations from group 1 that are larger than the largest observation in group 2. A second order M statistic for group 1 counts the number of observations from group 1 that are larger than the second largest observation in group 2. The math formula for this can be expressed as,

$$M_j^{(i)} = \sum_{k=1}^{n_j} 1(x_{j,k} \geq x_{3-j,(i)}) \tag{21.17}$$

This is the ith order M statistic for group j, the data given by $x_{j,k}$ is data from the jth group the kth observation. In this notation $x_{j,(k)}$ refers to data from the jth

group the kth largest observation from group j, n_j is the number of protein arrays in the jth and finally,

$$1_{(x_{j,k} \geq x_{3-j,(i)})} = \begin{cases} 1, & \text{if } x_{j,k} \geq x_{2-j,(i)}; \\ 0, & \text{otherwise.} \end{cases} \tag{21.18}$$

Using this formula a first order M statistic for group 1 would be defined as

$$M_1^{(1)} = \sum_{k=1}^{n_1} 1_{(x_{1,k} \geq x_{2,(1)})} \tag{21.19}$$

As can be seen, this counts the number of data points $x_{1,k}$, where $k = 1,2,\ldots,n_1$ in group 1 that are larger than the largest data point $x_{2,(1)}$ in group 2. A 3rd order M statistic for group 2 would be given by

$$M_2^{(3)} = \sum_{k=1}^{n_2} 1_{(x_{2,k} \geq x_{1,(3)})} \tag{21.20}$$

This counts the number of data points $x_{2,k,}$ where $k = 1,2,\ldots,n_2$ in group 2 that are larger than the 3rd largest data point $x_{1,(3)}$ in group 1.

In general, the larger the resulting value, the more likely that it is significant. To calculate an actual p-value for the M statistic we use combinatorics as follows:

$$P\left(M_j^{(i)} = m\right) = \frac{\begin{pmatrix} n_1 + n_2 - m - i \\ n_2 - i \end{pmatrix}\begin{pmatrix} m + i - 1 \\ i - 1 \end{pmatrix}}{\begin{pmatrix} n_1 + n_2 \\ n_1 \end{pmatrix}} \tag{21.21}$$

This gives the probability of randomly seeing a particular value for the $M_j^{(i)}$ that is being used. To calculate this in Microsoft Excel you would enter in the cell,

$$\begin{aligned} &= \text{COMBIN}(n_1 + n_2 - m - i,\ n_2 - i) * \text{COMBIN}(m + i - 1,\ i - 1) \\ &\ \ /\text{COMBIN}(n_1 + n_2, n_1) \end{aligned} \tag{21.22}$$

To calculate the p-value we calculate,

$$\text{p} - \text{value} = P\left(M_j^{(i)} \geq m\right) = \sum_{k=m}^{n_j} P\left(M_j^{(i)} = k\right) \tag{21.23}$$

TABLE 21.3
M Statistic p-Value Table

	0	1	2	3	4	5	6	7	8	9	10
1	1	.5000	.2368	.1053	.0433	.0163	.0054	.0015	.0004	.0001	.0000
2	1	.7632	.5000	.2910	.1517	.0704	.0286	.0099	.0027	.0005	.0001
3	1	.8947	.7090	.5000	.3142	.1749	.0849	.0349	.0115	.0027	.0004
4	1	.9567	.8483	.6858	.5000	.3250	.1849	.0849	.0349	.0099	.0015
5	1	.9837	.9296	.8251	.6750	.5000	.3281	.1849	.0894	.0286	.0054
6	1	.9946	.9714	.9151	.8151	.6719	.5000	.3250	.1849	.0704	.0163
7	1	.9985	.9901	.9651	.9106	.8151	.6750	.5000	.3250	.1517	.0433
8	1	.9996	.9973	.9885	.9651	.9151	.8251	.6858	.5000	.2910	.1053
9	1	.9999	.9995	.9973	.9901	.9714	.9296	.8483	.7090	.5000	.2368
10	1	1	.9999	.9996	.9946	.9946	.9837	.9567	.8947	.7632	.5000

Note: The left side of the plot is the order of the M statistic and across the top is the observed value. The values in the table are the corresponding p-values to four digits of significance.

To perform this calculation using Microsoft Excel just requires summing up over the appropriate cells and using the previous Excel formula.

In Table 21.3 we extend the example of having a 10 vs. 10 protein array experiment, calculating the p-values for all possible value of M statistics for all orders. Using this table we can, for example, determine the p-value for $M_1^{(1)} = 6$, as .0054. Similarly, $M_2^{(4)} = 5$ gives a p-value of .3250. Note that there is no difference in using the table for $M_1^{(i)}$ *or* $M_2^{(i)}$.

The Assumptions

Similar to the assumptions for the Bayesian Prevalence method, we require that the data from both groups be random, independent unbiased samples from the populations that are being tested. It is assumed that the sample can be compared and considered together, i.e. that the data has been normalized.

The Upside

This approach provides a method to determine hits (by being larger than some value in the other group), as well as an associated p-value.

This method can be modified slightly to make it more powerful. When searching for unique markers, a first order M statistic is most appropriate. However, the most appropriate order M statistic is less clear when searching for the largest difference in prevalence between two populations. For this we can employ a dynamic calculation called the "minimum M statistic." For each protein we calculate M statistics and associated p-values for each order. We then select the M statistic that gives the minimum p-value (and report the order and the p-value).

The Downside

The downside to this approach is common to almost all two group comparison methods for high density arrays; the multiple testing problem. False positives occur because of the large numbers of proteins being tested as hits. For instance, even at a p-value of 0.005, a 2,000 protein array would be expected to randomly and incorrectly identify 10 = 0.005*2,000 proteins as hits. These problems typically arise when the signals for hits are low. For this reason we have considered modifying the M statistic by adding additional restrictions to equation (21.17). We have used two specific modifications. The first is to count the value of the M statistic over a certain signal threshold. The second is counting only values more than some static amount of the order. Specifically the new formula for the modified M statistic is given as,

$$M_j^{(i)}(\text{above, between}) = \sum_{k=1}^{n_j} 1_{(x_{j,k} \geq x_{3-j,(i)} + \text{between})} \times 1_{(x_{j,k} \geq \text{above})} \tag{21.24}$$

Note this is the product of two indicator variables where the second indicator is defined as

$$1_{(x_{j,k} \geq \text{above})} = \begin{cases} 1, & \text{if } x_{j,k} \geq \text{above;} \\ 0, & \text{otherwise.} \end{cases} \tag{21.25}$$

This variable requires that the signal be above a certain signal level or a determined "detectable" signal. Adding the between variable requires a signal in the test group be above the ordered threshold value. When using the modified M statistic we use the same p-value calculation as for the unmodified M statistic. This is technically not the correct p-value because it does not take into account the additional modifications. The combinatorics that are used to derive the p-value cannot take into account the additional modifications, nor can the p-values be modified without additional assumptions. Therefore, this p-value should be considered an upper bound to the true p-value.

PREDICTION AND CLASSIFICATION

Prediction and classification problems are usually the true end point for biomarker profiling of disease or drug response. The first two steps of normalization and identification are viewed as steps to building up a diagnostic panel. With this panel in hand, the next step is to classify new samples. To do this correctly the new data need to be normalized. Note that special care must be paid to the application of the normalization technique to new data. Once this is done, the information about the identified markers can then be used to classify the new samples.

To use this classifier we need to estimate the prevalence of individual markers in the panel. We can use a naïve Bayes classifier to identify a panel of significant differential biomarkers to make diagnostic decisions. The naïve Bayes classifier assumes pairwise conditional independence in the prevalence of the individual biomarkers that make up the diagnostic panel.

Conditional independence means that,

$$P(A \cap B|C) = P(A|C)P(B|C) \tag{21.26}$$

The probability of A and B, given C, is equal to the product of the probability of A given C and the probability of B given C. Expanding this definition for pairwise conditional independence means that

$$P(M_i \cap M_j | C) = P(M_i|C)P(M_j | C) \tag{21.27}$$

for $i \neq j$ and $i, j = 1,2,\cdots,n$. This just means that the probability for any two events given C is equal to the product of the probability of the individual events given C.

It is worth pointing out that the pairwise conditional independence assumption is likely never really true, but it has been shown in many different studies ranging from predicting protein crystallization from sequence to detecting e-mail spam that the classification model still performs very well.[16,17]

For the sake of notation let:

C^+ represents positive diagnostic outcome, for example positive for cancer or successful treatment.

C^- represents the complement of C^+, i.e., a negative diagnostic outcome, such as not having cancer or unsuccessful treatment.

M_i is the state of the i[th] marker, here this can either equal 0 if the marker is negative or 1 if the marker is positive, here $i = 1,2,\ldots,n$ that there are n markers in the diagnostic panel.

The Naïve Bayes Classifier for a positive diagnostic outcome can thus be written as,

$$P\left(C^+ | \bigcap_{i=1}^{n} M_i = m_i\right) = \frac{P(C^+)\prod_{i=1}^{n} P(M_i = m_i|C^+)}{P(C^+)\prod_{i=1}^{n} P(M_i = m_i|C^+) + P(C^-)\prod_{i=1}^{n} P(M_i = m_i|C^-)} \tag{21.28}$$

The left part of the equation is the probability of having a positive diagnostic outcome given the observation of these results of markers.

Note that this equation gives the probability of a positive diagnostic outcome given the state of the individual markers of the diagnostic panel. Specifically, we break down the individual components of the above model and explicitly look at each individually.

For this model $P(C^+)$ is the probability of a positive diagnosis in the population. For example, this can be the rate of lung cancer in the general population. Conversely, $P(C^-)$ is the probability of a negative diagnosis in the population. These are mutually exclusive events that define the space, i.e.,

$$P(C^-) = 1 - P(C^+). \tag{21.29}$$

Additionally, we define that

$$P\left(M_i = m_i | C^+\right) = \begin{cases} \hat{p}_i^+, & \text{if } m_i = 1, \text{ i.e. } M_i^{\text{th}} \text{ marker is positive;} \\ 1 - \hat{p}_i^+, & \text{if } m_i = 0, \text{ i.e. } M_i^{\text{th}} \text{ marker is negative.} \end{cases} \tag{21.30a}$$

and

$$P(M_i = m_i | C^-) = \begin{cases} \hat{p}_i^-, & \text{if } m_i = 1, \text{ i.e. } M_i^{\text{th}} \text{ marker is positive;} \\ 1 - \hat{p}_i^-, & \text{if } m_i = 0, \text{ i.e. } M_i^{\text{th}} \text{ marker is negative.} \end{cases} \tag{21.30b}$$

where

$$\hat{p}_i^+ = \begin{cases} \dfrac{x^+}{n^+}, & \text{if a frequentist approach is used;} \\ \dfrac{x^+ + 1}{n^+ + 2}, & \text{if bayesian approach is used.} \end{cases} \tag{21.31a}$$

where x^+ is the number of samples positive from the diagnostic population and n^+ is the total number of samples looked at from the diagnostic population. Additionally,

$$\hat{p}_i^- = \begin{cases} \dfrac{x^-}{n^-}, & \text{if frequentist approach is used;} \\ \dfrac{x^- + 1}{n^- + 2}, & \text{if bayesian approach is used.} \end{cases} \tag{21.31b}$$

where x^- is the number of samples positive from the non-diagnostic population and n^- is the total number of samples looked at from the non-diagnostic population, additionally.

With these equations in hand, we can rewrite the naïve Bayes classifier to be:

$$P\left(C^{+}\middle|\bigcap_{i=1}^{n} M_i = m_i\right) = \frac{P(C^{+})\prod_{i=1}^{n}\left(\hat{p}_i^{+}\right)^{m_i}\left(1-\hat{p}_i^{+}\right)^{1-m_i}}{P(C^{+})\prod_{i=1}^{n}\left(\hat{p}_i^{+}\right)^{m_i}\left(1-\hat{p}_i^{+}\right)^{1-m_i} + P(C^{-})\prod_{i=1}^{n}\left(\hat{p}_i^{-}\right)^{m_i}\left(1-\hat{p}_i^{-}\right)^{1-m_i}} \tag{21.32}$$

Suppose that you have a 20 biomarker panel for disease, where for each marker the prevalence of the biomarker for the disease of 30% in the diseased population ($p_i^{+} = .3$ for $i = 1,2,\ldots,n$) and 10% in the normal population ($p_i^{-} = .1$ for $i = 1,2,\ldots,n$). Finally, let us assume that the disease is in .1322% of the population ($P(C^{+}) = .001322$) (note this is the national 2005 estimated breast cancer rate in females, age adjusted, in the U.S. according to the American Cancer Society). When applying the naïve Bayes classifier with this panel we get Table 21.4.

We can use this table to create a diagnostic test and judge how it will perform as a screening tool, as well as a companion diagnostic tool. For example, a PSA test for prostate cancer, which typically screens for PSA in excess of 4 mg/μl, which has a false positive rate of 70%–75%;[18] a mammogram has a false positive rate of 20%.[19]

The individual markers in this diagnostic panel are poor screening tools alone; if a single marker is positive, the probability of a diagnostic positive is .3956% (Note to calculate this plug into (21.32) with n = 1). However, if we take a panel of similar markers, we can see a significant performance increase. For example, with 5 out of 20 positive markers as a screening tool, there is a 0.74% chance of a positive diagnostic outcome, 70.06% chance of a true positive diagnostic outcome with a positive companion test with a 70% false positive rate, and 95.73% chance of a true positive diagnostic outcome as a positive companion test with a 20% false positive rate.

The Assumptions

This method has many assumptions. First, we assume that individual markers are pairwise independent. If this assumption is true, it means that being positive for any marker, based on being positive, gives no information on being positive for another marker, which seems not likely true in human physiology. However, Bayes classifiers have been shown to work even when these assumptions do not hold up.[16,17]

Additionally, it assumes that the probabilities are estimated correctly (they are unbiased and random samples from the population) and that the new samples are also unbiased samples from the same population.

The Upside

This is a powerful method for converting a diagnostic panel of individually poor markers into a good test. The method is relatively simple to apply, because it is just

TABLE 21.4
Probability Cut-Off Table

Number of Positive Markers	Bayes Classifier p-Value for Screening	Bayes Classifier p-Value for 70% FP Companion Diagnostic	Bayes Classifier p-Value for 20% FP Companion Diagnostic
0	.0000087	.002805	.025581
1	.0000335	.010733	.091949
2	.0001292	.040166	.280872
3	.0004983	.138978	.601037
4	.0019193	.383698	.853174
5	.0073628	.706002	.957289
6	.0278138	.902558	.988565
7	.0993840	.972772	.99701
8	.2985606	.992796	.999223
9	.6214636	.998122	.999798
10	.8636205	.999512	.999948
11	.9606691	.999874	.999986
12	.9894971	.999967	.999996
13	.9972557	.999992	.999999
14	.9992871	.999998	1
15	.9998151	.999999	1
16	.9999520	1	1
17	.9999876	1	1
18	.9999968	1	1
19	.9999992	1	1
20	.9999998	1	1

Note: The first column is the number of positive individual markers out of the panel of 20 markers, the 2nd column is the naïve Bayesian classifier probability. The third and fourth columns are the bayes classifier if the diagnostic panel is used as a companion tool with 70% and 20% false positive rate respectively.

a product of probabilities. The method also gives an overall probability of being diagnostically positive.

The Downside

Using ROC curves[20,21] is likely required to allow the user to determine the right cut-off using the naïve Bayes classifier, with an appropriate balance between false positive and a false negative rate.

A relatively large amount of data is required to get a good estimate of probabilities. It is very important that both data sets are unbiased random samples of the populations of interest.

Special care to normalization should be used for this and any classification method. Typical normalization methods are data-driven, meaning that they are normalized to the data in the test set. To work with new training data does not require just applying the same normalization algorithm, but requires using the results from the training set if the normalization is data-driven.

REFERENCES

1. Hartemink, A. et al., Maximum likelihood estimation of optimal scaling factors for expression array normalization, *Proceedings of the International Society for Optical Engineering*, 4266, 132, 2001.
2. Chebyshev, P.L., Des valeurs moyennes, *Journal de Mathmatiques Pures et Appliquées Series 2*, 12, 177, 1867.
3. Schweitzer, B., Predki, P.F., and Snyder, M., Microarrays to characterize protein interactions on a whole-proteome scale, *Proteomics*, 3, 2190, 2003.
4. Michaud, G.A. et al., Analyzing antibody specificity with whole proteome microarrays, *Nature Biotechnology*, 21, 1509, 2003.
5. Bolstad, B.M. et al., A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics*, 19, 185, 2003.
6. Irizarry, R.A. et al., Exploration, normalization and summaries of high density oligonucleotide array probe level data, *Biostatistics*, 4, 249, 2003.
7. Irizarry, R.A. et al., Summaries of Affymetrix GeneChip probe level data, *Nucleic Acid Research*, 31, e15, 2003.
8. Bolstad, B.M., Probe level quantile normalization of high density oligonucleotide array data, unpublished manuscript, http://bolstad.com/stuff/qnorm.pdf, 2001.
9. Qiu, J. et al., Development of natural protein microarrays for diagnosing cancer based on an antibody response to tumor antigens, *Journal of Proteomic Research*, 3, 261, 2004.
10. Yang, Y. et al., Development of a toxicogenomics *in vitro* assay for the efficient characterization of compounds, *Pharmacogenomics*, 7, 177, 2006.
11. Higo, M. et al., Identification of candidate radioresistant genes in human squamous cell carcinoma cells through gene expression analysis using DNA microarrays, *Oncology Reports*, 14, 1293, 2005.
12. Berkvens, D. et al., Estimating disease prevalence in a bayesian framework using probabilistic constraints, *Epidemiology*, 17, 145, 2006.
13. Lew, R.A. and Levy, P.S., Estimation of prevalence on the basis of screening tests, *Statistical Medicine*, 8, 1225, 1989.
14. Orr, K.A., O'Reilly, K.L., and Scholl, D.T., Estimation of sensitivity and specificity of two diagnostic tests for bovine immunodeficiency virus using bayesian techniques, *Preventive Veterinary Medicine*, 61, 79, 2003.
15. Austin, P.C., Brunner, L.J., and Hux, J.E., Bayeswatch: An overview of bayesian statistics, *Journal of Evaluation in Clinical Practice*, 8, 277, 2002.
16. Todd, B.S., Stamper, R., and Macpherson, P., A probabilistic rule-based expert system, *International Journal of Biomedical Computing*, 33, 129, 1993.
17. Zorkadis, V.M., Karras, D.A., and Panayotou, M., Efficient information theoretic strategies for classifier combination, feature extraction and performance evaluation in improving false positive and false negatives for spam e-mail filtering, *Neural Network*, 18, 799, 2005.

18. Keetch, D.W., Catalona, W.J., and Smith, D.S., Serial prostatic biopsies in men with persistently elevated serum prostate specific antigen values, *The Journal of Urology*, 151, 1571, 1994.
19. Hofvind, S., Thoresen, S., and Tretli, S., The cumulative risk of a false-positive recall in the Norwegian breast cancer screening program, *Cancer*, 101, 1501, 2004.
20. Pepe, M.S. and Longton, G., Standardizing diagnostic markers to evaluate and compare their performance, *Epidemiology*, 16, 598, 2005.
21. Pepe, M.S., Cai, T., and Longton, G., Combining predictors for classification using the area under the receiver operating characteristic curve, *Biometrics*, 62, 221, 2006.

22 Evaluating Precision and Recall in Functional Protein Arrays

Keith Robison

CONTENTS

INTRODUCTION

Protein kinases are key mediators of intracellular signal transduction cascades and are an important target class for pharmaceutical development. Two key challenges to fully understanding the role of individual kinases in cellular biology have been the difficulty in identifying the direct protein substrates phosphorylated by these enzymes and monitoring these phosphorylation events *in vivo*. For many protein kinases, no direct substrates have yet been reported and so the need for substrate identification methods is pressing.

Functional protein arrays (FPAs) offer an approach to proteome-wide identification of protein kinase targets.[1,2] Such arrays consist of proteins of the organism of interest spotted onto a solid support. These arrays can then be probed with the kinase of interest and phosphorylation detected via a number of strategies, such as incubation of the kinase with radiolabeled ATP or downstream detection using phosphospecific antibodies or fluorescent dyes.

A critical question for this nascent technology is how faithfully it reports the substrate profile of a kinase. Further experimental validation of kinase substrates *in vivo* generally necessitate low throughput experiments that may involve the expensive

and slow generation of custom reagents such as phosphospecific antibodies. Understanding the probability of success of such endeavors is a useful prerequisite to undertaking them. This chapter will examine the available data on FPA success and use computer simulations to further explore this topic. While the specific examples used are protein kinases and their substrates, the approach is equally applicable to other molecular interaction studies using FPAs.

HOW FUNCTIONAL ARE FPAS?

Protein kinase substrate hunts on functional protein arrays could go awry for a number of reasons. First, the folding state of most proteins on the array is unknown, and misfolding events could mask or destroy substrate binding motifs or could expose motifs that are safely buried *in vivo*. Second, the protein may be missing binding partners which normally mask potential kinase motifs. Such partners may also be critical to the recruitment of kinases to a particular substrate. Third, the process of making the array may hinder the protein. For example, either the protein's binding to the solid surface or modifications introduced for the expression and purification of the protein may interfere with proper phosphorylation. Fourth, and most complicated, is the post-translational state of the protein. For some substrate-kinase pairs, prior phosphorylation of the substrate is critical to constructing the correct phosphorylation site motif. In other cases, phosphorylation of a protein or an interacting partner is required for recruitment of a kinase to a substrate. Conversely, in some cases phosphorylation may destroy the recognition sequence for a kinase. Hence, some proteins may have a very large number of true *in vivo* states. The tumor suppressor p53 has 23 reported phosphorylation sites (author's compilation), or over 8 million (2^{23}) possible phosphorylation states. Other modifications, such as glycosylation and ubiquitination further increase the potential complexity. The likelihood of a protein being in a correctly modified, correctly folded, correctly partnered state drops further if the array protein has been expressed in a heterologous system. A further twist is that functional protein arrays, like two-hybrid approaches, completely divorce proteins from their temporal expression and cellular localization context. A number of experiments suggest that *in vivo* molecular targeting appears to rely on codes with the minimum amount of information required for success.[3,4] Hence, a divorce from a protein kinase from its cellular expression context will almost certainly lead to false positives.

Even with a representative sample of proteomic diversity, functional protein arrays present analytic challenges. Published studies with such arrays have used simple ratiometric comparisons of small[1,2,5] or no[6] replicate experiment numbers. Are these adequate to the task? Only by careful calibration can we assess the reliability of these approaches and determine whether more complex ones are warranted.

PRECISION, RECALL, AND THE COMPARATOR CHALLENGE

Precision and recall are common measures of search performance. Precision is the fraction of hits which are true hits whereas recall is the fraction of true hits recovered by a search (Figure 22.1). These concepts are closely related to those of false positive

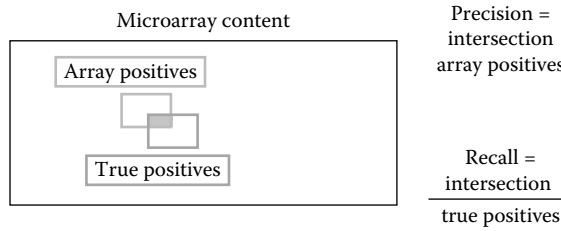


FIGURE 22.1 Graphical definitions of Precision and Recall.

and false positive rates. A false positive rate can be found by subtracting the precision from one and similarly the false negative rate is one minus the recall rate.

A great difficulty in evaluating these performance measures lies in identifying reference data sets with which to make a comparison. Protein functional array technology is needed because so few substrates are known, but the lack of substrates makes it difficult to calibrate the technology. For this analysis, three sources of information were used as calibration sets. First, for some kinases a significant number of substrates have been previously reported. Second, some authors have used various low throughput approaches to assess the quality of their results *in vivo*. Third, there exists an *in vivo* whole proteome substrate identification technology called ASKA (Analog Sensitive Kinase Allele; see below) which can be treated as a comparator.

Compilations of literature results can be labor-intensive to build, but once built are valuable tools for validation of new approaches. Publicly available phosphorylation databases are available[7–9] but may not be comprehensive. However, published literature results cannot be assumed to be a gold standard. First, some of these results may be erroneous, either because the original experiment was misinterpreted (for example, if the kinase preparation used contained multiple kinases) or due to confusion of overlapping biological names in the literature. Experimental conditions, such as over-expression of artificial constructs, may lead to phosphorylation events which do not occur naturally. Second, such lists are incomplete and worse have no estimate of their incompleteness. As a result, literature comparisons can be used to estimate recall but are useless for precision estimates.

Conversely, direct follow-up of positive results from an array experiment can give information about precision, but not recall. Using the same approaches to assess the literature true positives and false negatives from an experiment would give a better estimate and aid in interpreting the novel results, but such controls are not present in any of the existing kinase substrate array publications.

ASKA COMPARATOR, GET AN ANSWER?

ASKA (Analog Sensitive Kinase Allele) is an approach to identifying *in vivo* or in cellular extracts the specific substrates of a particular kinase. The kinase of interest is mutagenized to introduce an extra nook within the ATP-binding pocket, and a specific ATP analog is used which contains a bulky group which can fit in this nook. Other kinases are unable to use the analog due to steric clashes created by this bulk.

By using a radiolabeled analog, the substrates of the modified kinase can be specifically identified. The primary drawback to this elegant approach is the labor and difficulty in creating the modified kinase allele and reintroducing it into the correct cellular context. Because it is a whole proteome technique, ASKA results can yield estimates of both precision and recall when used as a comparator for an independent whole proteome technique such as FPAs. Four ASKA experiments are considered in this analysis.[10–13]

A human ASKA experiment by Larochelle et al. looked for substrates of Cdk7 in human nuclear extracts and found ten bands.[12] This implies, if (as is likely) Cdk7 activity is restricted to the nucleus, that Cdk7 has at most ten substrates in humans. This experiment illustrates how restricted some kinases are in their targets, and subsequently the stringency required for identifying substrates of such kinases. In this case, assuming a human proteome of 25,000 proteins (a low estimate ignoring alternative splicing), the probability of any one protein being a true Cdk7 substrate is 4×10^{-5} and substrate identification techniques with low precision (high false positive rate) are not going to succeed. A corresponding FPA experiment is not available, so illustrating the low end of kinase substrate abundance is the limit of this study's utility for FPA assessment.

The other ASKA experiments were performed in the budding yeast *Saccharomyces cerevisiae*. Two looked at the yeast cyclin dependent kinase Cdc28(Cdk1) and the influence of two cyclins on the substrate specificity of this enzyme.[10,13] A particularly important contribution of Ubersax et al. is to estimate the abundance of Cdk1 substrates at around 500.[13] Nearly 200 of these were initially subjected to validation,[13] with 150 being used in later analyses.[10] The follow-up experiment by Loog and Morgan further investigated the influence of different cyclins on the selection of substrates.[10] Dephoure et al. looked at a different cyclin-dependent kinase, Pho85, Pho85 and the influence of a selection of its cyclin partners on its substrate specificity.[11]

ARRAY EXPERIMENTS UNDER THE LENS

For the purposes of this analysis three protein array experiments will be examined. Two are published experiments on a (nearly) whole proteome yeast array, including one providing a look at about two thirds of all yeast kinases.[2] The other data is from an unpublished experiment from the author using an array containing about 2000 human proteins, or 8% of the human proteome. An additional experiment using an Arabidopsis array[5] was omitted because of the lack of extensive prior Arabidopsis kinase substrate data for comparison. This study did not perform any *in vivo* validation which could be used to estimate precision.

Mah et al. examined the kinase Dbf2. These arrays are produced in yeast, offering a greater probability of correct folding and partnering, and perhaps even a representative sampling of post-translational states Western blot and IVK followup.[1] Ptacek et al. used the same type of arrays as the first, but there were probed in duplicate with more than 80 yeast kinases.[2]

A third experiment (the author, unpublished results) probed a commercially-available human protein array (Invitrogen ProtoArray 1.0) with protein kinase A (PKA), an intensely-studied kinase which offers a large literature of identified

substrates. The human FPA contained approximately 8% of the human proteome and was produced in insect cells, and hence is less likely than the yeast arrays to faithfully represent the *in vivo* state of these proteins.

COMPARISONS

Table 22.1 shows the comparisons which can be made. For Ptacek et al.,[2] literature comparisons were made where 7 or more literature substrates had been reported. The table illustrates the sparseness of data at this early stage of kinase substrate identification by FPAs. Furthermore, in most cases we can estimate precision or recall but not both. What is very striking is the wide range of values observed for both values. While a recall as high as 33% is observed, for other kinases no known substrates were recovered. Precision values are fewer but similarly scattered, ranging from 1% to 59%. Undoubtedly some of the problem are the low numbers sample sizes; in only three cases do we have even twenty examples. Hence, the 0% cases would jump to over 5% if a single additional true positive were found (such as by further scrutiny of the existing literature) which was also a hit. So all of these estimates are inherently crude.

A limitation of ASKA combined with immunoprecipitation is a potential inability to detect phosphorylation of low abundance proteins.[11] The availability of yeast protein abundance data[14] enables looking for a significant relationship between the abundance of an FPA hit for Pho85-Pho80 or Pho85-Pcl1[2] and whether it was identified in the ASKA experiment.[11] However, using a T-test on the log10 (molecules-per-cell) values in a T-test yields P-values of 0.612 and 0.073 respectively, suggesting that the difference between the FPA and ASKA experiments is not a simple matter of FPA detecting low abundance true positives missed by ASKA.

SIMULATING SUCCESS AND FAILURE

An important question to ask is what do these numbers mean in a real world situation? Are FPAs a practical approach for identifying human protein kinase substrates, and if so what might alter the chances of success. One way to explore this question is through computer simulation. For this simulation, success is defined as taking ten hits from the array experiment (or all hits if fewer than ten) and finding at least one is a true *in vivo* substrate. This definition is arbitrary, but does describe a follow-up campaign that is both substantial and practical.

Three sizes of protein array were chosen for simulation. The smallest, 2000, is approximately the size of the first commercially available human array which was used for the PKA experiment. The second, 5000, represents the largest currently available protein array. The remaining size, 10,000, represents (anticipated) future expansion. Each array is modeled as a random selection of proteins from a 25,000-protein proteome, with a similar random selection of false positive and false negative hits on each array. The number of potential false positives, true positives, and false negatives in the array are calculated from set precision and recall values and a set number of true substrates in the proteome; by cycling through combinations of these values the entire space can be explored. So each round of simulation involves set

TABLE 22.1
Summary of Recall and Precision Estimates for Functional Protein Array Analyses of Protein Kinase Substrates in Human and Yeast

Array Experiment	System	Kinase	Comparator	True Substrates	Precision	Recall	Notes
Robison	Human	PKA	Literature	18		11%	
Mah et al.[1]	Yeast	Dbf2-Mob1	Western blot and IVK of immunoprecipitates	17	59%		
Ptacek et al.[2]	Yeast	Cdc5	Literature	7		0%	
		Cdc28	Literature	15		0%	ASKA data excluded
		Cka1	Literature	17		0%	
		Cla4	Literature	9		0%	
		Fus3	Literature	12		10%	
		Ipl1	Literature	12		0%	
		Pho85	Literature	15		20%	
		Snf1	Literature	16		6%	
		Tpk1	Literature	48		33%	
		Cdc28-Clb2	ASKA with immunoprecipitates[10]	150	23%	7%	
		Pho85-Pcl1	ASKA with immunoprecipitates[11]	21	10%	10%	
		Pho85-Pho80	ASKA with immunoprecipitates[11]	18	2%	33%	
		6 kinases	Gel shift / reduced phosphorylation of IP from kinase-deleted strain		9%		

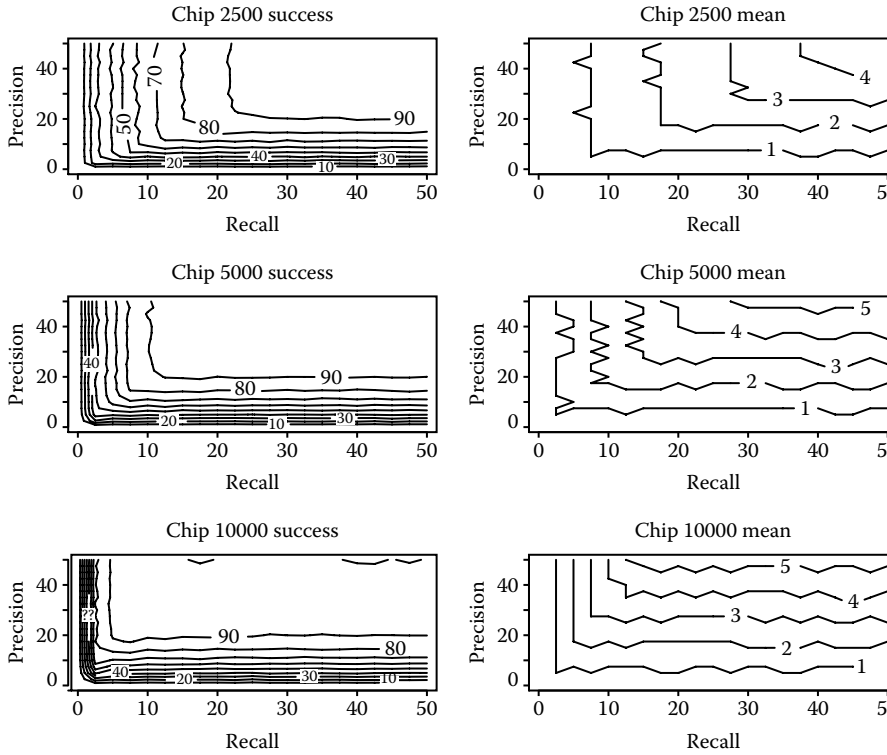


FIGURE 22.2 Contour plots showing the probability estimates of obtaining at least one substrate (left side) and the mean number of substrates found (right side) for a kinase with 100 true targets and chip sizes of 2500 (top row), 5000 (middle row) and 10000 (bottom row) and 10 hits (or all hits if fewer than 10) from the array subjected to *in vivo* testing.

values for array size, number of true substrates in the entire proteome, precision rate, and recall rate, and then at least 2000 array experiments are generated within that parameter space. After the virtual array is run, 10 hits (or all if fewer) are randomly selected; if one of these is a true positive, then that campaign is judged a success. For any given parameter combination, the percent of campaigns yielding success is calculated. The mean number of substrates found per campaign is also tabulated.

Figure 22.2 presents a contour map of the success probabilities estimated by this simulation in the case of a kinase with 100 true substrates in the human proteome. A number of human kinases have approximately 100 or more known substrates, including Akt, Cdc1, Cdc2, CK2, PKA, and PKC (author's compilation), so this would plausibly represent a large number of kinases in the proteome. On the other hand, as in the case of Cdk7, some kinases may have much smaller numbers of true substrates. A similar set of contour maps for 25 true substrates is presented in Figure 22.3. It should be remembered that the lowest value explored by the simulation for either precision or recall was 0.025 (2.5%), and so contour lines plotted with lower values are artifactual (as is some of the irregularity of the contours).

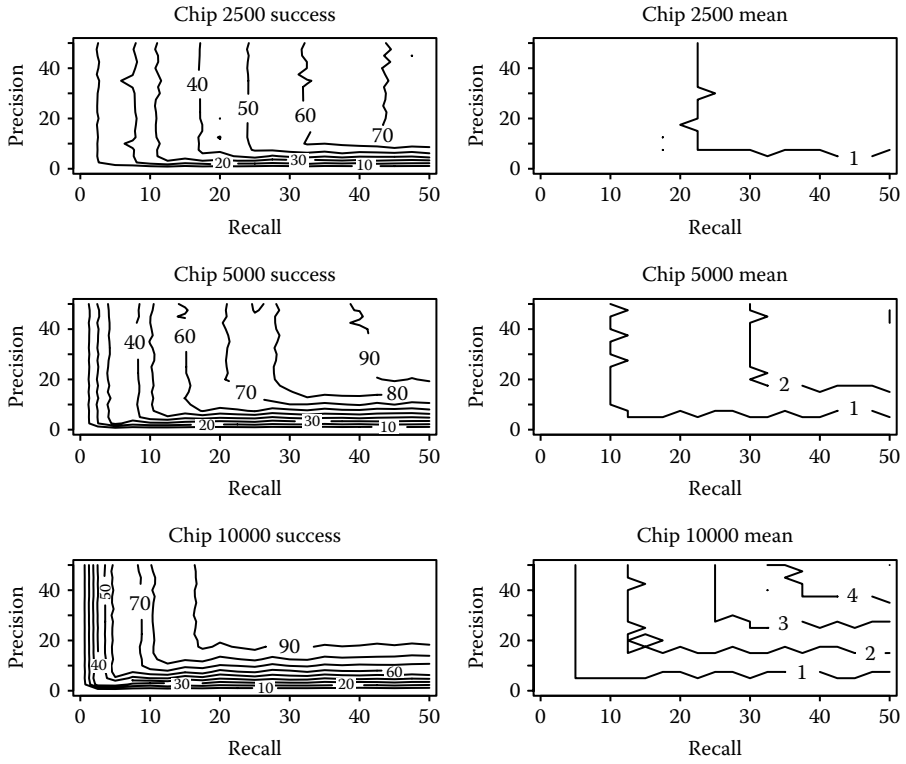


FIGURE 22.3 Contour plots as in Figure 22.2, but for a kinase with 25 true substrates.

Quite clear from the plots is the strong effect of recall on the overall success of a campaign. For example, with a 5000-protein array and 100 true positives, a recall of 0.1 (10%) limits the success rate to about 50% regardless of the precision value, but a precision of 0.1 allows a wide variety of different success rates depending on the recall rate. This is not to suggest that precision is irrelevant; particularly in the case of 25 substrates and small array sizes. But, once the precision exceeds 1 over the sample size of 10 it is not longer constraining. Since precision is the fraction of true positives expected in a sample, this result is unsurprising. The right hand side of the figure shows the other key impact of precision: the number of substrates likely to be identified.

One key parameter in these simulations is itself generally unknown: the number of true *in vivo* substrates of a particular kinase. For some kinases, such as PKA or Akt, large numbers of substrates have been identified and it is unlikely that saturation has occurred, so some estimate is possible. For largely uncharacterized kinases, however, it is simply an estimate which an investigator must make based on their own hunches and tolerance for risk of failure. Some kinases may have extremely focused biological roles resulting in extremely limited substrate repertoires, as in the case of Cdk7.[12] In a few cases, experimental estimates of substrate numbers may be available from ASKA or two-dimensional electrophoresis studies. A conservative

estimate is that any kinase with large numbers of substrates is likely to be relatively well characterized, and so uncharacterized kinases are unlikely to have large numbers of substrates. However, motif finding experiments in large phosphorylation site databases suggest several motifs which do not correspond to any known kinase, and so may represent the motifs of as-yet-to-be-characterized kinases with relatively large substrate repertoires.[15]

In a real-life situation, it may be possible to use prior knowledge to make this final draw biased in favor of true substrates. For example, if the localization of a protein is known only those candidate substrates which share the same localization would be included. Since large catalogs of protein localization are available from bioinformatics culling of the literature as well as focused proteomics efforts. However, such filtering is not possible if the location of kinase action is not known, and may provide only a modest boost in accuracy.

In a similar manner, Ptacek et al.[2] noted that for some of their kinases the substrates were highly enriched for particular categories of proteins, which sometimes correlated with the known function of the kinase in question. Such overlaps can be a powerful way to bias the choice of proteins for *in vivo* validation, but may not always be present. Ptacek et al. also demonstrated an overrepresentation of certain circuit motifs in the integration of their FPA data with other high throughput data. Examples of these circuit motifs include a kinase phosphorylating two proteins which interact with each other and a kinase phosphorylating both a transcription factor and a target of that transcription factor. When such information is available, it can be extremely valuable for prioritizing hits for follow-up. However, in many cases no such clues may be available for a previously uncharacterized kinase in an organism such as human where protein-protein interaction and transcription factor networks are still very sparsely elucidated.

So are FPAs a suitable technology for protein kinase substrate identification? It depends on the kinase, the ability to attempt validation on many substrates, and a tolerance for the risk of failure. For example, with the current commercially available human array (5000 proteins; 20% of proteome) and 100 true substrates in the proteome, even if both precision and recall are around 10% the probability of success with 10 hits followed up is in the range of 50 to 60%. So even for kinases with relatively limited substrate repertoires and relatively modest precision and recall values (which have been seen in real experiments), the probability of success is better than half.

CONCLUSIONS AND SUMMARY

The results shown here demonstrate that functional protein arrays are likely to be useful tools for kinase substrate research, particularly as the content on these arrays grows. The approaches presented here should assist a researcher in deciding whether an FPA-based approach is likely to work.

The technology of functional protein arrays are in their infancy. What sorts of approaches will drive their maturation?

First, better statistical approaches should enhance the ability to recover true positives and suppress false positives. Published FPA experiments have tended to use simple

statistics such as Z-scores (standard deviations from the mean) computed from control samples coupled with ad hoc voting systems using the replicate kinase treated samples. This approach ignores the variance of the kinase treated samples and fails to fully utilize the combined variance of the untreated and treated samples. Such an approach also fails to use all of the available information when multiple kinases are used to probe a set of arrays in parallel or in series (see below). Furthermore, such approaches completely ignore the multiple testing problem; if we run enough spots on the array, by chance we will ultimately see a positive result even if there are no true positives. More sophisticated statistical models, such as T-tests, ANOVA, and SAM, are capable of utilizing more of the available information on noise[16,17] and hence should yield superior results. A large body of literature has examined the statistical treatment of genome-scale studies on RNA and DNA arrays, and these methods would be a good place to start. The particular noise characteristics of FPAs, which are likely to be coupled both to array production methods as well as specific applications such as kinase substrate identification or protein interaction analysis, will require investigation in order to select the specific methods and tune them.

A second goal for further FPA research should be to investigate deliberately manipulating the post-translational state of the arrays by dephosphorylating them and specifically phosphorylating them with specific kinases. Feasibility of dephosphorylation has been reported, but results of such experiments have not been reported.[2] Ultimately, such experiments should be coupled to MS/MS identification of the specific phosphorylation state present in the original protein mixture and the phosphorylation state after each stage of treatment. This would assist in understanding to what degree the phosphorylation state of array proteins is interfering with the identification of true substrates. Evidence for phosphorylation state being an issue comes from a published experiment searching for interactors with 14-3-3,[6] a protein which primarily recognizes specific phosphorylated motifs.[18] None of the 20 positives identified and validated by Satoh et al. were shown to be phosphorylation-specific interactions, nor did the experiment recover any of the 11 known 14-3-3 interactors present on the array.[6]

Another area ripe for exploitation, particularly when coupled to sophisticated statistical models, is to carefully explore the effects of experimental conditions on the array results. A recent array based interaction study performed this very elegantly, using multiple concentrations of query proteins to explore the affinity profile of interactors to ERBB family kinases.[19]

A fourth key area of research is understanding how differences in FPA production affect results. For example, all of the FPA experiments considered here involved expression (in a heterologous system in the case of human) and purification off-chip and spotting onto the chip with a contact printer.[2] An alternative strategy is to skip purification and spot crude lysates directly onto the chip, with an affinity capture material capturing only the desired protein.[20] Yet another alternative is to synthesize proteins *in situ* using mammalian *in vitro* transcription and translation.[21] Exploring how these different alternatives affect the state of proteins and how these changes translate into differences in recall and precision would be highly valuable.

Foremost, though, is the need for more *in vivo* validation information on proteins identified as kinase substrates by FPA. Some of this will come from the further

accumulation of substrates in the literature, but more directed approaches are needed. Particularly interesting would be additional cross-comparisons of ASKA results with FPA results in the same system coupled to MS/MS exploration of the actual phosphorylation state of hits from the two methods. These would help pin down the degree to which multiple phosphorylations are enabling ASKA to identify low abundance substrates.

Finally, it should be noted that this particular analytic strategy is not specifically limited to kinase substrate identification. Many of the same issues haunt any protein interaction study on protein arrays. Kinase substrates were chosen for this particular study both because of the author's interests but more critically because far more published data is available for unfocused arrays for kinase substrate identification than for general interaction studies. Studies cross-validating high-throughput various interaction detection technologies (such as FPAs, two hybrid approaches, and high-throughput immunoprecipitation) with low-throughput ones are needed here as well.

ACKNOWLEDGMENTS

The author wishes to acknowledge many valuable discussions with Stan Letovsky, Mike Pickard, Arijit Chakravarti, and Eric Lightcap at Millennium; Steve Roels at Millennium for critical assistance with figure preparation, Alan Ruttenberg at Millennium for critical reading of the manuscript, Janie Merkel and Paul Predki at Invitrogen for discussions and the execution of the PKA experiment, and Mike Snyder of Yale University for prepublication access to Ptacek et al.

REFERENCES

1. Mah, A.S. et al., Substrate specificity analysis of protein kinase complex Dbf 2-Mob1 by peptide library and proteome array screening, *BMC Biochem*., 6, 22, 2005.
2. Ptacek, J. et al., Global analysis of protein phosphorylation in yeast, *Nature*, 438, 679, 2005.
3. Zarrinpar, A., Park, S.H., and Lim, W.A., Optimization of specificity in a cellular protein interaction network by negative selection, *Nature*, 426, 676, 2003.
4. Schneider, T.D. et al., Information content of binding sites on nucleotide sequences, *J. Mol. Biol*., 188, 415, 1986.
5. Feilner, T. et al., High throughput identification of potential Arabidopsis mitogen-activated protein kinases substrates, *Mol. Cell. Proteom.*, 4, 1558, 2005.
6. Satoh, J.I., Nanri, Y., and Yamamura, T., Rapid identification of 14-3-3-binding proteins by protein microarray analysis, *J. Neurosci. Meth.*, 152, 278, 2005.
7. Hornbeck, P.V. et al., PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation, *Proteomics*, 4, 1551, 2004.
8. Lee, T.Y. et al., dbPTM: An information repository of protein post-translational modification, *Nucleic Acids Res*., 34, D622, 2006.
9. Diella, F. et al., Phospho.ELM: A database of experimentally verified phosphorylation sites in eukaryotic proteins, *BMC Bioinform.*, 5, 79, 2004.
10. Loog, M. and Morgan, D.O., Cyclin specificity in the phosphorylation of cyclin-dependent kinase substrates, *Nature*, 434, 104, 2005.
11. Dephoure, N. et al., Combining chemical genetics and proteomics to identify protein kinase substrates, *Proc. Natl. Acad. Sci. USA*, 102, 17940, 2005.

12. Larochelle, S. et al., Dichotomous but stringent substrate selection by the dual-function Cdk7 complex revealed by chemical genetics, *Nat. Struct. Mol. Biol.*, 13, 55, 2006.
13. Ubersax, J.A. et al., Targets of the cyclin-dependent kinase Cdk1, *Nature*, 425, 859, 2003.
14. Ghaemmaghami, S. et al., Global analysis of protein expression in yeast, *Nature*, 425, 737, 2003.
15. Schwartz, D. and Gygi, S.P., An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets, *Nat. Biotechnol.*, 23, 1391, 2005.
16. Meiklejohn, C.D. and Townsend, J.P., A Bayesian method for analysing spotted microarray data, *Brief Bioinform.*, 6, 318, 2005.
17. Allison, D.B. et al., Microarray data analysis: From disarray to consolidation and consensus, *Nat. Rev. Genet.*, 7, 55, 2006.
18. Wilker, E. and Yaffe, M.B., 14-3-3 proteins — a focus on cancer and human disease, *J. Mol. Cell. Cardiol.*, 37, 633, 2004.
19. Jones, R.B. et al., A quantitative protein interaction network for the ErbB receptors using protein microarrays, *Nature*, 439, 168, 2006.
20. Boutell, J.M. et al., Functional protein microarrays for parallel characterisation of p53 mutants, *Proteomics*, 4, 1950, 2004.
21. Ramachandran, N. et al., Self-assembling protein microarrays, *Science*, 305, 86, 2004.

23 Visualization of Protein Microarray Data

Kevin Clancy

CONTENTS

INTRODUCTION

The generation of proteomic data sets is becoming increasingly common, while the analysis of such data remains in its infancy. Experience in handling genomic data, however, can help guide this challenge. This experience teaches us that efficient analysis, exchange and dissemination of proteomics data will require standardized methods for data storage and representation.

The increase in size and complexity of the various gene and genome sequence databases has been well documented over the last 10 years. Now proteomics — the identification and study of proteins, usually in sets defined by some biological context — is maturing with the development of a wide range of high-throughput analysis methodologies. These experiments typically consist of one or more separations performed on samples, often involving electrophoretic, chromatographic, or affinity-based techniques, followed by mass spectrometric or fluorescent detection and

quantitation. More recently, protein array-based approaches have been added to the arsenal of high-throughput proteomics technologies.

The representations of gene, genome, and transcriptome data have been largely standardized, and databases and tools for their analysis are widely used. Standardized sequence data formats, for instance, include the common GenBank/EMBL/DDBJ sequence files and SwissProt protein sequence representations. Three dimensional structural data is standardized utilizing PDB or NDB file formats. Transcriptional data in the form of microarray data has been standardized around the MIAME data format.[1] Many databases of protein pathways and protein interactions exist, most notably KEGG,[2] DIP,[3] and BIND[4] in the academic arena and GeneGo (www.genego.com) as an example of a commercial offering.

The situation around the analysis, storage, and retrieval of proteome data is less developed, largely because the field is nascent. Because of the dynamic nature of proteomics data it can be difficult to fully define the key data from different types of experiments. For example, there are many different subsets of the proteome of an organism, just as there are many different RNA transcription and turnover patterns, which may be classified by cellular or tissue type and condition. Proteome experiments produce data that need to be placed in context to provide a full biological understanding of their significance. Classification and origin of the sample, preparation of the sample, means analysis, and data acquisition all have an impact on the interpretation of the results for pools of samples. Without the ability to compare multiple experiments, results become much less useful to the scientific community at large than they could otherwise be. The context-sensitive nature of proteomic data also means that the investigator needs to gather a richer set of metadata (data about the data) than is required for basic genetic sequence data sets.

Effective data sharing will be essential to enable scientists to combine and interpret different types of experimental data. For instance, a genomics experiment may implicate a particular set of proteins. Follow-up analyses using proteomics approaches, such as 2D gel electrophoresis, biochemical assays, yeast two hybrid, co-localization, immunoprecipitation, protein array, or *in vivo* imaging techniques, may be desired. The scientist will build *de novo* experimental models from several pieces of experimental data. In some cases, these experiments may have already been performed, and reside in various databases. Scientists may piece together experimental information with existing pathway information in the public domain or begin to build a new pathway. Finally, they may want to use these models to further test their hypotheses.

This chapter will review the types of biological data and analyses that scientists will typically need: ontologies and standards to describe protein microarray experiments, issues involved in the visualization and mining of pathway and experimental data, and the use of systems biology markup language to help the scientist access a wide variety of programs with their own data.

What Types of Data and Analyses Do We Need To Track?

Much has been made of the explosion of data over the last 20 years. Scientists have steadily been taking advantage of this data to expand the scope of the types of analyses that they now routinely think about. With the development of microarray technologies, scientists encountered increased data handling issues. Many of the types of data handling

challenges for the protein microarray filed have been issues for the last 10 years in the fields of genomics and transcriptomics. However, proteomic arrays offer a further level of experimental and analysis complexity because several functionally different types of experiments may be performed upon one array, including protein–protein interaction and diagnostic studies, enzymatic and small molecule studies, and antigen and immunology profiling studies. Furthermore, subtle variations in experiments, including temperature, buffer composition, lot and age of the array, handling and storage of the array, etc. can have important consequences for experimental results. In addition, the nature of the interaction of a spotted protein target with a labeled protein probe is considerably more complex than the simpler well established kinetics governing nucleic acid hybridizations. Finally, the means of experimental data capture can vary depending upon application, extending from more traditional autoradiographic analyses to fluorescent analyses. Hence, data acquisition programs need to be able to capture a wider range of experimental data than is traditionally found in DNA microarray experiments.

The different types of data and results that an experimentalist will typically want to track are shown in Figure 23.1. These include microarray design, experimental design, microarray data acquisition, result storage, and data retrieval. In addition, investigators may require access to public and private databases; descriptions of experimental specific methods for visualization and analyses; access to public and private algorithms and data mining methodologies, incorporation of public sequence databases, canonical data, and methods for export of experimental data.

Traditional ways of managing laboratory data and experiments (lab notebooks and a mix of software) can become seriously overburdened when being used to track microarray experiments. Laboratory Information Management System (LIMS) software, however, is specifically designed for such tasks. LIMS software can offer a host of useful capabilities, including:

- Standardized data management and recording (such as one-stop storage of various array architectures and storage of lot specific array data important in later experimental analyses)
- Standardized storage and retrieval of experimental data (including designation of which application an array was used for, experimental details of how the array was used (particularly if protocols vary from each other), QA of experimental data and measurement of experimental reproducibility)
- Standardized data analysis (such as inclusion of pre-canned tools for analyses of data, interfaces to conveniently bring in third party applications, and storage of various stages of data analyses for later retrieval)
- Publication of data in a uniform format (including the ability to link to or retrieve public data for further analyses, storage or linkage to important downstream analyses, the ability to import and export of array data in a standardized format, the ability to share databases as an alternative mode of data exchange and report generation).

There are several commercially available array analysis products, as well as those developed and promoted via open source and community initiatives or developed

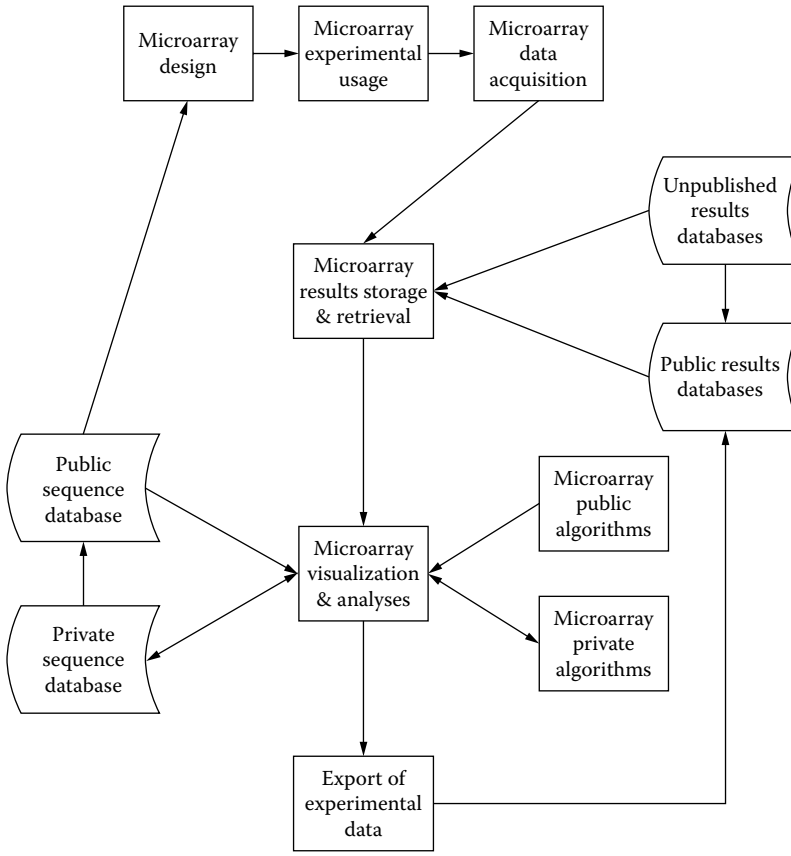


FIGURE 23.1 Idealized generic data flow in a protein array experiment. Data from both public sources and data from the scientist's private, unpublished experiments are managed within this LIMS. Different modules and associated data are accessible to humans via graphical viewers or to other programs through programmable interfaces. Software modules, as shown in rectangles, describe anticipated functional needs for the scientist. Local and public data sources that the system will access are indicated by curved boxes.

by governmental agencies such as NCBI or EMBL. Table 23.1 presents a table of options that are available to investigators. However, an important limitation of these offerings is that they are oriented toward the DNA microarray field. It is to be expected that protein array specific offerings will become more commonly available to researchers over the next couple of years. The next section describes some of the ways of standardizing and managing the exchange of data in such a LIMs.

Ontologies and Data Standards

One need that quickly arises in microarray data analysis is the desire to retrieve information related to a "hit" from a large number of different data sources. For instance, proteins on arrays can be identified by a GenBank ID or a manufacturer specific ID. The scientist may want to extract the protein sequence for the gene and

TABLE 23.1
Selection of LIMS Products for Microarray Experiments

Product		Company/Institute	Features
Acuity 2.0	http://www.axon.com/GN_Acuity.html	Axon Instruments	Various visualization tools; normalization, hierarchical, k-means, k-medians clustering with many different similarity metrics, SOM, PCA, gene shaving. Scripting engine for customizable analysis through VBScript, JavaScript or ActiveX objects store data in relational database, Supports Microsoft SQL Server 2000, ODBC-compliant
AMAD	http://www.microarrays.org/software.html	UCSF	Flat file, web driven database system written entirely in PERL and javascript, and intended for use with microarray generated data. Storage, retrieval, and extraction of microarray data by means of a centralized web based server. Interoperative with Cluster and Treeview softwares
ARGUS	http://vessels.bwh.harvard.edu/software/argus/default.htm	Dr. Michael A. Gimbrone Jr lab; Harvard University	Database software system designed to process, analyze, manage, and publish microarray data
ArrayDB	http://genome.nhgri.nih.gov/arraydb/	NHGRI	Interactive user interface for the mining and analysis of microarray gene expression data. All of the analyzed expression data and the clones used in the experiment, are stored in a relational database
Array Informatics	http://lifesciences. perkinelmer.com	Perkin Elmer Life Sciences	Direct data links to SpotArray, ScanArray, and QuantArray microarray instrumentation for increased automation capabilities; Barcode automation provides increased throughput; various data analysis and visualization tools

(continued)

TABLE 23.1 (CONTINUED)
Selection of LIMS Products for Microarray Experiments

Product		Company/Institute	Features
BASE (BioArray Software Environment)	http://base.theplu.se	Department of Oncology, Lund University	BASE is a database server to manage the massive amounts of data generated by microarray analysis. Manages biomaterial information, raw data and images, and provides integrated and "plug-in"-able normalization, data viewing and analysis tools. The system also has array production LIMS features; support MIAME and MAGE-ML
CloneTracker 2	http://www.biodiscovery.com	Biodiscovery	Manages and processes all information associated with fabrication of spotted microarrays using popular commercial robotic systems
Genedirector 2	http://www.biodiscovery.com	BioDiscovery	Comprehensive microarray data management system that organize, archive, retrieve, analyze and query all microarray data throughout the entire workflow; built-in LIMS for project management; Relation Database Design deployed on Oracle; Security control; integrated with Clonetracker, Imagene and Genesight softwares for various stage of data analysis.
GeNet 4.1	http://www.silicongenetics.com	Silicon Genetics	Dynamically archive and visualize expression data from a wide range of technologies and organisms.
Genetraffic 1.2	http://www.iobion.com	Iobion Informatics	Store raw data and images in a PostgreSQL relational database. MIAME compliant. Provides data management and data validation tools.

GeneX	http://genex.sourceforge.com	GeneX is an Open Source project based in SourceForge.net	GeneX is an Open Source database system.
Genowiz	http://www.ocimumbio.com	Ocimum Biosolutions	MIAME compliant database. Manages planning, execution, and storage and analysis of microarray experiments. Includes data classification and pathway analysis.
Longhorn Array Database	http://www.longhornarraydatab ase.org	Section of Molecular Genetics and Microbiology, Institute for Cellular and Molecular Biology, University of Texas at Austin	MIAME compliant microarray database that operates on PostgreSQL and Linux. It is a fully open source version of the Stanford Microarray Database (SMD)
LIMaS (Laboratory Information Management for Array Systems)	http://www.limas.har. mrc.ac.uk	Mammalian Genetics Unit, Harwell UK.	MIAME compliant java and relational database backend Trackes laboratory processes for expression experiments and expression data produced from image analysis
MADAM (MicroArray DAta Manager)	http://www.tigr.org	The Institute of Genomic Research (TIGR)	MADAM loads and retrieves microarray data to and from a local database. It provides data entry forms, data report forms and additional applications necessary to maintain microarray data for further analysis.
ArrayDB	http://www.arraygenetics.com	Array Genetics	A web-based database system for archiving and exchanging DNA microarray data (e.g., in .DAT, .CEL and .EXP formats).

(continued)

TABLE 23.1 (CONTINUED)
Selection of LIMS Products for Microarray Experiments

Product		Company/Institute	Features
maxd (Manchester Array Express Database)	http://www.bioinf.man.ac.uk	Microarray group,Machester Bioinformatics	A data warehouse and visualization environment for genomic expression data.
NOMAD	http://ucsf-nomad.sourceforge.com	UCSF	An open source system for storing and querying the results of microarray experiments.
Partisan ArrayLIMS	http://www.clonediag.com	Clondiag	MIAME compliant LIMS for microarrays: array design, manufacturing, experiments, sample management, analysis, open interfaces, imaging tools, data analysis
Phoretix Array Professional	http://www.phoretix.com	Nonlinear Dynamics	A combination of Phoretix Array v3.0 and Phoretix Array Database v2.0
Rosetta Resolver 3.0	http://www.rosettabio.com	Rosetta Biosoftware	The Rosetta Resolver system combines advanced analysis software, a high-capacity database, and high-performance server framework in one enterprise-wide tool.
Scierra Microarray Laboratory Workflow System	http://www5.amershambiosciences.com	Amersham Biosciences	A complete management system for microarray experiments and gene expression data.

Stanford Microarray Database (SMD) package	http://genome-www5.stanford.edu/Microarray	Stanford University	SMD stores raw and normalized data from microarray experiments, as well as their corresponding image files. Software provides interfaces for data retrieval, analysis and visualization.
Vector NTI Suite Xpression PathBlazer Prospector	http://www.invitrogen.com	Invitrogen Corporation	Vector NTI suite is a robust protein and DNA sequence analysis package managing data from genomic, short sequence and protein databases. It has a full range of sequence analysis tools Xpression is a software suite to manage storage and analysis of Xpression experiments. PathBlazer is a Petri net based pathways analysis software. Prospector is a protein microarray expression analysis software.

Source: Data derived from Y.F. Leung of Harvard University's Department of Molecular and Cellular Biology. http://ihome.cuhk.edu.hk/~ b400559/arraysoft_database.html

then identify motifs in the sequences that are present in other motif databases. Links for major motif databases can be found in the SwissProt record, but more specialized motifs may not be. The investigator may also want to know if the protein is associated with any diseases. In this case, they might look in the OMIM database for such information (which may or may not be part of the GenBank or Swissprot records). Finally, the scientist might want to analyze the protein as part of a pathway. For this, he or she needs to start comparing a variety of potential identifiers to the identifiers used in the pathway database of interest. The challenge is that the pathway databases have a variety of purposes, from all inclusive tracking of generalized protein pathways to specializing in the recording of data on interactions between in a given organism using a particular experimental technique. The need for managing and tracking of large amounts of data from different data sources can quickly become a non-trivial part of the investigator's time and efforts.

The investigator can also make use of data sources, such as GeneCards (www.genecards.org), which combine data from many different sources into a unified record, or can use specialized programs such as gene annotation tools to correlate information from different databases into a combined record. Regardless, scientists typically want to manage protein information in terms of their putative products and functionalities. This is where ontologies become important.

Scientists typically know about ontologies via the more traditional taxonomic applications. Here every living organism is placed into a descending hierarchical grouping of kingdom, phylum, class, order, family, genus, and species. Genus and species form the name of the organism and the hierarchical nature of the data means that organisms can be compared and analyzed in a standardized fashion. Many organisms have commonly used names that are actually ambiguous. The taxonomic system allows one to be very specific about the type of organisms being referring to.

Gene Ontologies and the GO Consortium

Gene ontologies provide a similar set of terms for describing genes and their products. An early form of ontological classification is the venerable enzymatic classification (EC) system. Under this system, each enzyme is allocated a four-digit EC number, the first three digits of which define the reaction catalyzed and the fourth of which is a unique identifier (serial number). Each enzyme is also assigned a systematic name that uniquely defines the reaction. The recommended names and EC number are referred to in publications. This has a number of benefits: it can eliminate ambiguities in the literature caused by investigators inadvertently using the same name for different enzymes. It can make the literature databases searching more efficient. The EC number can also be used to find ancillary information, such as genes, sequences, properties, and structures in other databases. However, this system became more complex as more organisms were sequenced and their gene products were released to the community. It became necessary to track the origin of the gene as well as EC numbers and other data pertinent to their investigation.

The Gene Ontology (GO) Consortium[5] was set up in 1999 to provide a common framework to describe genes and gene products. The work of the GO consortium allows its members to provide a common set of terms when annotating genomes and allows members to provide unambiguous gene descriptions, simplify database queries by use of a standardized vocabulary, and simplify cross-species comparisons.

The GO consortium provides three levels of gene description:

1. Molecular Function, such as catalytic or binding activities at the molecular level
2. Biological Process, where a series of molecular functions occur in a particular order to effect a biological result
3. Cellular Component, where the protein is part of the cell or a larger assembly or organelle within the cell.

Any given protein can have a number of molecular functions, such as catalytic activities or binding activities. It can also be involved in a number of biological processes, and it can be localized in one or more areas of the cell. A unifying principle within the GO annotations is the entry fields that contain:

- The GO ID, the unique numerical identifier for the entry
- Synonyms, the alternative names for the same term
- Parents and children that reference this annotation
- Modification date containing information on the most recent change to the entry

The data in GO can be thought of as being hierarchical, with more general terms or data branching down into more specific terms or data. Unlike the taxonomic hierarchy, the GO hierarchy permits more than one parent or child. This is necessary to capture the multifunctional aspects of proteins and protein data. The flow of data is always top to bottom in the hierarchy, and this directional principle can be used to prepare one directional graphs or directed acyclic graphs of the data. This, in turn, becomes very useful for searching through the data, as will be discussed later.

Microarray Ontologies and the MGED Society

Shortly after the initial DNA microarray publications, scientists and manufacturers quickly realized the need for standardized methods of describing and replicating the details of microarray experiments. The Microarray Gene Expression Data (MGED) Society[6] was formed in 1999 to establish microarray data annotation standards and create databases to manage microarray data. MGED is an international organization of biologists, computer scientists, and data analysts that aims to facilitate the sharing of microarray data.

The current focus of MGED is establishing standards for microarray data annotation and exchange, facilitating the creation of microarray databases and related

software implementing these standards, and promoting the sharing of high-quality, well-annotated data within the life sciences community. A longer-term goal is to extend the mission to other functional genomics and proteomics high-throughput technologies. MGED's activities are largely organized around six working groups:

1. MIAME — The formulation of the minimum information about a microarray experiment required to interpret and verify the results.
2. MAGE — The establishment of a data exchange format (MAGE-ML) and object model (MAGE-OM) for microarray experiments.
3. Ontologies (OWG) — The development of ontologies for microarray experiment description and biological material (biomaterial) annotation in particular.
4. Transformations — The development of recommendations regarding microarray data transformations and normalization methods.
5. RSBI WGs — Reporting Structure for Biological Investigations Working Groups (RSBI WGs) — A single point of focus for Toxicogenomics, Nutrigenomics and Environmental Genomics domains of application.
6. MISFISHIE — The development of the Minimum Information Specification For *In Situ* Hybridization and Immunohistochemistry Experiments required to interpret and verify the results.

The MIAME standard is designed to describe the minimum information that should be recorded during a microarray experiment in order to allow the data to be understood and to provide enough detail to allow the experiment to be replicated elsewhere. The standard assists in the exchange of data and provides a basis for development of microarray data repositories. Many softwares, journals and repositories comply with MIAME guidelines and it is anticipated that protein microarrays will use many parts of this standard while evolving novel parts to deal with the added complexities of protein data.

The MIAME standard has evolved around two broad areas: array design and experimental descriptions. The organization of the various sub components can be seen in Figure 23.2.

Much of the structure for DNA array and experimental descriptions already matches needs for describing protein arrays. The main points of difference for protein arrays revolve around the handling of data for reporters, the biological sequences that act as targets for the labeled materials that the array is exposed to. In the case of the experimental description, the standard will need to be extended to include information on the type and purpose of the experiment being performed, which will in turn govern data on the labeling or detection modalities used, the nature of the controls, and how they should be used for analysis.

The most recent released versions of the MIAME standard have been extended to include management of spotted protein arrays and transcriptional oligonucleotide arrays.

As can be seen under the listing of the working groups sponsored by MGED, there are working groups for microarray-specific ontologies, OWG, and for specification of software tools for handling and transmitting microarray based information, MAGE.

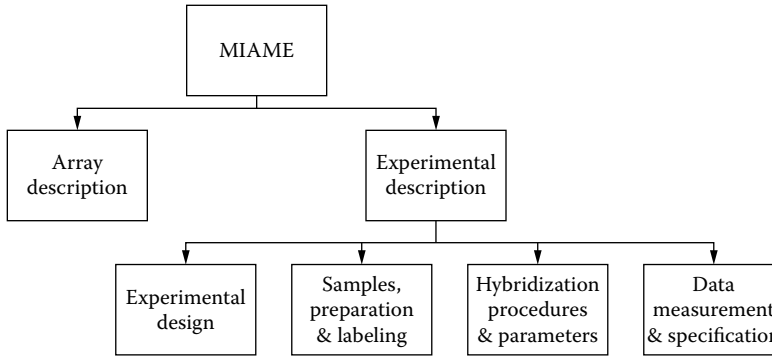


FIGURE 23.2 A simple schematic representation of how the MIAME standard organizes and manages relationships between the different parts of microarray architecture and experimental design. The LIMS depicted in Figure 23.1 shows a design implementation of the MIAME design guidelines.

The MGED Ontology working group provides standard terms for the annotation of microarray experiments, which in turn enables structured queries of unambiguous elements of the experiments. Thus, the work of this group is to turn all the elements relating to a microarray experiment into a hierarchical ontology organized into classes of data, in a similar fashion to that seen with the GO annotations. For descriptions of biological material (biomaterial) and certain treatments used in the experiment, terms may come from external resources that are specified in the ontology. Software programs utilizing the ontology are expected to generate forms that can be used to annotate and populate microarray databases directly, or to generate files in the established MAGE-ML format. So the ontology will be used directly by investigators annotating their microarray experiments as well as by software and database developers and therefore will be developed with these very practical applications in mind.

The MAGE working group is responsible for two main goals. The first is the development of an optimal software model for storing and managing microarray data that is termed the MAGE-OM. This software architecture was developed by breaking down the MIAME standard into software functionalities to correctly handle and pass microarray information within both a computer program and a data repository. The advantage of this software model is that it checks the types of data being entered into the program as it is being entered and can act as a quality assurance and control on the type of data entered and captured during the performance of an array experiment. The second goal of the group has been the creation of a data exchange format based upon the development of a markup language — MAGE-OM. In this case, the group uses the ontology group's work to develop a representation of the data that can be used to exchange data between computers and investigators.

With the addition of protein arrays to the available technologies, the workgroups within the MGED Society will now need to examine the models that they currently

have and decide whether it will be sufficient to replicate and extend existing work to protein arrays or whether it will be necessary to branch off of current work and develop projects to specifically handle protein array data.

HUPO AND MIAPE

The Human Proteome Organization (HUPO) is a second body that will be of importance to protein microarray users. HUPO was formed in 2001 as an international consortium of national proteomics research associations, government researchers, academic institutions, and industry partners. HUPO promotes the development and awareness of proteomics research and establishes scientific collaborations between HUPO members and initiatives. Of interest to microarray researchers is the HUPO Proteomics Standards Initiative[7](PSI), which defines community standards for data representation in proteomics to facilitate data comparison, exchange, and verification. The PSI was founded in 2002 and this initiative has developed standards for two key areas of proteomics: mass spectrometry and protein–protein interaction data. It also develops an integrative General Proteomics Format for the full representation of a proteomics experiment. PSI's Molecular Interaction Standard (MI) for molecular interactions is widely accepted as a standard for the representation of molecular interactions, and is implemented by protein interaction databases such as BIND (www.bind.ca), DIP (http://dip.doe-mbi.ucla.edu/), HPRD (http://www. hprd. org/), and MINT (http://mint.bio.uniroma2.it/mint/Welcome.do). A second activity of this group concerns the development of standardization formats for mass spectroscopy data. The minimum information about proteomics experiments (MIAPE)[8] is concerned with many areas that are of interest to the protein array analyst, including study design, sample generation and sample handling.

Both the MGED and HUPO organizations are in agreement about minimum reporting requirements that overlap with each other. It can be expected that the members of these organizations will coordinate to agree to common standards which will in turn help the users of data to assemble compatible data sets from different domains and software developers to produce software that can manage such diverse data, identify potential quality assurance problems in comparing data sets and manage the storage and retrieval of such data for later analysis.

Revisiting Figure 23.1, we can see how ontologies and standards help to simplify the process of managing protein array data. MIAME provides a framework for describing the types of features that an experimentalist will need to track within a complex experiment. Ontologies help to decrease the redundancy in data, whether it is the representation of information about genes and proteins from GO or the features associated with elements of an array experiment from the OWG. Ontologies can be expressed as part of a markup language, such as MAGE-ML, which can be used for formatting data queries or sending data between computers. Initiatives like MAGE-OM help lay the foundations of the databases and software that are needed to store and handle data for microarray experiments. Finally, complementary initiatives like MIAME and MIAPE help define the shared features and identify differences in these modern complex experimental methodologies, which can in turn be used by scientists and software developers to develop appropriate means of analyzing and comparing different sets of data.

An interesting development in web technology has been the description of the semantic web.[9] This is a new initiative being undertaken by the W3C to further develop the representation of knowledge on the web. The concept of the semantic web is driven by the two elements of knowledge representation and agents. Knowledge representation is performed by providing (1) an ontology describing a set of data and its relationships, (b) a resource description framework, which provides a triplet of data representing subject, properties, and values associated with subject and properties and which can be represesnted in a flexible fashion by XMLs, and (c) a universal resource indicator, which provides a naming convention to identify where data objects are located on a computer, a local network of the world wide web.

The combination of these technologies can result in a user creating a software agent that can go to a number of data repositories, gather data, and perform analyses, and then return the results of such analyses to the user. For example, a query for a sample in an experiment might look like "urn:lsid:uniprot.org:enzymes:2.7.2.3," where "urn:lsid" refers to the local source of data, in this case a LIMS database, "uniprot.org" identifies that the query should be directed to the UniProt database, and "enzymes:2.7.2.3" directs the search against enzymes that have the EC functional classification of 2.7.2.3. In the context of a proteomics researcher's LIMs system, it would provide a number of powerful capabilities to researchers, allowing them to have a number of LIMs dedicated to different specific experimental modalities, such as protein arrays, mass spectroscopy, yeast two hybrid experiments, etc. It could also provide a framework that would enable researchers to query between these different databases and extract appropriate data and analyses.

These technologies are probably a few years away from realization. However, it is interesting to look at how the biotechnology company Invitrogen is utilizing an early version of such technologies. This is illustrated in Figure 23.3. The salient points from this figure are the utilization of ontologies to manage the complexity of the genes and proteins and their subsequent applications; the combination of web and local databases to store, manage and exchange data; the use of defined interfaces and data formats to query and obtain data; and the use of common interfaces to allow users to exchange data from different applications and databases for comparative purposes.

MINING AND VISUALIZATION OF PROTEIN ARRAY DATA

Up to now we have been concerned with the storage and management of array data. However, users will also want to also be able to mine data for interesting relationships. This can encompass simply performing mathematical or statistical analysis of data to utilizing public sources of pathway and interaction data to compare and contrast their experimental results. In this section we will provide an overview of the main approaches to mining and visualizing experimental data.

Data Mining and Statistical Approaches to Proteomics

Data mining is the discipline of analyzing large sets of data for patterns and trends. It allows scientists to find unexpected relationships in the data and to summarize

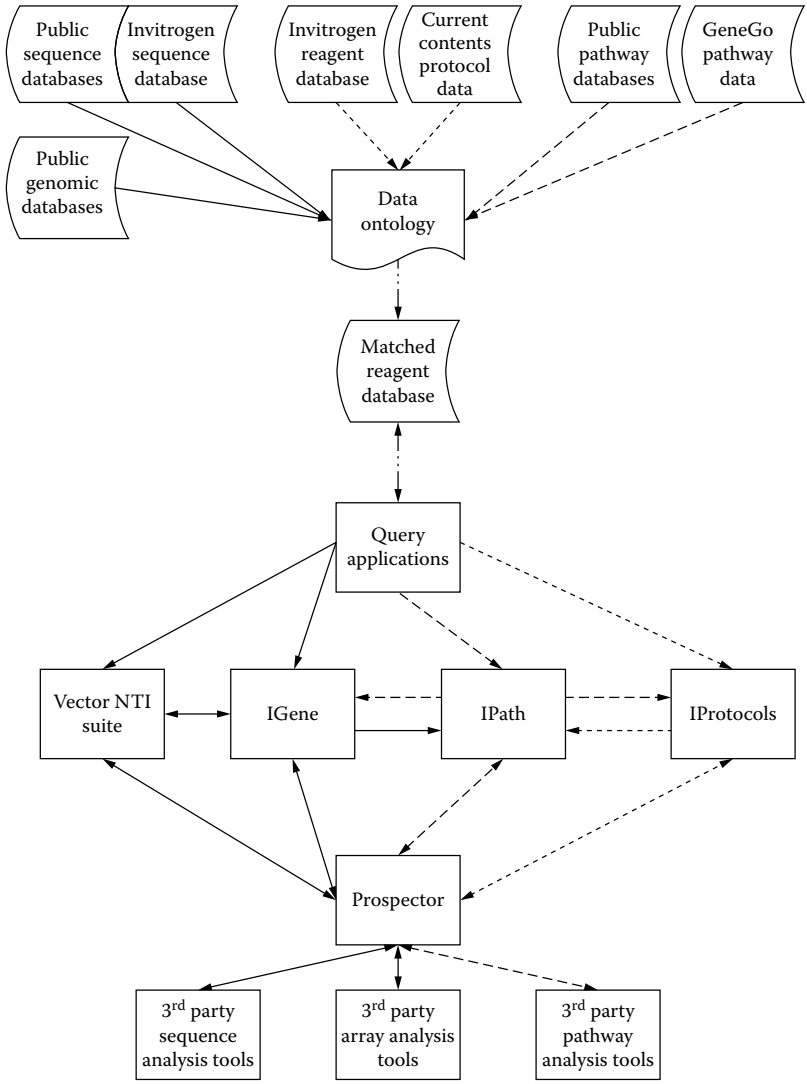


FIGURE 23.3 Schematic view of invitrogen software databases to integrate sequence, pathway and protocol information for protein array analysis via an ontology. Diagram illustrates how a software architecture incorporating both public and private data may be used to develop a semantic web platform to support web and desktop based analyses. A mixture of both public and private sequence, protein pathway and experimental protocols are incorporated into the Matched Reagents database, from which an ontology is derived. A query layer interface is developed to manage query and retrieval of data from the database. A series of web tools interact with this query layer to assist users with queries of the data based upon genes, pathways, or protocols of interest. Prospector, a protein microarray analysis tool, is demonstrated as an intermediary desktop web-enabled software that can be developed to facilitate analysis of protein microarray data as well as providing links to genes, pathways and protocols of interest via the native web-based applications. Protein microarray data or records retrieved via Prospector from web-based tools can in turn be stored and manipulated by appropriate third-party software.

data in ways that are understandable and useful to the scientist. The identification of relationships in data is termed modeling or pattern identification.

The process of data mining can be thought of as falling into a number of approaches. Scientists will typically use one of more of these approaches depending upon their initial understanding of their data set. These approaches include:

- **Exploratory Data Analysis,** where the goal is to explore the data without any preconceived ideas as to what relationships the data will contain. In biology, this is often referred to as a 'fishing expedition.'
- **Descriptive Modeling,** where the goal is to create a descriptive model accounting for all of the data. This is one of the more common starting points for scientists working with proteins that are part of a pathway.
- **Predictive Modeling,** where the goal is for any variable to be able to predict the known values of all other variables associated with it. This is another common starting point for scientists interested in working with known genes or proteins.
- **Discovering Patterns and Rules,** where the goal is to develop models of the data that can be used to test hypotheses and validate models of how biological systems are performing relative to the *in silico* model. Here the goal is to find exceptions or deviations from the model, which can lead to clarification of the model or discovery of new aspects of the model system that were not previously understood.
- **Retrieval by Content,** where the goal is to take a pattern of relationship within the data and see how many other instances of it can be found within the dataset. For biologists, this can mean looking for patterns in known sets of data and applying them to novel genes or novel organisms and seeing if similar relationships are found. This can lead to identification of conserved biological functions between known and new uncharacterized genes or between normal or disease states within tissues or organisms.

Data mining has a very heavy mathematical component, particularly in statistics. One of the advantages of microarrays is that they often generate large data sets, so many of the statistical analyses methods can be applied to this data. The application of statistical analysis to DNA microarray data is well described. This is important, as it suggests that many preexisting data mining approaches can be applied to protein microarray data sets.

Another aspect of data mining is the use of visualization techniques for data analysis. Comprehension of data is often facilitated by its presentation in a graphical rather than a numerical fashion. Biologists can again take advantage of a very large body of investigation into processes and methods for the display and visualization of data. Eisengraphs have been used for DNA microarray analysis and can be informative for protein array data. Use of basic histographs, dot plots, and scatterplots are extensively used in primary data visualization. More complex data types can be represented in contour maps, pie charts, star charts and self-organizing maps.

Temporal data can be represented by three-dimensional graphs, where time is one of the axes for the data.

Array experiments are often qualitative and not highly quantitative. Unless the data demonstrates sufficient qualitative strength to be identified by statistical means, interesting relationships in the data may be lost in the general background noise of the analysis. Here, one would wish to include biological data at an early stage to help establish significance due to known biological data based upon other studies or findings, which would in turn help filter out known relationships in data from the background data.

The display of large amounts of data can be daunting both computationally and for the scientists examining the output of such plots. Methods of displaying data with several dimensions typically rely upon the use of methods such as principal component analysis to reduce the complexity of the data to the most informative portions of the data. However, it is possible for subtle but important relationships within the data to be lost by the use of such techniques. Finally, the display of temporal aspects of complex data sets of several components can quickly become very complex and hard to inspect by eye. If the computer were able to display not only the statistical aspects of the data set but also to include known or inferred biological data, such as pathway data, such relationships might be more easily detectable.

To perform these more advanced analyses, scientists must first identify the types of data that would support their research and render it into a format that facilitates computational searching, assignment of quality to data, identification of search and optimization methods, and modeling or identifying patterns in the data. We will now examine Natural Language Processing (NLP) and Graph Theory as two such technologies to facilitate basic data analysis.

Natural language Processing and Biological Data Analysis

Natural language Processing[10] (NLP) has developed as a means of using computers to process human speech and writing with the ultimate goal of developing automated systems that can correctly extract context and sense from these sources. To do this, two main tasks have been identified:

1. To process written text, computers must be able to use a combination of lexical, syntactic, and semantic knowledge of a given language as well as correctly manage and process any required real-world information.
2. To process spoken language, the computer must use all the above information as well as incorporating knowledge about phonology and any other ambiguities in how a given language is spoken or understood.

A number of databases cover different aspects of protein function, protein–protein interactions, regulatory genetic networks, or signaling pathways. In addition, a large number of electronic journals and other electronic data sources contain large amounts of biological data that would be useful to incorporate into such analyses. Databases such as KEGG, DIP, BIND and TRANSPATH (www.transpath.com) all have well

defined data formats that can be used as a source of data for development of NLP systems. One point to note is that the figures or diagrams are generated by these databases are not searchable by NLP approaches. The lack of structure currently associated with generation of figures and diagrams of pathway or interaction data typically excludes them from utility in these types of searches.

Scientific papers are published within a fairly uniform set of guidelines, but certain portions of papers tend to be denser in experimental data such as results and figure legends, while other portions tend to be more discursive, such as the introduction and discussion sections of papers. MedLine is an example of a database of abstracts that could be searched for information on protein interactions. This makes it a good source for the development of NLP technologies.

A number of approaches have been developed over the last few years to use NLP methodologies for mining protein data. One such system, used by the MedScan information-extraction system, is shown in Figure 23.4. The advantage of such a system is the accuracy with which it can extract meaningful biological relationships from abstract data. This engine and others like it have been applied to many of the pathway databases as well. The production of XML-based output means that the data can be parsed and queried by downstream applications or used to store the results of such analyses into associated databases. In addition, the XML is amenable for parsing with public or private ontologies. Thus, the ontologies that the scientist may use for the LIMs system described earlier can be used to parse and extract MedLine and pathway data.

One utility of such an NLP engine and its product is the ability to parse data for quantitative and qualitative relationships in the data. Thus, one can analyze abstract data for the numbers of independent publications that refer to a given interaction. Such measurements can be stored within the associated database and weights can be given to the association of proteins with substrates, with other proteins, with compartments within the cell, and so on. Such weights can then be applied to pathway databases to provide qualifiers on the degree of support of each individual step of a given pathway. Such weights can be used in statistical studies either as measurements of distance and correlation or as building blocks for the development of weight matrices, probability based Markov Models and other complex statistical data structures. Such data representations can then be used as a portion of earlier statistical calculations and the accuracy of detection and resolution of biological data can be enhanced.

A major advantage of the NLP-based approach is that of reducing both lexical and contextual complexity in data. As we saw earlier, one major advantage of ontologies is the reduction of synonyms. An NLP engine that can automatically detect synonyms and reduce them to a single root term is a tremendous asset, particularly as it reduces the reliance of scientists on the services of experts to develop and verify sets of terms within ontologies. NLP engines can instead highlight difficult-to-resolve synonyms allowing experts to concentrate on such problems while correctly sorting and assigning more easily resolved data.

One concern with NLP-derived data is that the strength of correlations between terms can often just be a confirmation of existing data. Finding novel or less published pathways can be an issue in such systems. In such cases the scientist must

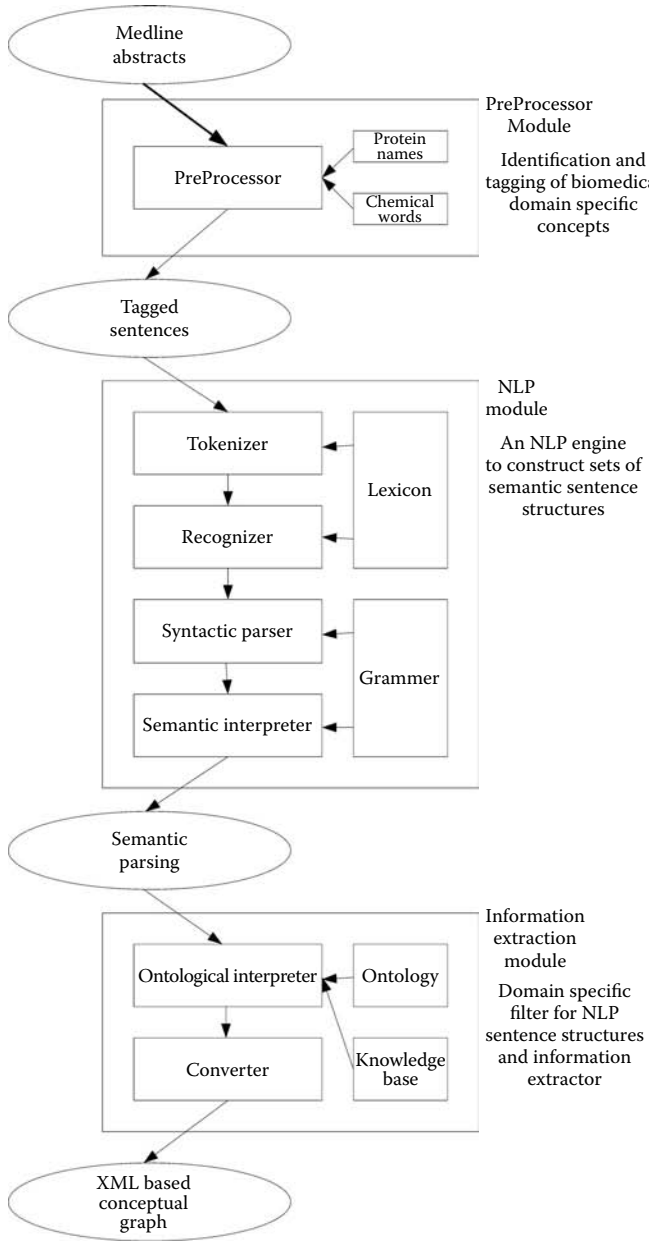


FIGURE 23.4 Ariadne Genomic's MedScan NLP Engine as an example of a full sentence parser for biological data. This diagram illustrates the functionality present in the three modules of the MedScan NLP engine. The PreProcessor module manipulates and identifies biologically relevant data from MedScan abstracts. The NLP module performs the task of identifying relationships between biological terms. The Information Extraction module interprets semantically parsed text for ontological relationships and produces an XML output that can be stored or further manipulated. Further details of the functionality of the engine is based on Daraselia et al.[20]

resort to utilizing the NLP engine output in the absence of weighted terms, which can result in large search sets to analyze.

An additional point concerns the inclusion of a scientist's own pathway data, particularly when the scientist wants to develop a novel pathway or create variations on an existing pathway. In this case, the NLP engine should be designed to allow for the inclusion on privately developed data. This in turn means that the scientist must utilize a methodology for standardizing the creation and input of such novel data into the NLP engine. When querying such data, the scientist must be able to select a combination of private and public data and if necessary assign precedence in its reporting and analysis. Finally, scientists must be able to report and publish their data in a standardized fashion. Currently, there are no agreed standards for reporting of such NLP-derived data that is recognized by the major research journals. This situation will change over the next few years as this type of data becomes more important in research efforts.

Another difficult-to-model aspect of biological data is the presence of concurrent and time-related events. Unless such terms are explicitly included and sought for in the data set, they can frequently be overlooked. So NLP data sets tend to be an excellent resource for identifying the likely correspondence of two terms with each other but may miss additional data of scientific interest. For instance, two proteins that act upon a small molecule may be successfully found but the speed with which the two proteins catalyze the molecule or the fact that this is occurring in two different time periods may be overlooked by the NLP engine.

NLP engine-derived data can be made more amenable to mathematical and statistical analysis by conversion into a graphical representations. Graphical representations can be used to model aspects of biological data that are not captured by NLP engines. Use of graphical approaches with biological data will be examined in the next section.

Graph Theory, Petri Nets, and Biological Data Analysis

Graph theory is one of the fastest growing areas of mathematics today. It offers many advantages to scientists for representing and analyzing biological data. A graph is represented as a series of points, termed vertices. Lines or edges are used to connect vertices together. Edges can be connected by edges that point in one or both directions. Graphs using such directed edges are known as digraphs. Different types of graphs exist based upon the means used to connect vertices with edges. One of the advantages of graph theory is that it correlates very closely to set theory. Thus, a graph can be used to connect sets, and sets can in turn be used to create matrices, thus permitting analysis of graphs with linear algebra techniques.

Petri Nets[11] are a promising tool for describing and studying systems that are characterized as being concurrent, asynchronous, distributed, parallel, nondeterministic, or stochastic. As a graphical tool, Petri nets can be used as a visual communication aid similar to flow charts, block diagrams, and networks. In addition, data about the state of the system, such as whether all the catabolites, cofactors, and enzymes are assembled prior to catalysis, are used in these nets to simulate the dynamic and concurrent activities of systems. As a mathematical tool, it is possible to set up state equations, algebraic equations, and other mathematical models governing the behavior of systems.

In Petri net models, there are places, transitions, and arcs. Transitions and places are specialized vertices. A place contains a token representing information about the system. Transitions model activities that can change the state of the system. If a transition changes, this affects the token associated with a place. Arcs go between places and transitions, representing the flow of information within the graph.

As an example, an initiation place may contain tokens representing an enzyme and a cofactor. For the system to change, the initiation place will need a small molecule token to be introduced via a small molecule containing transition to change its state. Such a token is introduced to the initiation place via an output arc coming from a small molecule transition to the initiation place. When the initiation place receives the small molecule, it changes its tokens to zero and passes the catalytic complex to a new catalytic transition via an input arc. Completion of catalysis is represented by passing tokens to the catalytic completion place from the catalytic transition via another output arc. This event may occur only if the completion place is already empty of its tokens. At the completion place, the enzyme and cofactor may be returned by to the initial place via another input and output arc connected via a recycling transition to the initiation place. The initial place gains tokens for the newly returned enzyme and cofactor and now waits for the addition of the next catabolite via another small molecule transition.

The concept of timing can be introduced to study performance, dependability and competition between places, and transitions within the Petri net. This is typically done via firing delays. If the delay is random, the Petri net is termed stochastic. The net can also have immediate transitions with no delays, exponential transitions with exponential delays, and deterministic transitions with fixed delays. This ability to modulate the activities associated with points and transitions makes Petri nets very suitable for representing protein pathways as well as other biological mechanisms.

The first attempt to use Petri nets for modeling biological pathways was made in 1993,[12] giving a method to represent metabolic pathways. This was soon expanded to model metabolic networks. Subsequently, several enhanced Petri nets have been used to model biological phenomena. The stochastic Petri net has been applied to model a variety of biological pathways, such as the response of the δ^{32} transcription factor to a heat shock.[13]

An example of an immediate transition using Petri net-based pathway analysis software can be seen in Figure 23.5. The PathBlazer software was designed against the KEGG, DIP, BIND, Transpath, and Biocyc (www.biocyc.org) data formatted into a proprietary XML representation. Individual components of pathways were termed components. Components were grouped together into reactions. Reactions could in turn be grouped into pathways. Experiments were used to represent experimental data derived from PPI experiments or array data. GO ontologies were used to order the data, reduce the complexity of synonyms within the data sets and facilitate searching and retrieval of data sets. All data was stored with a Petri net representation, so querying the software resulted in assembly of Petri nets within the parameters of the search query. Results of such queries were in turn exportable in an XML representation or displayed within the PathBlazer viewer software. The software thus allowed microarray users to analyze expression or PPI data vs. published or private pathways to confirm or support their hypotheses.

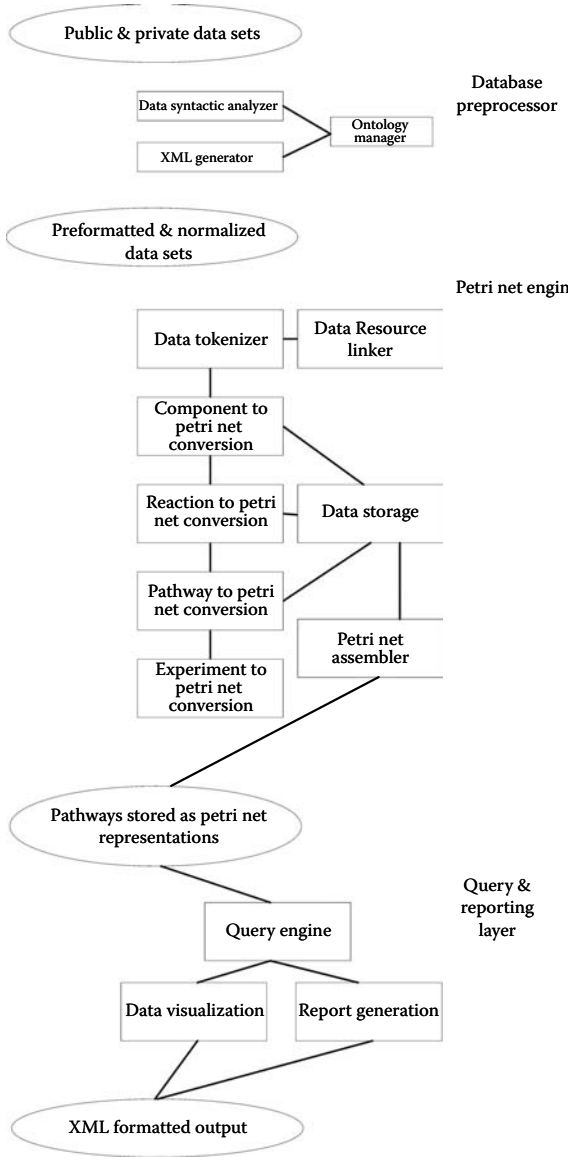


FIGURE 23.5 PathBlazer NLP Engine Architecture. Illustrates the NLP engine used to produce stochastic Petri net representations of pathway data in PathBlazer. Data from KEGG, DIP, Bind, Transpath, or private data sets are processed by the database preprocessor in the first step to generate an XML document describing components, reactions, pathways, and experiments. Preformatted and normalized XML is then analyzed using Petri net engine to develop stochastic Petri net representations of pathway data, which are stored in a local database. Pathway queries are performed via the Query and Reporting engine, using the Petri net engine to assemble networks from available data and sending these representations back to the query and reporting engine for visualization and report generation.

Conclusions about Needs for Data Analyses

The ability to use ontologies with Petri net, NLP, and statistical representations presents a natural transition between these different types of analyses. Primary experimental data from the LIMs can be analyzed and characterized with mathematical and statistical packages. NLP engines can be used to mine databases of pathways and publications to further validate these basic analytical models. Petri nets can be used to render such data in a mathematically and graphically friendly fashion. Ontologies can be used to reduce the complexity of data sets by identifying synonyms and reducing the numbers of terms and objects that software and databases need to store and analyze. Proteomics researchers can use these diverse but complementary tools to build up and develop the representation of their data, provide evidence for known or novel pathways, and provide data ready to apply to modeling technologies. The development of commercial or open source software that permit scientists to export and import data in documented formats between different applications is an important need.

The next step for proteomics scientists concerns the modeling of their data so that it can be developed, tested, and shared with other researchers. This requires the ability to standardize the development and distribution of such models. This introduces the concept of systems biology, which will be the topic of the final section.

SYSTEMS BIOLOGY AS A COMMON PLATFORM TO DEVELOP AND EXCHANGE BIOLOGICAL MODELS

Systems Biology originally arose from the modeling of biological systems, such as predator/prey data, competitive growth of bacteria on a substrate, and evolution of viruses to immunological detection. Such simulations concentrated on abstracting large biological systems into their smallest components. With the onset of the large sets of biological data in the mid to late 1990s, the field evolved to include the representation of large molecular data sets. Scientists quickly realized that sharing large complex models was problematical in a number of ways. First, scientists needed to work with multiple data sets in multiple software packages because there were no means of communicating information in a standardized fashion. Frequently, open source and commercial developers developed their own data formats and exchange mechanisms — or not, as was suitable to their project. In addition, published models were not in a standardized format. So scientists using two different modeling environments had great difficulty in being able to represent the same model in two different software packages. As no one package was superior to all others, scientists wanted to be able to use the software they thought was best, but needed a model representation that would allow them to transition between one software and the next.

When simulators were no longer supported, the models in that simulator became unusable. Hence, there was the need to port models accurately to other newer simulators as needed.

From this dissatisfaction with modeling systems data, the Systems Biology Markup Language (SBML) was born.[14] The SBML consortium is an international representation of academic, government, commercial software vendors and biotechs,

and pharmaceutical companies of all sizes. SBML is a data format to represent computational models such that different software systems can communicate and exchange models in a standardized format. Thus, different tools can have the same representation of the same model, errors in input, and translating the models are reduced or eliminated and all scientists have a common starting point for analyses and simulations.

CellML[15] is a second standard built around an approach of composing systems of equations by linking together the variables in those equations. CellML declares biochemical reactions explicitly and encapsulates arbitrary components into modules. Its focus is on a component-based architecture to facilitate model reuse and the mathematical description of these models.

SBML and CellML represent different but complementary approaches to solving the same general problem. They were initially developed independently, but the developers of both languages are now making the languages more interoperable. SBML Level 2 borrows a number of approaches from CellML, making the formats that much easier to translate between each other.[16]

More than 90 different software packages are compliant with SBML. One example that may be of interest to proteomics researchers is Cytoscape.[17] Cytoscape is a bioinformatics software platform for visualizing molecular interaction networks and integrating these interactions with gene expression profiles and other state data. Additional features are available as plugins, including network and molecular profiling analyses, new layouts, additional file format support, and connection with databases. Plugins can be developed by anyone using the Cytoscape open software architecture. Figure 23.6 shows a view from the Cytoscape software. This modern software architecture of core software modules supplemented by community developed supplemental software modules has been in place for many years now. Notable use of such development practices can be seen through the development of OpenBio projects like BioPerl[18] and BioJava.[19] A significant advantage of such OpenSource drive projects over use of commercial software is that these projects exist as long as the participants are prepared to contribute to it. BioPerl has been under active development since 1992, which is considerably longer than most of the surviving commercial software companies in the bioinformatics field. A drawback of OpenSource projects is that support and help can be very ad hoc and dependent upon the time and good will of the developers in these projects. SBML and its supporting softwares are likely to be around for a long time simply due to the needs of the participants for this type of community effort.

The SBML standard is sufficiently robust to be able to take the various data transformations from initial analysis to NLP analysis to Petri net analysis, and provide appropriate place holders and data types to handle the accumulated data. Proteomics researchers can take their accumulated data and input it into SBML either using defined file formats and community provided SBML generating tools or by constructing their own in compliance with the SBML standard. Once scientists have ported their data into SBML, they can then use any of these 90 tools for further analysis, they can publish their model in a journal, exchange their model with colleagues, extend the model to include data from other researchers or databases, and so on. By having their model available in SBML, researchers can take advantage of new analyses or visualization tools as they become available.

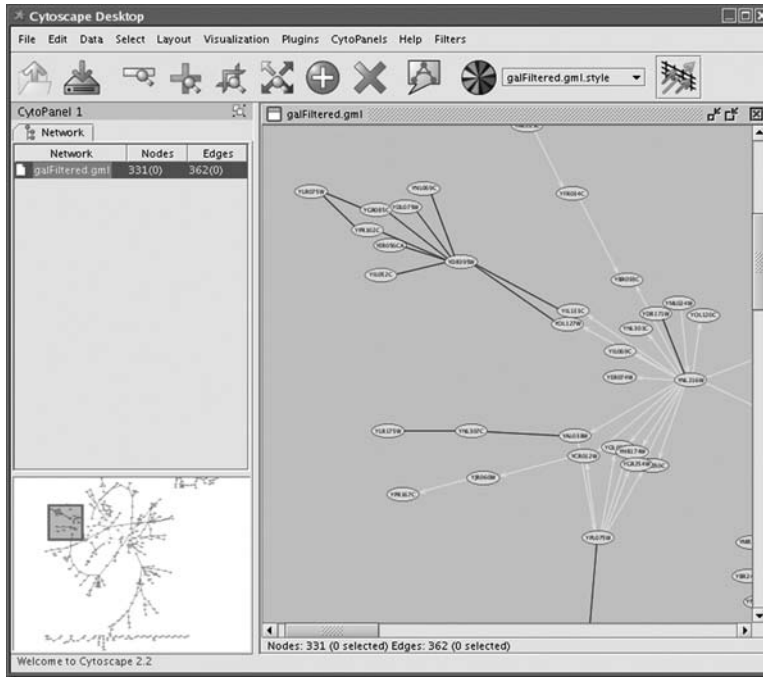


FIGURE 23.6 Screenshot of the Cytoscape Software Package.[17] The screenshot shows the main window of Cytoscape, displaying a network for protein–protein and protein–DNA interactions among 331 yeast genes. See color insert following page 236.

CONCLUSIONS

Many options are available to proteomics researchers who wish to exhaustively analyze and publish their data and expect to have other researchers able to utilize it as part of their own research. Proteomics researchers are able to take advantage of many of the lessons learned by the sequence, genomics, and microarray communities over the last 10 years. This should facilitate the adoption and utility of protein microarrays in common research use. However, researchers must be prepared to characterize and record details of their experimental systems with great care. Doing this in electronic format, such as adopting a LIMS system or Electronic Notebook technologies, will greatly facilitate their later analysis of their data. Researchers can expect to apply a number of different methods to analyzing their data. Ontologies are a unifying technology in this regard and scientists should make every effort to use these extremely valuable tools. Application of NLP to the investigators' data will allow them to take advantage of textual sources of experimental data, whether in the form of electronic journal publications or databases of pathway information. Use of Petri nets and graph-based techniques will allow investigators to prepare data for modeling in a mathematically rich data representation while also allowing them to represent complex data sets containing different experimental components in a visually comprehensible fashion.

Finally, use of the SBML standard will permit the user to participate in a larger community effort of preparing models of biological systems.

REFERENCES

1. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkan., U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., Vingron, M., Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*, 29(4):365–71, 2001
2. Goto, S., Bono, H., Ogata, H., Fujibuchi, W., Nishioka, T., Sato, K., Kanehisa, M., Organizing and computing metabolic pathway data in terms of binary relations, *Pac Symp Biocomput* 175–86, 1997.
3. Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M., Eisenberg, D., DIP, the Database of Interacting Proteins: A research tool for studying cellular networks of protein interactions, *Nucleic Acids Res*, 30(1):303–5, 2002.
4. Bader, G.D., Donaldson, I., Wolting, C., Ouellette B,F., Pawson, T., Hogue, C.W., BIND–The Biomolecular Interaction Network Database, *Nucleic Acids Res,* 29(1):242–5, 2001.
5. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium, *Nat Genet,* 25(1):25–9, 2000.
6. Whetzel, P.L., Parkinson, H., Causton, H.C., Fan, L., Fostel, J., Fragoso, G., Game, L., Heiskanen, M., Morrison, N., Rocca-Serra, P., Sansone, S.A., Taylor, C., White, J., Stoeckert, C.J. Jr., The MGED Ontology: A resource for semantics-based description of microarray experiments. *Bioinformatics*, 22(7):866–73 (Epub) 2006.
7. Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S.G., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C., Apweiler, R., The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data, *Nat Biotechnol,* 22(2):177–83, 2004.
8. Orchard, S., Hermjakob, H., Julian, R.K. Jr., Runte, K., Sherman, D., Wojcik, J., Zhu, W., Apweiler, R., Common interchange standards for proteomics data: Public availability of tools and schema, *Proteomics,* 4(2):490–1, 2004.
9. Jenssen, T.K., Hovig, E., The semantic web and biology, *Drug Discov Today*, 7(19):992, 2002.
10. Swanson, D.R., Searching natural language text by computer. Machine indexing and text searching offer an approach to the basic problems of library automation. *Science*, 132:1099–104, 1960.

11. Pinney, J.W., Westhead, D.R., McConkey, G.A., Petri Net representations in systems biology, *Biochem Soc Trans*, 31(Pt 6):1513–5, 2003.
12. Reddy, V.N., Mavrovouniotis, M.L., Liebman, M.N., Petri net representations in metabolic pathways, *Proc Int Conf Intell Syst Mol Biol*, 1:328–36, 1993.
13. Voss, K., Heiner, M., Koch, I., Steady state analysis of metabolic pathways using Petri nets, *In Silico Biol*, 3(3):367–87, 2003.
14. Ginkel, M., Gor, V., Goryanin, I.I., Hedley, W.J., Hodgman, T.C., Hofmeyr, J.H., Hunter, P.J., Juty, N.S., Kasberger, J.L., Kremling, A., Kummer, U., Le Novere, N., Loew, L.M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E.D., Nakayama, Y., Nelson, M.R., Nielsen, P.F., Sakurada, T., Schaff, J.C., Shapiro, B.E., Shimizu, T.S., Spence, H.D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., Wang, J., SBML Forum. The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models, *Bioinformatics,* 19(4):524–31, 2003.
15. Lloyd, C.M., Halstead, M.D., Nielsen, P.F. CellML: Its future, present and past, *Prog Biophys Mol Biol*, 85(2-3):433–50, 2004.
16. Schilstra, M.J., Li, L., Matthews, J., Finney, A., Hucka, M., Le Novere, N., CellML2SBML: Conversion of CellML into SBML, *Bioinformatics*, 22(8):1018–20, 2006. (Epub 2006 Feb 10.)
17. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T, Cytoscape: A software environment for integrated models of biomolecular interaction networks, *Genome Res,* 13(11):2498–504, 2003.
18. Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock, M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D., Birney, E., The Bioperl toolkit: Perl modules for the life sciences, *Genome Res*, 12(10):1611–8, 2002.
19. Mangalam, H., The Bio* toolkits — A brief overview, *Brief Bioinform*, 3(3):296–302, 2002.
20. Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A., Mazo, I., Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*, 20(5):604–11, 2004. (Epub 2004 Jan 22.)

Index

C

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

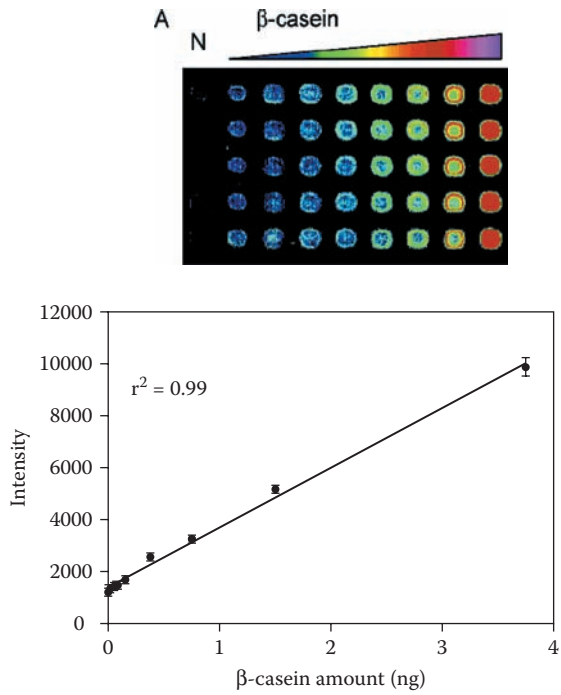
V

W

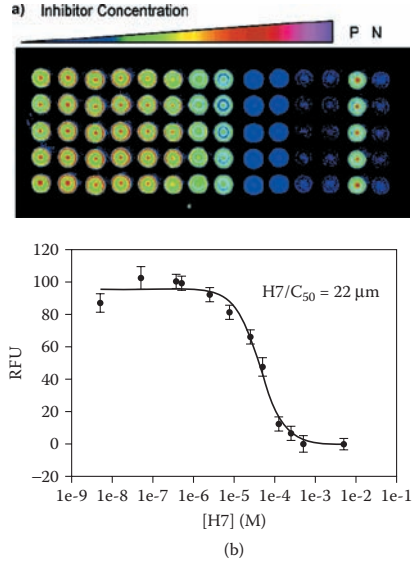
X

Y

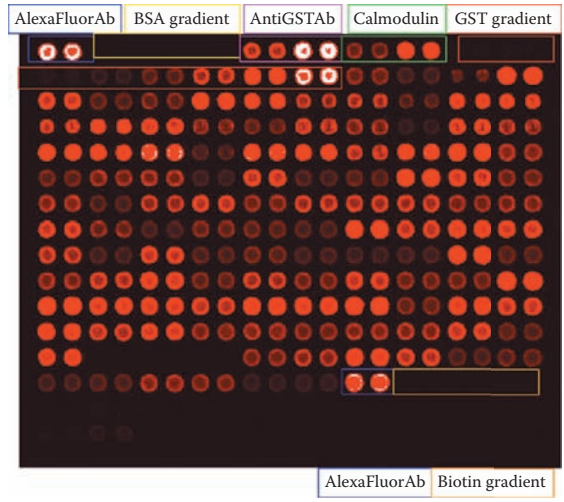
Z



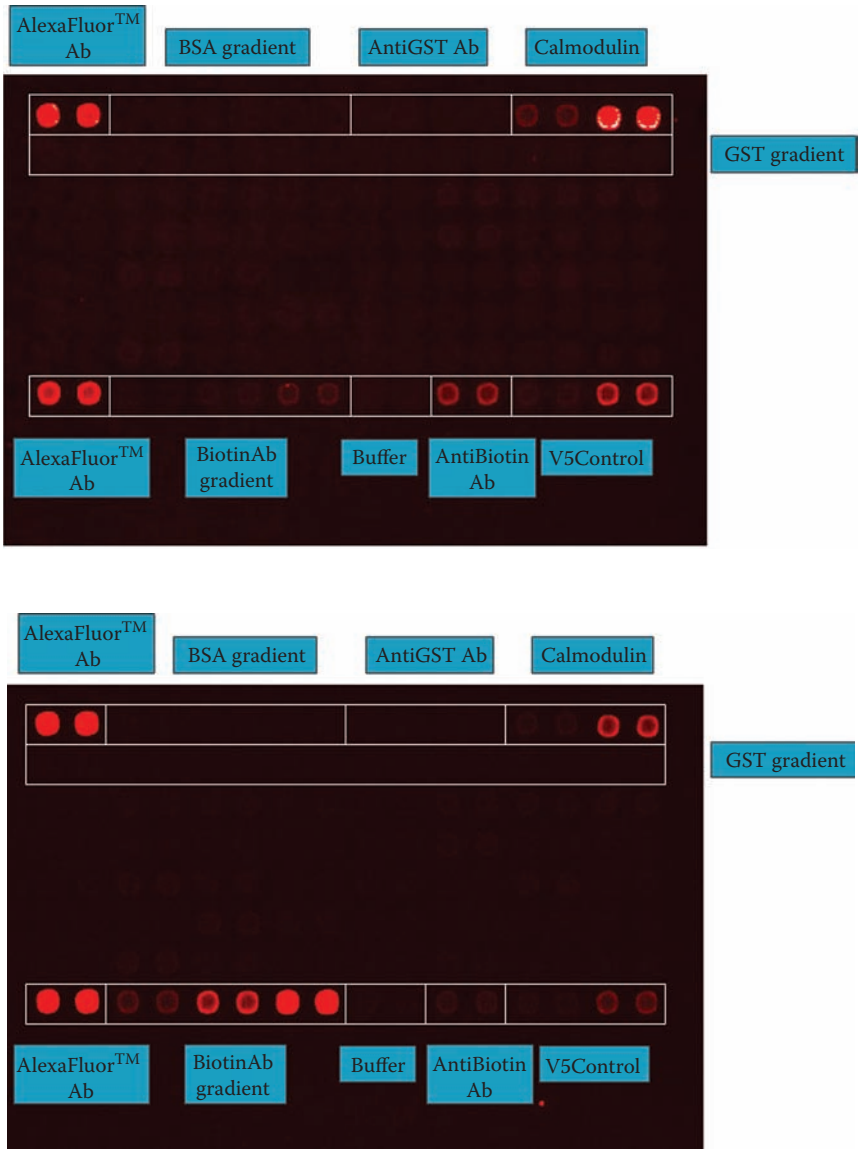
COLOR FIGURE 5.11 Linearity of phosphoprotein detection with Pro-Q Diamond dye within a sol-gel-derived microarray. Panel A shows the fluorescence intensity of the protein gradient on the array. Panel B shows the correlation between signal intensity and amount of protein. [Reproduced with permission from Ref. 68. Copyright 2005 American Chemical Society].



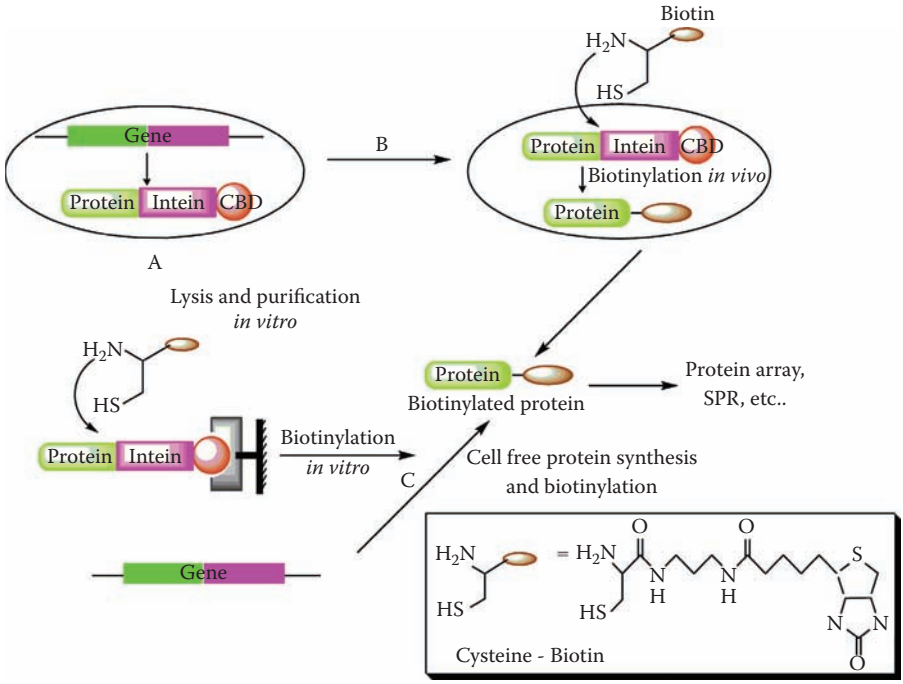
COLOR FIGURE 5.12 (a) H7 IC50 assay performed on a PKA/kemptide array. Inhibitor concentration increases from left to right, resulting in decreased fluorescence intensity due to inhibition of the phosphorylation reaction. N is the BSA negative control, P is the b-casein positive control. (b) IC50 curve generated from the H7 inhibition assay. Background signals from the negative control sample were subtracted and the data was normalized to the maximum intensity obtained in the absence of inhibitor. [Reproduced with permission from Ref. 68. Copyright 2005 American Chemical Society].



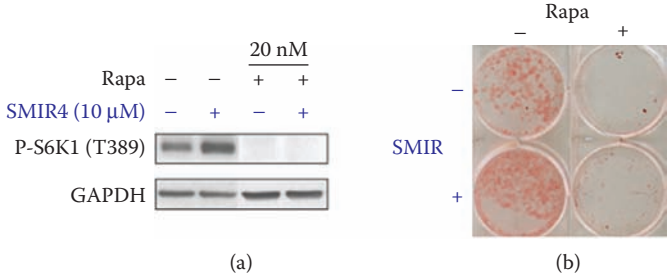
COLOR FIGURE 6.8 The protein array was probed with an anti-GST antibody followed by an AlexFluor 647 labeled secondary antibody. 8A is an image of the entire array and 8B is one of the 48 subarrays. Control proteins are included in every subarray and shown in the boxes.



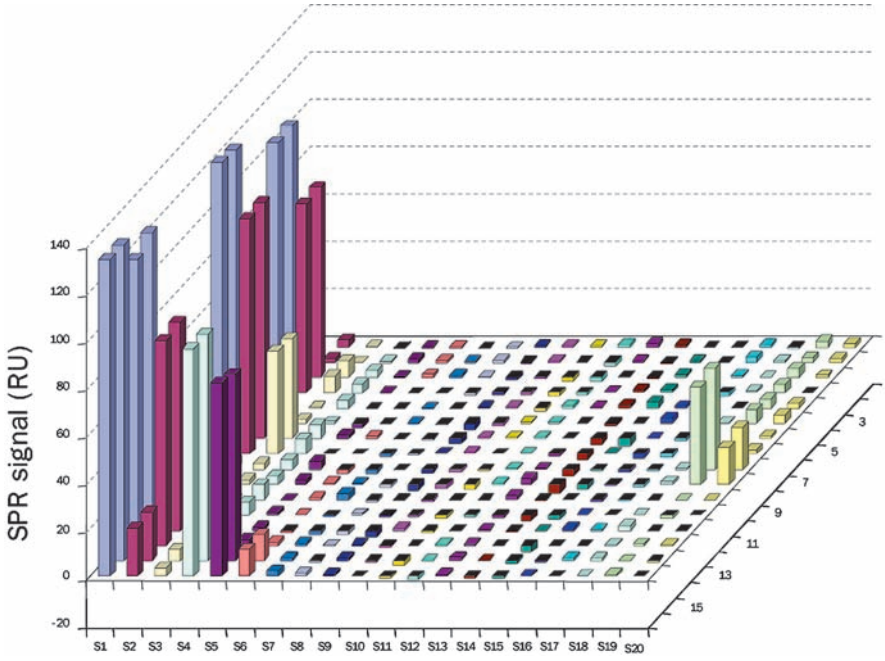
COLOR FIGURE 6.9 Protein arrays from every batch are tested for functionality. Calmodulin kinase was used as a probe to detect its interaction with Calmodulin printed on each array. (a) is a subarray image of the probed array detected with Alex Fluor 647 labeled anti-V5 antibody, and (b) detected with Alex Fluor 647 labeled streptavidin.



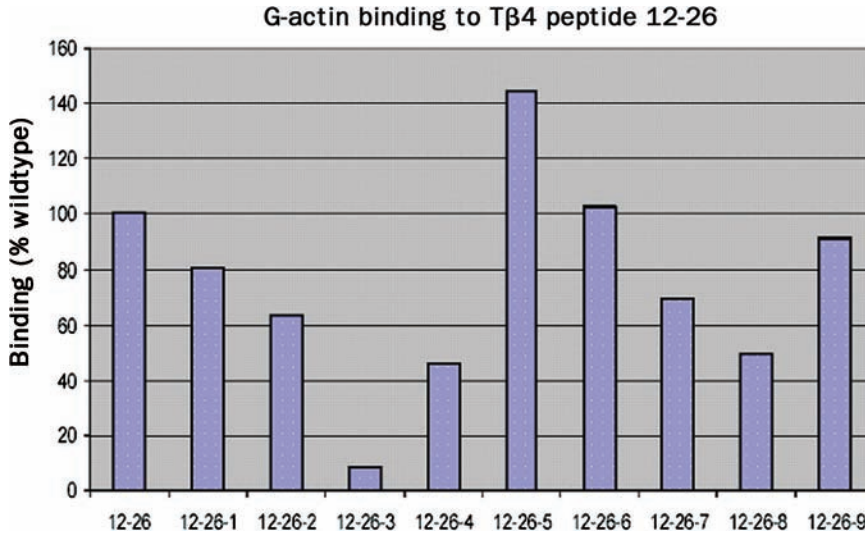
COLOR FIGURE 7.1 Three intein-mediated protein biotinylation strategies: (A) *in vitro* biotinylation of column-bound proteins; (B) *in vivo* biotinylation in live cells; (C) cell-free biotinylation of proteins.



COLOR FIGURE 14.2 (a) SMIR4 effect on S6K1 detected by Western blot analysis using the phosphorylation status of Thr-389 as a readout. (b) SMIR effect on adipogenesis. 3T3-L1 cells are treated with a differentiation cocktail (insulin + IBMX + dexamethasone) in the presence and absence of SMIR, and adipogenesis is assayed by a simple staining method using oil red O, which stains the lipid droplets in differentiated adipocytes.[95]

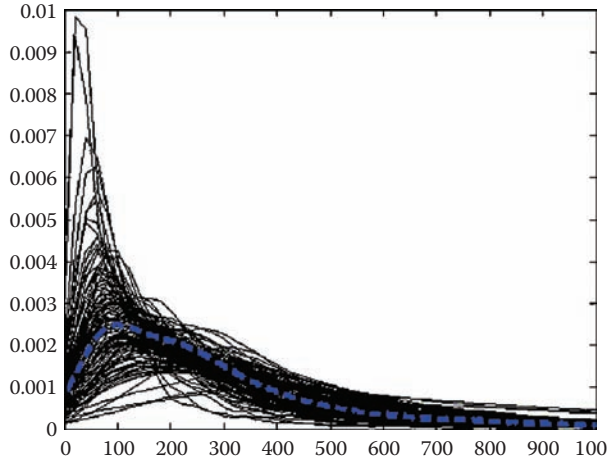


(a)

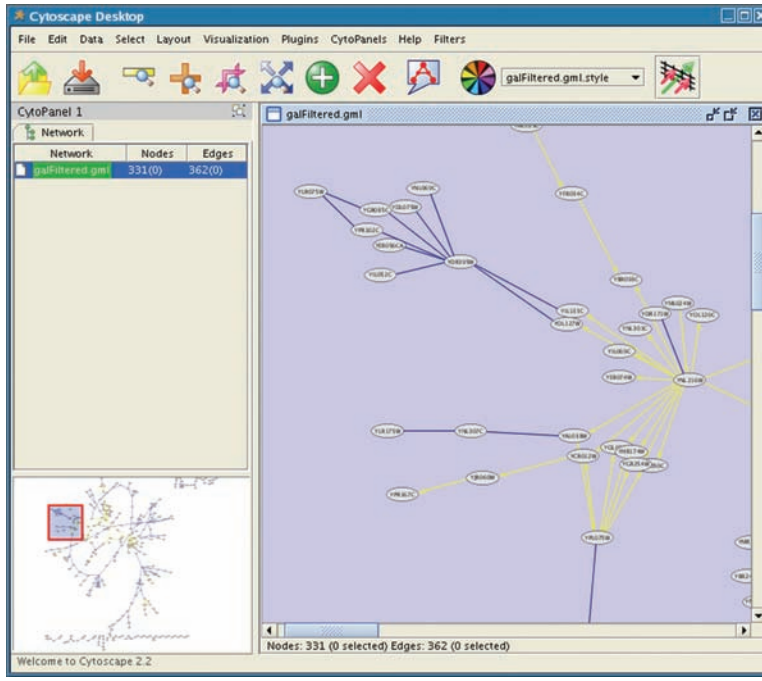


(b)

COLOR FIGURE 10.3 (A) Select which peptides immobilized in an array best bind a protein in solution. (B) Spot peptides containing a single alanine substitution at one amino acid position. Pinpoint which amino acid(s) is (are) critical for the interaction.



COLOR FIGURE 21.2 The heavy dashed line is the resulting quantile normalized distribution all of 127 chips worth of data; the other lines are the distributions of the individual 127 protein arrays.



COLOR FIGURE 23.6 Screenshot of the Cytoscape Software Package.[17] The screenshot shows the main window of Cytoscape, displaying a network for protein–protein and protein–DNA interactions among 331 yeast genes.

Chemistry

Functional Protein Microarrays in Drug Discovery

As central actors in most biological functions, proteins are the subject of intense study. This, in turn, has driven the development of increasingly sophisticated approaches for the study of proteins, which, in recent years, has extended to proteomic level methodologies. At the same time, the number of publications in the field has increased exponentially. However, until now, no book has addressed all aspects of functional microarrays in a coherent and integrated fashion. **Functional Protein Microarrays in Drug Discovery** provides an up-to-date overview of the field and the background required to actually design and develop arrays or perform and analyze array experiments.

Features

- Explores all aspects of functional protein microarrays, including basic principles, methods, and applications
- Discusses the generation of functional protein content
- Describes both standard and state-of-the-art fabrication methods
- Reviews current and next generation approaches to assay detection
- Addresses computational issues, bioinformatics, data analysis and standards, and business aspects

While the field's early successes have set the stage for the rapid growth now being witnessed, it is not without its challenges. Indeed, these challenges are to be expected in a fast moving interdisciplinary endeavor such as this, where molecular biology, protein chemistry, bioinformatics, engineering, and physical sciences intersect. The first integrated reference for functional protein microarrays, this book helps you not only meet the challenges but also excel in your field.

9809

ISBN 0-8493-9809-6

90000

9 780849 398094

www.crcpress.com

CRC Press
Taylor & Francis Group
an informa business
www.taylorandfrancisgroup.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
270 Madison Avenue
New York, NY 10016
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK