METABOLOME ANALYSES

Strategies for Systems Biology

Edited by
Seetharaman Vaidyanathan,
George G. Harrigan and
Royston Goodacre

METABOLOME ANALYSES:
Strategies for Systems Biology

METABOLOME ANALYSES: Strategies for Systems Biology

Edited by

Seetharaman Vaidyanathan
School of Chemistry, The University of Manchester, UK

George G. Harrigan
Pfizer, Chesterfield, MO, USA

Royston Goodacre
School of Chemistry, The University of Manchester, UK

Library of Congress Cataloging-in-Publication Data

A C.I.P. Catalogue record for this book is available
from the Library of Congress.

ISBN-10: 0-387-25239-8 e-ISBN-10: 0-387-25240-1 Printed on acid-free paper.
ISBN-13: 978-0387-25239-1 e-ISBN-13: 978-0387-25240-7

© 2005 Springer Science+Business Media, Inc.
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, Inc., 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now know or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks and similar terms, even if the are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1 SPIN 11054030

springeronline.com

Dedication

To my parents (SV),

To Beth, Sean and Evan
(GGH),

To Elizabeth, Tamara and
Rhozzum Connor (aka. Pickles)
(RG)

Contents

Dedication v

Contributing Authors xi

Foreword xvii

Acknowledgments xxi

1. Introduction 1
Seetharaman Vaidyanathan, George G. Harrigan and Royston Goodacre

2. Towards integrative functional genomics using yeast as a reference model 9
Juan I. Castrillo and Stephen G. Oliver

3. Metabolomics for the assessment of functional diversity and quality traits in plants 31
Robert D. Hall, C.H.Ric de Vos, Harrie A. Verhoeven, Raoul J. Bino.

4. Metabolomics: a new approach towards identifying biomarkers and therapeutic targets in cns disorders 45
Rima Kaddurah-Daouk, Bruce S. Kristal, Mikhail Bogdanov, Wayne R. Matson, M. Flint Beal

5. Comparative metabolome profiling using two dimensional thin layer chromatography (2DTLC) 63
Thomas Ferenci and Ram Maharjan

6. Capillary electrophoresis and its application in metabolome analysis 83
Li Jia and Shigeru Terabe

7. Metabolite profiling with GC-MS and LC-MS 103
Ralf Looser, Arno J. Krotzky, Richard N. Trethewey

8. The application of electrochemistry to metabolic profiling 119
David F. Meyer, Paul H. Gamache and Ian N. Acworth.

9. Differential metabolic profiling for biomarker discovery 137
Haihong Zhou, Aaron B. Kantor and Christopher H. Becker

10. NMR-based metabonomics in toxicology research 159
Laura K. Schnackenberg, Richard D. Beger, and Yvonne P. Dragan

11. Methodological issues and experimental design considerations in metabolic profile-based classifications 173
Bruce S. Kristal, Yevgeniya Shurubor, Ugo Paolucci, Wayne R. Matson

12. Modelling of fungal metabolism 195
Helga David and Jens Nielsen

13. Detailed kinetic models using metabolomics data sets 215
Jacky L. Snoep, Johann M. Rohwer

14. Metabolic networks 243
Eivind Almaas, Zoltán N. Oltvai and Albert-László Barabási

15. Metabolic networks from a systems perspective 265
Wolfram Weckwerth, Ralf Steuer

16. Parallel metabolite and transcript profiling 291
Alisdair R. Fernie, Ewa Urbanczyk-Wochniak and Lothar Willmitzer

17. Fluxome profiling in microbes 307
Nicola Zamboni and Uwe Sauer

18. Targeted drug design and metabolic pathway flux 323
László G. Boros and Wai-Nang Paul Lee

19. Metabonomics in the pharmaceutical industry 337
Eva M. Lenz, Rebecca Williams and Ian D. Wilson

20. How lipidomic approaches will benefit the pharmaceutical industry 349
Alvin Berger

21. Metabolites and fungal virulence 367
Edward M. Driggers and Axel A. Brakhage

Index 383

Contributing Authors

Ian N. Acworth
ESA Inc. 22 Alpha Road, Chelmsford, MA 01824, USA

Eivind Almaas
Center for Network Research and Department of Physics, University of Notre Dame, Notre Dame, IN 46556, USA

Albert-László Barabási
Center for Network Research and Department of Physics, University of Notre Dame, Notre Dame, IN 46556, USA

M. Flint Beal
Weill Medical College of Cornell University, 525 East 68 St., NY 10021, USA

Christopher H. Becker
SurroMed, Inc.,1430 O'Brien Drive, Menlo Park, CA 94025, USA

Richard D. Beger
Division of Systems Toxicology, 2, National Center for Toxicological Research, Food and Drug Administration, Jefferson, AR 72079-9502, USA

Alvin Berger
Icoria Inc. (formerly Paradigm Genetics, Inc), 108 Alexander Dr., Research Triangle Park, NC, 27709, USA

Raoul J. Bino
Plant Research International, Business Unit Bioscience, P.O. Box 16, 6700 AA, Wageningen, The Netherlands

Mikhail Bogdanov
Weill Medical College of Cornell University, 525 East 68 St., NY 10021, USA

László G. Boros
SIDMAP, LLC, 10021 Cheviot Drive, Los Angeles, CA 90064, USA

Axel A. Brakhage
Institute of Microbiology, University of Hannover, Schneiderberg 50, D-30167, Hannover, Germany

Juan I. Castrillo
The University of Manchester, School of Biological Sciences, The Michael Smith Building, Oxford Road. Manchester M13 9PT, UK

Helga David
Center for Microbial Biotechnology, BioCentrum-DTU, Technical University of Denmark, DK-2800 Kgs Lyngby, Denmark

Yvonne P. Dragan
Division of Systems Toxicology, 2, National Center for Toxicological Research, Food and Drug Administration, Jefferson, AR 72079-9502, USA

Edward M. Driggers
Microbia, Inc., 320 Bent St., Cambridge, MA 02141, USA*

**Current address:* Ensemble Discovery Corp., 99, Erie St., Cambridge, MA 02139, USA

Thomas Ferenci
School of Molecular and Microbial Biosciences, University of Sydney G08, N.S.W. 2006, Australia

Alisdair R. Fernie
Max-Planck-Institute für Pflanzenphysiologie, Am Mühlenberg 1, 14476 Golm, Germany

Paul H. Gamache
ESA Inc. 22 Alpha Road, Chelmsford, MA 01824, USA

Royston Goodacre
School of Chemistry, The University of Manchester, Faraday Towers, Sackville Street, P.O. Box 88, Manchester M60 1QD, UK

Robert D. Hall
Plant Research International, Business Unit Bioscience, P.O. Box 16, 6700 AA, Wageningen, The Netherlands

George G. Harrigan
Pfizer, Chesterfield, MO 63017, USA

Li Jia
Graduate School of Material Science, University of Hyogo, Kamigori, Hyogo, 678-1297, Japan

Rima Kaddurah-Daouk
Metabolon Inc. 800 Capitola Dr., Suite 1, Durham NC 27713, USA*

**Current address*: Duke University Medical Center, Department of Psychiatry, Box 3950, Durham NC 27710, USA

Aaron B. Kantor
SurroMed, Inc.,1430 O'Brien Drive, Menlo Park, CA 94025, USA

Bruce S. Kristal
Departments of Biochemistry and Neuroscience, Weill Medical College of Cornell University, 1300 York Ave, NY 10021, USA
and
Dementia Research Service, Burke Medical Research Institute, 785 Mamaroneck Ave, White Plains, NY 10605, USA

Arno J. Krotzky
metanomics GmbH and Co. KGaA, metanomics Health GmbH, Tegeler Weg 33 10589 Berlin, Germany

Wai-Nang Paul Lee
SIDMAP, LLC, 10021 Cheviot Drive, Los Angeles, CA 90064, USA

Eva M. Lenz
Dept. of Drug Metabolism and Pharmacokinetics, Mereside, Alderley Park, Macclesfield, Cheshire SK10 4TG, UK

Ralf Looser
metanomics GmbH and Co. KGaA, metanomics Health GmbH, Tegeler Weg 33 10589 Berlin, Germany

Ram Maharjan
School of Molecular and Microbial Biosciences, University of Sydney G08, N.S.W. 2006, Australia

Wayne R. Matson
ESA, Inc., 22 Alpha Road, Chelmsford, MA 01824, USA

Wayne R. Matson
ESA, Inc., 22 Alpha Road, Chelmsford, MA 01824, USA

David F. Meyer
ESA Inc. 22 Alpha Road, Chelmsford, MA 01824, USA

Jens Nielsen
Center for Microbial Biotechnology, BioCentrum-DTU, Technical University of Denmark, DK-2800 Kgs Lyngby, Denmark

Stephen G. Oliver
The University of Manchester, School of Biological Sciences, The Michael Smith Building, Oxford Road, Manchester M13 9PT, UK.

Zoltán N. Oltvai
Department of Pathology, Northwestern University, Chicago, IL 60611, USA

Ugo Paolucci
Dementia Research Service, Burke Medical Research Institute, 785 Mamaroneck Ave., White Plains, NY 10605, USA

C.H. Ric de Vos
Plant Research International, Business Unit Bioscience, P.O. Box 16, 6700 AA, Wageningen, The Netherlands

Johann M. Rohwer
Triple-J group for Molecular Cell Physiology, Department of Biochemistry, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa

Uwe Sauer
Institute of Biotechnology, Swiss Federal Institute of Technology (ETH) Zürich, 8093 Zürich, Switzerland

Laura K. Schnackenberg
Division of Systems Toxicology, 2, National Center for Toxicological Research, Food and Drug Administration, Jefferson, AR 72079-9502, USA

Yevgeniya Shurubor
Dementia Research Service, Burke Medical Research Institute, 785 Mamaroneck Ave., White Plains, NY 10605, USA

Jacky L. Snoep
Triple-J group for Molecular Cell Physiology, Department of Biochemistry, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa
and
Molecular Cell Physiology, Vrije Universiteit, Amsterdam, The Netherlands

Ralf Steuer
University, Potsdam, Nonlinear Dynamics Group, Am Neuen Palais 10, 14469 Potsdam, Germany

Shigeru Terabe
Graduate School of Material Science, University of Hyogo, Kamigori, Hyogo, 678-1297, Japan

Richard N. Trethewey
metanomics GmbH and Co. KGaA, metanomics Health GmbH, Tegeler Weg 33 10589 Berlin, Germany

Ewa Urbanczyk-Wochniak
Max-Planck-Institute für Pflanzenphysiologie, Am Mühlenberg 1, 14476 Golm, Germany

Seetharaman Vaidyanthan
School of Chemistry, The University of Manchester, PO Box 88, Manchester M60 1QD, UK

Harrie A. Verhoeven
Plant Research International, Business Unit Bioscience, P.O. Box 16, 6700 AA, Wageningen, The Netherlands

Wolfram Weckwerth
Max-Planck-Institute of Molecular Plant Physiology, 14424 Potsdam, Germany

Rebecca Williams
Dept. of Drug Metabolism and Pharmacokinetics, Mereside, Alderley Park, Macclesfield, Cheshire SK10 4TG, UK

Lothar Willmitzer
Max-Planck-Institute für Pflanzenphysiologie, Am Mühlenberg 1, 14476 Golm, Germany

Ian D. Wilson
Dept. of Drug Metabolism and Pharmacokinetics, Mereside, Alderley Park, Macclesfield, Cheshire SK10 4TG, UK

Nicola Zamboni
Institute of Biotechnology, Swiss Federal Institute of Technology (ETH) Zürich, 8093 Zürich, Switzerland

Haihong Zhou
SurroMed, Inc.,1430 O'Brien Drive, Menlo Park, CA 94025, USA

Foreword

The value of obtaining information on entire classes of analytes is now widely recognized among biological researchers. This unbiased ('omic) approach allows for observation of whole systems, and it is being employed in myriad applications spanning the entire spectrum of biology. There is, of course, no substitute for the hypothesis-driven experiment in validating new concepts. With an 'omics approach, however, it is possible to develop hypotheses for testing from an astonishingly complete understanding of a system and to monitor the results of hypothesis-driven experiments in a far more comprehensive fashion. Unbiased research was developed and most enthusiastically embraced by the genomics community. Looking back on the 'omic revolution from the future we might expect to observe that genomics defined a new course for biological research and made many fundamental advances in biological knowledge. It would not be surprising, however, to find that most of the practical tools developed through 'omics research were developed by applying the principles of genomics to profiling metabolites. Metabolites are particularly valuable for practical applications because they represent the integrated consequence of endogenous metabolism and the response to environmental stimuli. Thus, metabolic profiling provides a method for gaining insight into how biological entities function and into how they adapt or fail in the context of their surroundings. Profiling metabolites is not a new concept- metabolites have been used as useful indices of phenotype for many decades- but the improved analytical and informatic technologies exponentially increase the power of the approach. Research fields that have and will continue to benefit greatly from metabolomic profiling include functional genomics, nutrition, metabolic disease research, clinical care, drug discovery and development, agricultural biotechnology

and toxicology to name a just few. A major advantage for metabolic profiling over other 'omic strategies in advancing our understanding of these fields is that metabolites are inherently linked to phenotype and, importantly, 100 years of biochemical knowledge has been assembled around biochemical pathways. This latter point should allow a much faster translation of profile data to knowledge than is possible with genomics.

Advances in metabolic profiling have been driven in large part by improved analytical and informatics capabilities. The previous volume of this book outlined several of the primary technologies for profiling metabolites including mass spectrometry and NMR. While mass spectrometry and NMR will continue to serve as the core technologies for broad-based metabolic profiling schemes, the goals of metabolic profiling (generating quality data on a wide variety of metabolites simultaneously) do not favor any analytical platform over another. Older chromatographic platforms are equally likely to find use in this field, depending on the biological applications. This edition contains further examples of techniques and applications for spectrometry and NMR, but also contains several examples of new analytical technologies. While the advances in metabolic profiling capabilities are undeniable, the next phase of development for the field should encourage a broad range of researchers to adopt this obviously powerful research strategy. Only proof-of-principle biological results can accomplish this, and it is these examples the current practitioners of metabolic profiling should pursue.

While metabolic profiling has many advantages over genomics and proteomics in terms of utility, it is not without its own set of pitfalls and tradeoffs. Metabolites possess such an astonishingly broad spectrum of physical and chemical properties that no single analytical platform has, or is likely to, accurately quantify and identify all metabolites simultaneously from a biological sample. This fact forces some degree of compromise on the part of researchers, who can choose to trade quantitation for analytical breadth or *vice versa.* In general, research striving to be as inclusive as possible, and therefore sacrificing some degree of accuracy or the identification of compounds, is termed unbiased metabolomics. Research striving to be as accurate as possible on a known subset of the metabolome is termed focused metabolomics. There are also difficulties in the interpretation of data once they are generated. High-content datasets are notoriously prone to produce false discoveries as a result of the number of predictors relative to the degrees of freedom, and metabolic profiling is not exempt from this problem. As metabolic profiling matures, innovative solutions to these problems need to be developed.

Since the publication of the previous volume of this book, the National Institutes of Health announced the NIH Roadmap which outlines the key

themes and initiatives the NIH feels will advance public health in the coming years (Zerhouni, 2003). Among the initiatives singled out in the Roadmap for attention and, critically, public funding is metabolomic research and analytical technology development. The fact that the NIH has chosen to publicly back the concept of metabolic profiling and to commit to funding the development of new technologies is an indication that the field is entering a new phase of development and growth. The growing interest in metabolic profiling in the academic community is another sign that the field is beginning to mature. A keyword search on PubMed using the common terms for metabolic profiling demonstrates the rapid acceleration of publication in the field. While the number of papers meeting these search criteria (just shy of 1,000 as of this writing) lags far behind similar results for genomics, transcriptomics and proteomics, there are many signs that metabolome analyses will catch up in the coming years. Several prominent peer-reviewed publications are actively recruiting manuscripts involving metabolomic research and the new journal *Metabolomics* will begin publishing manuscripts in early 2005. These developments point to a recognition of metabolic profiling/metabolome analyses as an emerging, and important, new field.

It is undeniable that, at the time of this printing, capital investment in biochemical profiling and the publications produced by the approach lag far behind those for genomics, transcriptomics or proteomics. There are many encouraging indications that this disparity will not persist for long. The adoption of biochemical profiling as a central discovery platform should accelerate dramatically as more researchers enter the field, as access to grant money and investments continues to increase, and as proof-of-principle biological results develop and become widely recognized.

Zerhouni E. The NIH roadmap. *Science* 302: 63 (2003).

Steven M. Watkins

President and CSO
Lipomics Technologies, Inc,
West Sacramento, CA 95691

Acknowledgments

SV thanks the University of Manchester and the UK BBSRC for the opportunity and financial assistance. Contributions to the cover design by Sukanya is gratefully acknowledged, as is the help provided by present and past members of the research group, including Irena Spasic, Consuelo Lopez-Diez and Steve O'Hagan, at various times during the compilation of this volume. GGH acknowledges Margann Wideman of Pfizer for her continued support. RG would like to thank the University of Manchester and the UK BBSRC for allowing the academic freedom and financial assistance to investigate metabolic profiling. Heartfelt thanks are also expressed to all present and past members of the research group for their hard work and enthusiasm. Needless to say the editors are greatly indebted to all the authors for their invaluable contributions, without whom this volume would not have been possible.

Chapter 1

INTRODUCTION

Metabolome analyses for systems biology

Seetharaman Vaidyanathan[1], George G. Harrigan[2] and Royston Goodacre[1]
[1]School of Chemistry, The University of Manchester, Faraday Towers, Sackville Street, P.O. Box 88, Manchester M60 1QD, UK. [2]Pfizer, Chesterfield, MO 63017, USA

We are currently in a phase of scientific enquiry that is increasingly driven by the need to analyse biological systems much more holistically. Much of the excitement with respect to this need is due to the realization among practitioners of the traditional reductionist approach, including biochemists and molecular biologists, that there is more to biological systems than can be adequately accounted for by reductionist enquiries alone. Although not entirely novel, a 'systems' perspective in biology affords challenges and prospects which are only now being fully addressed in detail.

Tracking changes in the metabolic complement of the system (the low molecular weight component – the metabolome) that relate to its behaviour is progressively gaining momentum (Oliver *et al.*, 1998; Tweeddale *et al.*, 1998; Fell, 2001; Fiehn, 2001; ter Kuile and Westerhoff, 2001; Harrigan and Goodacre, 2003; Goodacre *et al.*, 2004; Kell, 2004). This particular aspect forms the subject matter of this edited volume. Following in the footsteps of its predecessor (Harrigan and Goodacre, 2003), this volume is compiled to give an overview of the scientific activity that is in progress in this particular field of enquiry. It is by no means comprehensive, but is aimed at capturing the excitement of the current practitioners of the field and relates to their experiences. In keeping with this objective, the authors' views are preserved and presented with minimal edits. Consequently, while the appearance of similar views strengthens its foundation, the appearance of conflicting views only reflects the growing nature of the field and emphasizes the need for active discussions that are inevitable in any emerging field.

1. THE PANOMICS ROUTE TO SYSTEMS BIOLOGY

The central dogma of molecular biology over the last few decades has advocated that the flow of information from the genes to function (or phenotype) is linear and is translated through transcripts, then proteins and finally metabolites. Most scientists have tended to analyse these in isolation with little emphasis on cross-talk between these different levels of molecular organisation. By contrast, the central dogma of systems theory dictates that there is more to a system than the sum of its parts. Indeed, the interaction of a system's parts can result in an emergent state that is not adequately accounted for by investigating the parts independently of each other (Weiner, 1948; Bertalanffy, 1969). Systems biology thus attempts to account for biological system behaviour that cannot be adequately explained by investigations at the molecular level alone (Ideker *et al.*, 2001; Kitano, 2001). Two routes to the evolution of this thinking within biological scientific enquiry can be identified (Levesque and Benfey, 2004; Westerhoff and Palsson, 2004) – i) the panomics route that relies on the generation of high-throughput data on the components of the system (the parts list) and ii) *in silico* routes that attempt to provide information on the interactions that the parts of the system might be involved in to effect a function.

The panomics route to systems biology has its roots in molecular biology. Molecular biology investigations over the past few decades have resulted in the identification of the molecular make-up of cells and the construction of a likely route to the storage, replication, processing and execution of information within cells. A linear hierarchy, in which information is stored in DNA, processed by RNA and proteins, and executed by proteins and metabolites, has become the basis for our understanding of cellular function. Consequently, it has become essential to catalogue these molecular entities in order to understand system behaviour. The genomic era ushered in large-scale DNA sequencing of living organisms, with the aim of explaining biological complexity and versatility in terms of genetic make-up. However, it is now known that whilst a few thousand genes can code for a eukaryotic cell (6000 for yeast (Goffeau *et al.*, 1996)), only two to three times as many is required to construct an entire multicellular organism (Bird *et al.*, 1999) and as little as five times more is required to construct a human being (McPherson *et al.*, 2001; Venter *et al.*, 2001). In addition, discoveries such as short-term information storage in proteins (Bray, 1995), the significant role of post-transcriptional and post-translational modifications in cell function, and the existence of metabolite-mediated regulation of cell function (Winkler *et al.*, 2004), now serve to question the rigor of classically defined hierarchical organisation and illustrate the limitations of genomic

enquiries. Clearly, it has become essential to catalogue other players in the cell factory to define gene function in the post-genomic era. This has now given birth to trancriptomes, proteomes and metabolomes, each relating to the make up of the cell associated with the respective components, RNA, proteins and metabolites.

Whilst transcriptomic and proteomic investigations are facilitating gene-function and annotation efforts, metabolomic investigations are lagging behind. An overview of the gains to be had by directing investigations at the metabolome level is provided in the following three chapters which address microbial (Chapter 2), plant (Chapter 3) and animal (Chapter 4) systems. These chapters also set the scene by providing an indication of the scope and context of metabolome analyses as applicable to different biological systems

Castrillo and Oliver (Chapter 2) elegantly provide the justification and need for directing enquiries at the metabolome level, taking a microbial system, the 'well characterized' yeast, as their model system. The complexity and metabolic diversity of plants, especially with respect to secondary metabolites, offers unique challenges to the characterization of their metabolomes. Hall and colleagues introduce us to some of these aspects in Chapter 3, and discuss metabolome analyses as applied to plant systems. In the following chapter Kaddurah-Daouk and colleagues give an insight into the application of metabolome analyses to the identification of (surrogate) biomarkers and therapeutic targets in animal systems, elaborating on issues pertaining to the study of disorders of the central nervous system.

1.1 Strategies for capturing metabolome-wide changes

Various strategies and challenges pertaining to the tracking of metabolome-wide changes in different biological systems under different application contexts are discussed in the next seven chapters (Chapters 5-11). Most strategies for capturing comprehensive metabolomic data employ a separation technique followed by sensitive detection, typically using mass spectrometry (MS). Separation techniques include two-dimensional thin layer chromatography (2D-TLC), capillary electrophoresis (CE), gas-chromatography (GC) and liquid chromatography (LC). Whilst the objective in such strategies is to capture comprehensive metabolome-wide changes, often the nature of the techniques and sample preparation protocols bias the type of metabolites detected, restricting the analyses to sub-metabolomes. Ferenci and Maharjan discuss the development and application of 2D-TLC in the context of profiling microbial metabolomes (Chapter 5). This is an economically viable solution, useful for comparing metabolomes. CE strategies are discussed by Jia and Terabe (Chapter 6), with respect to, but by no means restricted to, microbial metabolomes. In Chapter 7, Trethewey

and colleagues give an overview of current practices in GC-MS and LC-MS approaches to profiling metabolomes, as applicable to plant, microbial and health care investigations. The development and application of electrochemical techniques in combination with LC separations is discussed in Chapter 8 by Ackworth and collegues. Zhou and colleagues elaborate on the application of LC-MS strategies in Chapter 9 with emphasis on biomarker discovery using MS, within a clinical and drug discovery and developmental context.

Whilst comprehensive analysis would be informative for gaining metabolome-wide knowledge of the system, there are instances when capturing dominant changes in the metabolome through the detection of changes in a few metabolites as biomarkers can provide sufficient information for identifying system wide disturbances. These are usually effected with fingerprinting approaches that involve the direct detection of the system-wide changes with minimal sample pre-treatment or analyte separation, usually with the application of MS, nuclear magnetic resonance (NMR), Fourier transform infrared (FT-IR) or Raman spectroscopies (Harrigan and Goodacre, 2003; Goodacre *et al.*, 2004). In Chapter 10, Beger and colleagues discuss analytical strategies using NMR, highlighting its application in toxicology investigations.

A characteristic feature of 'omic approaches is the parallel and simultaneous high-throughput analysis of several analytes. This places unique demands on experimental design, with the requirement for careful considerations of biological, analytical and data processing issues. Kristal and colleagues (Chapter 11) elaborate on some of these issues and share the lessons they have learnt from metabolic profiling of a model nutritive system in animals.

2. METABOLIC INTERACTIONS FROM A SYSTEMS PERSPECTIVE – THE *IN SILICO* ROUTE TO SYSTEMS BIOLOGY

A metabolomic "parts" list will benefit functional genomic investigations, and can be associated with system-level perturbations. However, knowledge of gene function or, as identified earlier, a catalogue of all the genes, transcripts, proteins and metabolites associated with a system is unlikely to suffice in explaining system behaviour. In addition to establishing which components are involved in a given cellular or biological event, systems-level understanding requires information on how the different components interact to influence system behaviour. A second route to

systems biology (Levesque and Benfey, 2004; Stelling, 2004; Westerhoff and Palsson, 2004) that deals with *in silico* analysis of cellular processes and systems-level data that aim to capture system structure and dynamics can also be identified. At the metabolome level, this route promises to provide information on metabolic interactions from a systems perspective.

In Chapter 12, David and Nielsen focus their discussion on the construction, properties and application of genome scale models developed for fungal systems, and debate their significance in gaining systems level understanding of cellular function. Snoep and Rohwer (Chapter 13) present kinetic modeling of biological systems and elaborate on the concept of metabolic control analysis.

It is now increasingly recognized that complex entities such as biological systems can be represented as networks, the large-scale behaviour of which, if predicted, would enable the understanding of systems behaviour. Complex interactions of intracellular molecules can be captured by this network concept. Oltvai and colleagues (Chapter 14) discuss metabolic networks, presenting the underlying principles, approaches, and utilization of such information regarding these networks. It has been observed with plant systems that metabolites tend to vary in concert with other metabolites. The resulting correlation in metabolite levels within a data set can be used to construct metabolic correlation networks that can be useful in understanding systems behaviour. Weckwerth and Steuer discuss this aspect in Chapter 15. Another *in silico* route to understanding system behaviour is to combine information available from different 'omic platforms to look for patterns that can be associated with systems behaviour. Fernie and colleagues take this route and describe the pair-wise analysis of transcript and metabolite profiles to study potato tuber metabolism and discuss the potential of this approach in Chapter 16.

Metabolic flux ratio analysis can provide information of metabolic network operation, as opposed to network composition. In Chapter 17, Zamboni and Sauer describe flux ratio analysis and discuss the potential of comparative fluxome profiling, illustrating this type of analysis in microbial systems.

3. THE PATH AHEAD – CONCLUDING REMARKS

The final four chapters (Chapters 18-21) deal with the application of metabolome analyses in different contexts to summarize the potential scope of the technique in different application areas. Boros and Lee, in Chapter 18, detail the utility of stable isotope-labeled approaches (SIDMAP) in capturing metabolic changes. They show how SIDMAP can provide valuable

information in investigations of the effect of endogenous and exogenous agents on intermediary metabolism in tumor cells, and debate the role of metabolic profiling in targeted drug design. In the next chapter (Chapter 19), Lenz and colleagues provide an overview of metabonomic investigations in the pharmaceutical industry and discuss the potential this approach holds in toxicological studies and the study of disease models. Lipids constitute a significant proportion of the metabolic complement of biological systems, and play key roles in its functioning. Berger, in Chapter 20, explains why and how this subset of the metabolome contributes to our understanding of system behaviour. In the final (but by no means less important) chapter of the volume (Chapter 21), Driggers and Brakhage discuss the role of metabolic profiling in the study of fungal virulence and show the value of combining metabolome level data with transcriptome level information for assessing this system.

By now, one aspect of *Systems Biology* can be well appreciated, i.e., that it is *an integrative approach.* The route to obtaining systems level information, be it through molecular investigations or through global analysis of networks and interactions, is clearly complementary, and metabolome level data will have to be analysed alongside data obtained from other 'omic platforms to make meaningful observations on system-wide behaviour. Without doubt, data integration and bioinformatics tools for countering the challenges posed by such integration of data from different platforms will have to be addressed before meaningful interpretations can be made. Not withstanding, the potential in profiling metabolomes and investigating metabolome-wide network behaviour in understanding systems behaviour is clearly evident. We hope that this volume convinces you of this exciting potential and that you enjoy reading it!

REFERENCES

Bertalanffy Lv. *General System Theory, Foundations, Development, Applications*, George Braziller, New York, 1969.

Bird DM *et al.* The *Caenorhabditis elegans* genome: A Guide in the post genomics age. *Annu. Rev. Phytopathol.*, 37: 247-265 (1999).

Bray D. Protein molecules as computational elements in living cells. *Nature*, 376: 307-312 (1995).

Fell DA. Beyond genomics. *Trends Genet.*, 17: 680-682 (2001).

Fiehn O. Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp. Funct. Genom.*, 2: 155-168 (2001).

Goffeau A *et al.* Life with 6000 genes. *Science*, 274: 546-567 (1996).

Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.*, 22: 245-252 (2004).

Harrigan GG, Goodacre R. *Metabolic Profiling: Its role in biomarker discovery and gene function analysis*, Kluwer academic publishers, Boston (2003).

Ideker T, Galitski T, Hood L. A new approach to decoding life: Systems Biology. *Annu. Rev. Genomics Hum. Genet.*, 2: 343-372 (2001).

Kell DB. Metabolomics and systems biology: making sense of the soup. *Curr. Opin. Microbiol.*, 7: 296-307 (2004).

Kitano H. *Foundations of Systems Biology*, MIT Press, Cambridge, MA, 2001.

Levesque MP, Benfey PN. Systems Biology. *Curr. Biol.*, 14: R179 (2004).

McPherson JD *et al.* A physical map of the human genome. *Nature*, 409: 934-941 (2001).

Oliver SG, Winson MK, Kell DB, Baganz F. Systematic functional analysis of the yeast genome. *Trends Biotechnol.*, 16: 373-378 (1998).

Stelling J. Mathematical models in microbial systems biology. *Curr. Opin. Microbiol.*, 7: 513-518 (2004).

ter Kuile BH, Westerhoff HV. Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett.*, 500: 169-171 (2001).

Tweeddale H, Notley-McRobb L, Ferenci T. Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool ("metabolome") analysis. *J. Bacteriol.*, 180: 5109-5116 (1998).

Venter JC *et al.* The sequence of the human genome. *Science*, 291: 1304-1351 (2001).

Weiner N. *Cybernetics or control and communication in the Animal and the Machine*, MIT Press, Cambridge, MA (1948).

Westerhoff HV, Palsson BO. The evolution of molecular biology into systems biology. *Nat. Biotechnol.*, 22: 1249-1252 (2004).

Winkler WC *et al.* Control of gene expression by a natural metabolite-responsive ribozyme. *Nature*, 428: 281-286 (2004).

Chapter 2

TOWARDS INTEGRATIVE FUNCTIONAL GENOMICS USING YEAST AS A REFERENCE MODEL

Metabolomic analysis in the post-genomic era

Juan I. Castrillo and Stephen G. Oliver
School of Biological Sciences. The Michael Smith Building. University of Manchester.Oxford Road. Manchester M13 9PT, UK.

1. INTRODUCTION

Metabolites have been the subject of investigation since the early stages of modern biology. Thus, classical studies on identification of enzymes and metabolic intermediates performed in yeast in the 1920s-1930s (e.g. Embden-Meyerhoff unified theory of glycolysis, citric acid cycle, AMP, ATP) constitute the foundations of modern enzymology and biochemistry (Lehninger, 1975; Alberts *et al.*, 2002). The main interest of these studies focused on the elucidation of the complete map of central metabolic pathways and intermediary metabolites of an organism. This objective, satisfactorily fulfilled for the case of a few organisms (bacteria, yeast), may constitute a major task in more complex organisms (e. g. plants, mammalian cells), with particular metabolites (e.g. secondary metabolites and regulatory compounds) still to be identified. For the case of eukaryotes, yeast central metabolic pathways and methods for determination of metabolites are used as a reference from which to approach more complex biological systems (Gancedo and Gancedo, 1973; Saez and Lagunas, 1976; Rose and Harrison, 1987-1995; Fell, 1997; Alberts *et al.*, 2002).

The current 'genomic revolution' is generating large amounts of valuable information, primarily in the form of new genome sequences and genome-wide expression data (microarray-transcriptome data), with significant

advances on proteome studies as well (Castrillo and Oliver, 2004 and references therein). However metabolomics, the comprehensive analysis of the complete pool of cellular metabolites (the 'metabolome') closely interacting with the other genomic levels, and directly reflecting the cell's phenotype, is sometimes inadvertently overlooked in post-genomic studies (Adams, 2003; Harrigan and Goodacre, 2003; Goodacre *et al.*, 2004).

In the new post-genomic era studies will progressively have to evolve from the punctual, isolated discovery of biological information to the integration of present and new data in a structured manner, towards the comprehension of the cell as a global entity in which different genomic levels (genome, transcriptome, proteome, metabolome, Oliver *et al.*, 1998; Castrillo and Oliver, 2004) exert their respective functions not independently but interacting coordinately with the others, through specific regulatory mechanisms, direct response to the environmental conditions, in an integrative, 'Systems Biology' perspective (Kitano, 2002; Kafatos and Eisner, 2004).

The purpose of this chapter is to present a comprehensive view of metabolomics as an essential, intrinsic component of integrative studies in the post-genomic era. In the first section of the chapter basic metabolic profiling techniques and applications will be presented. In the second part, relevance of metabolites and metabolic regulation will be reported, along with new mechanisms involving participation of metabolites in global expression and regulatory control. Finally in the last section attention is focused on the favourable characteristics of yeast as a reference model organism for integrative genomic approaches, including metabolomics, for application in Systems Biology studies.

2. METABOLIC PROFILING. EXPERIMENTAL STRATEGIES AND APPLICATIONS

2.1 Methods of analysis of metabolites: Requirements.

The metabolic state of a cell is defined by the identity and concentrations of both intracellular and extracellular metabolites present or acting upon the cell. These will vary in a tightly regulated way in response to the environmental or developmental changes. In order to establish a reliable picture of a cell's metabolic state, covering a wide range of metabolites, comprehensive and efficient methods are required. This is intrinsically difficult due to the heterogeneity of different families of metabolites, their high reactivity (i.e. the turnover rates of intermediary metabolites range from

several seconds to milliseconds; Fell, 1997), and the different ranges of concentrations over which they exert their physiological effects (Table 1 and references therein).

Table 1. Ranges of internal and external metabolite concentrations. Physiological ranges of selected groups of yeast and fungal metabolites (Gancedo and Gancedo, 1973; Atkinson and Mavituna, 1991; Martinez-Force and Benitez, 1991; de Koning and van Dam, 1992).

Metabolites	Range	
	(aerobic)	(anaerobic)
Internal intermediary metabolites		
Glycolytic intermediates (aerobic – anaerobic)	mM	μM
Amino acids	mM	
Nucleotides (AMP, ADP, ATP)	mM	
Vitamins	[μM – mM]	
External metabolites/compounds		
Substrates/nutrients (C, N, P, S sources, mineral salts trace elements, vitamins)	[μM – mM]	
Products (e.g. ethanol, acetate, organic acids)	[μM – mM]	
Secondary metabolites (amino acids, peptides, other signalling molecules, e.g. heterocyclic compounds)	[nM – μM]	

In vivo studies can be applied in limited cases (e.g. fluorescence spectrophotometry, dual beam spectrophotometry or NMR; Fell, 1997), but in the majority of cases it will be necessary to work with extracts and, if the measurements are to truly represent the situation within the living cell, a number of requirements have to be fulfilled. These requirements have been established through the work of several researchers (e.g. Saez and Lagunas, 1976; De Koning and van Dam, 1992; Fell, 1997; Hajjaj *et al.*, 1998; Castrillo *et al.*, 2003) and they can be summarized as:

1) *Fast sampling.* Due to the low turnover rates of metabolites fast sampling (including extracellular medium and cells) coupled to methods to stop further reactions and fix the concentration of metabolites (quenching) is mandatory (Theobald *et al.*, 1993; Fell, 1997; Lange *et al.*, 2001).

2) *Quenching of metabolites.* A number of different methods are used, including rapid drop to low temperatures (-40 °C or lower), sudden pH change or mixing with organic solvents (Fell, 1997; Hajjaj *et al.*, 1998; Castrillo *et al.*,2003; Mashego *et al.*, 2003; Villas-Boas *et al.*, 2003).

3) *Efficient extraction of internal metabolites.* Due to their heterogeneity, there is no universal method that allows the extraction of all metabolites with maximum efficiency. Extraction is usually performed at neutral pH in mixtures of organic compounds (e.g. chloroform) or in boiling ethanol, in order to obtain a representative sample of the variety of chemically

compatible metabolites (e.g. soluble metabolites) present in the cell (Gonzalez *et al.,* 1997; Villas-Boas *et al.,* 2003).

4) *Concentration step.* The quenching and extraction steps result inevitably in the dilution of the metabolites, whose concentration can fall below the sensitivity limit of subsequent analytical techniques. A concentration step is, therefore, necessary. This is usually performed by evaporation of the solvent. After that, the extracts can be stored for short periods at -80 °C but, since different types of metabolites can exhibit different stabilities, immediate analysis is strongly recommended (Castrillo *et al.,* 2003).

5) *Preparation of the sample and analyte determination.* Due to the different ranges of concentrations of metabolites (Table 1) and the dilution and concentration steps inherent to the extraction method, the preparation of the sample from the concentrated extract has to be carefully designed to allow determination of the largest group of metabolites within the dynamic range and sensitivity of the analytical technique to be used. Among the most extensively used are: enzymatic and immunoassays methods (Fell, 1997; Gonzalez *et al.,* 1997), NMR (Brindle *et al.,* 1997; Griffin, 2004), and mass spectrometry methods (e. g. electrospray ionization mass spectrometry, ES-MS; Vaidyanathan *et al.,* 2001; Allen *et al.,* 2003). These can be used with high versatility, either individually (e.g. direct infusion electrospray mass spectrometry; Castrillo *et al.,* 2003) or combined with selected chromatographic techniques (e.g. GC-MS, GC-Q-ToF-MS; Villas-Boas *et al.,* 2003), coupled to tandem mass spectrometry (MS/MS) or even combined with the use of substrate labelling with stable isotopes (e.g. isotopomer ratio analysis of labelled extracts using LC-ES-MS/MS; Mashego *et al.,* 2004). More recently, a significant improvement in the level sensitivity has been obtained by the development of a new mass spectrometry technique, Fourier Transform Ion Cyclotron Mass Spectrometry (FT-ICR) which opens the possibilities to new advanced metabolome studies (Aharoni *et al.,* 2002).

The requirements listed above allow the extraction and analysis of a number of cell metabolites in order to obtain a global picture of the metabolic state of the cell (by high-throughput analysis of global external and internal metabolic profiles). However, eukaryotic cells, like yeast, contain a number of compartments and the internal metabolites are not uniformly distributed among them. For advanced studies, including quantification of metabolites in specific cellular compartments or free and bound metabolites, specific assumptions of relative volumes of water in these different compartments, in addition to well-designed strategies for organelle isolation and analysis regimes are required (Fell, 1997, Farre *et al.,* 2001).

2.2 Metabolic profiling of internal and external metabolites: Applications.

The concentrations and variations in the levels of metabolites reflect the metabolic state of the cell, and the metabolome is considered the closest level of analysis to the cell's phenotype (Oliver, 1997; Trethewey *et al.*, 1999; Raamsdonk *et al.*, 2001). Hence, metabolic profiling is applied to evaluate variations in metabolic states, competing favourably with, or being complementary to, other 'omic techniques (Adams, 2003; Harrigan and Goodacre, 2003). Metabolic profiling of internal metabolites (metabolic fingerprinting) is currently being used in a wide variety of organisms (yeast, plants, mammalian cells) for different applications (Trethewey *et al.*, 1999; Fiehn *et al.*, 2000; Raamsdonk *et al.*, 2001; Watkins and German, 2002). Metabolic profiling of external metabolites (metabolic footprinting) is being increasingly used (Allen *et al.*, 2003; Kell and Mendes, 2000), and more discoveries are sustaining their physiological relevance, not only in microorganisms (Petroski and McCormick, 1992; Demain, 1998) but also in human cell biology (Hebert, 2004). In functional genomics studies, new methods for metabolic profiling in different organisms (Fiehn *et al.*, 2000; Watkins and German, 2002; Adams, 2003) are used for the elucidation of the function of new genes and metabolic pathways (Teusink *et al.*, 1998; Raamsdonk *et al.*, 2001; Trethewey, 2001; de la Fuente *et al.*, 2002; Weckwerth and Fiehn, 2002). For applied purposes metabolic profiling is used in the investigation of molecules for nutritional assessments (e.g. studies on the interaction of diet and health, or for the assessment of GM foods), evaluation of health and disease states (biomarkers, e g. in cancer cells) for application in diagnostics, as indicators of disease progression and for the screening of new drugs (Griffin *et al.*, 2001; Schilter and Constable, 2002; Watkins and German, 2002; Fiehn and Spranger, 2003; Griffin and Shockcor, 2004; Lee and Boros, 2003; Heaton *et al.*, 1999,; Kaddurah-Daouk and Kristal, 2001; Stockton *et al.*, 2002).

3. METABOLOMIC STUDIES IN FUNCTIONAL GENOMICS

3.1 Role of metabolism and metabolites in Functional Genomics: Regulation.

Primary metabolism can be defined as the coordinated biochemical conversion of substrates through tightly regulated metabolic pathways in

order to generate energy and building blocks for growth and the maintenance of cellular functions. It is usually divided into catabolism and anabolism with participation of common amphibolic reactions (Lehninger, 1975; Castrillo and Oliver, 2004). Based on this definition only, the role of metabolism and metabolites in Functional Genomics could be underestimated, and be considered of secondary importance to the flow of genetic information and the regulation of gene expression. In the flow of information from gene (DNA) to RNA to proteins (e.g. enzymes, which catalyse the specific metabolic reactions) metabolites could be regarded as inert molecules with negligible participation in regulation. However, a comprehensive revision on participation of metabolites in regulation and control offers a more complete perspective of the importance of metabolomics in Functional Genomics, as can be seen from the following observations:

1) *Central metabolic pathways.* Internal metabolites exert rapid short-term regulation of metabolic fluxes by modulation of enzymatic activity. The changes in fluxes along the major metabolic pathways have long been reported to be tightly regulated by the concentration of specific internal metabolites (e.g. fructose-1,6-diphosphate, ATP, ADP, citrate) through rapid activation and inhibition of key enzymes by reversible covalent modification as well as by allosteric effects (metabolic effectors; see e.g. Monod *et al.*, 1963; Fell, 1997; Muller *et al.*, 2003; Plaxton, 2004). These key metabolites (e.g. sugar-phosphates, adenylates, cAMP), which collectively regulate carbohydrate metabolism, have no direct involvement in carbon regulation of gene expression. In these cases, assimilation of carbon nutrients is regulated by specific sensing and signal transduction pathways involving other specific protagonists.

2) *External signals - metabolite sensors.* A cell has to maintain the stability of the intracellular environment (homeostasis) in response to changes in the external conditions. The nature and variations of levels of external metabolites (i.e. substrates, sometimes called catabolites; products; other external compounds) constitute the primary level of environmental information (signals) detected by the cell through its specific sensing mechanisms (usually by means of metabolite-protein interactions, ligand-receptor at the membrane level; Hancock, 1997).

3) *Signal transduction pathways - internal metabolites.* Once an external signal (presence, absence or change in metabolite concentrations) is detected, intracellular signal transduction pathways are triggered (Hancock, 1997; Sprague *et al.*, 2004). In the widely accepted model of mechanism, the metabolite binds to a specific protein which can modify other regulatory proteins post-transcriptionally, resulting in changes in the levels and/or mechanisms of action of other regulatory proteins (e.g. transcription factors)

leading to tightly regulated changes in gene expression (i.e. groups of genes are selectively up-regulated whereas others are markedly down-regulated). In addition to this model, more evidence is progressively appearing which supports a relevant role of internal metabolites (e.g. phosphate, cAMP, inositol phosphate), in signal transduction pathways, participating closely with protein cascades and regulatory proteins (e.g. transcription factors; Hancock, 1997; Gancedo, 1998; Hansen and Johannesen, 2000; Auesukaree *et al.,* 2004; Sprague *et al.,* 2004).

The nutrient assimilation pathways (e.g. carbon, nitrogen, phosphate and sulphur assimilation pathways) constitute reference examples of regulation via signal transduction pathways. These routes are of central importance for efficient assimilation of substrates while keeping internal homeostasis. External concentrations of these metabolites are carefully monitored and their assimilation is tightly regulated at the level of gene expression. A remarkable aspect is that each class of metabolites (carbohydrates, nitrogen compounds, amino acids, lipids) has its own signal transduction mechanisms and they modulate a different set of cellular genes (although the signal transduction pathways may share specific components; Sprague *et al.,* 2004). Even for a given metabolite (e.g. glucose), the signal transduction pathway that detects a high concentration can be different than the one that detects a limiting concentration. The signal transduction pathways and their underlying mechanisms are the subjects of intensive investigations that are specific for each substrate. Relevant examples are, studies on carbon catabolite repression (Gancedo, 1998; Zaragoza *et al.,* 1999); nitrogen catabolite repression (Fafournoux *et al.,* 2000); phosphate (Pi) assimilation (Auesukaree *et al.,* 2004), as well as sulphur assimilation and the role of intracellular sulphur compounds in transcriptional regulation (Hansen and Johannesen, 2000; Sellick and Reece, 2003).

4) *Role of excreted metabolites* Secondary metabolites are produced by specific routes that are different from those of the central metabolic pathways, mostly operating after the phase of active growth and under conditions of nutrient deficiency. These excreted metabolites can perform functions in cell signalling, or as external inducers or autoinducers. They can govern the behaviour and differentiation of the cells in a colony (morphological differentiation, sporulation; Petroski and McCormick, 1992; Horinouchi and Beppu, 1995; Demain, 1998; Roncal and Ugalde, 2003). They usually act via receptor proteins, which repress chemical and morphological differentiation into aerial mycelia or spores. They normally act at very low concentrations (nM, μM) (Table 1) once a critical concentration (threshold) is reached.

All these studies confirm the relevance of the metabolites together with DNA, RNA and proteins in the global biological response of the cell, and the

importance of not overlooking the metabolome in Functional and Systems Biology studies (see next sections). Moreover, new mechanisms by which metabolites can control gene expression (e.g. by direct interaction with mRNA-riboswitches, without participation of proteins), or that can lead to post-translational histone modifications have been reported (Cech, 2004; Dong and Xu, 2004). These and other novel mechanisms constitute new challenges to be incorporated to the global picture of Functional Genomics.

3.2 Metabolomic studies in Functional Genomics: State of the art and new challenges.

A global perspective of the different levels of functional genomic analysis (genome, transcriptome, proteome and metabolome; Oliver, 1997) including the flow of genetic information (from DNA to RNA to proteins, with their interrelations with metabolites) and the main regulatory relationships between them and the environment is presented in Figure 1. The role exerted by the metabolome through their interaction with the other biological entities is presented, including the most recently discovered mechanisms referred in this chapter. For a good review on new mechanisms and nature of gene regulation see Choudhuri (2004).

From this picture, an essential characteristic of Functional Genomics emerges, which is the coordinated integration of different levels and individual networks in the cell, in direct communication with the environment, in a system that is intrinsically rich in complexity.

The first stages of functional genomics studies have been primarily characterized by the generation and optimisation of genome-wide strategies for the global study of the different genomic levels (usually genome, transcriptome and proteome only), in different organisms (e.g. yeast, plants, -see Fiehn *et al.*, 2000; Kell and King, 2000; Raamsdonk *et al.*, 2001; Adams, 2003; Griffin, 2004). Some combined studies that include different individual genomic approaches have been performed. In many cases, these studies have been directed towards the identification of overlooked genes or genes associated to specific protein activities (Kumar *et al.*, 2002; Chen *et al.*, 2003), while others have focussed on the elucidation of direct correlations between two different 'omic levels (ter Kuile and Westerhoff, 2001; Yoon and Lee, 2002; Urbanczyk-Wochniak *et al.*, 2003).

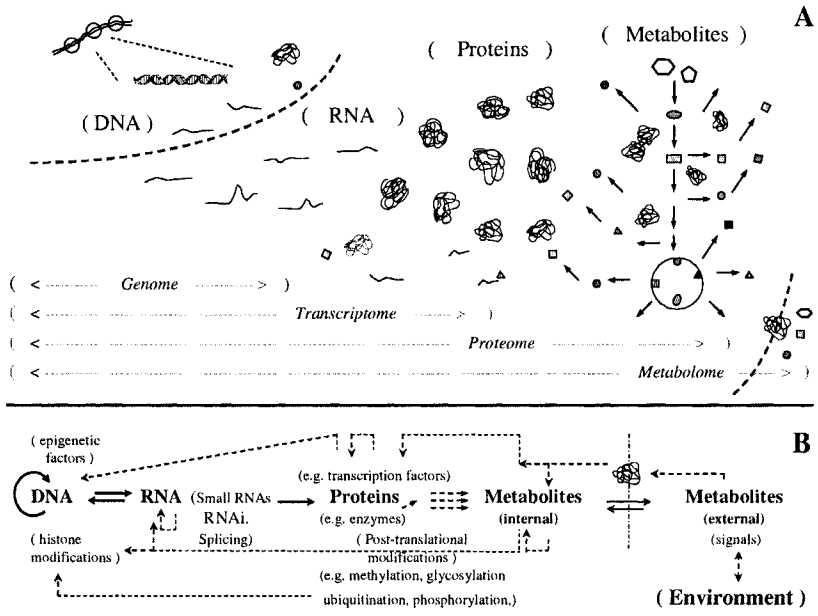


Figure 1. Functional genomics. Levels of study and interrelations at the regulatory level. A) Visual representation of levels of genomic information in the cell. B) Regulatory relationships between genomic levels: Flow of genetic information, from DNA to RNA and proteins and their relationships with metabolic entities and the environment, including latest discoveries in post-transcriptional and post-translational mechanisms (e.g. RNA interference, riboswitches, histone modifications) (Castrillo and Oliver, 2004; Choudhuri, 2004).

The new studies in the post-genomic era, however, will have to embrace recent discoveries and increased complexity, such as the existence of other functional elements (not only ORFs) in the DNA sequence (promoters, transcriptional regulatory sequences, intergenic regions; e.g. the ENCODE project: ENCyclopedia Of DNA Elements - http://www.genome.gov/10005107), epigenetic mechanisms, post-transcriptional and post-translational modifications (e.g. RNA splicing, RNA interference, histones methylation, and ubiquitination). The metabolome has an essential role in this new complexity of interrelated communication networks between 'omic levels (many of whose circuits are still to be elucidated) as the basis of the global biology of the cell (Fell, 2001; Ideker *et al.*, 2001; Castrillo and Oliver, 2004). Among the most intriguing mechanisms and new challenges for metabolomic studies in the post-genomic era are:

1) Metabolites regulating gene expression via protein-metabolite interactions. Interesting examples are a recently reported study on the modulation of transcription factor function by proline (Sellick and Reece, 2003), or more complex effects such as glucose-mediated phosphorylation converting a transcription factor from a repressor to an activator (Mosley *et al.,* 2003). 2) Metabolites regulating gene expression via binding to RNA, bypassing proteins (riboswitches). The metabolite binds to an RNA molecule (metabolite-RNA interaction) that is not translated (Cech, 2004; Winkler *et al.,* 2004). Although metabolite-binding RNA domains are present in genes of eukaryotes (Sudarsan *et al.,* 2003) the extent of this regulatory mechanism is still to be determined. 3) In a recent breakthrough in the field, the role of intergenic regions (formerly considered non-coding DNA regions) in amino acid assimilation pathways has been demonstrated. Thus, in *Saccharomyces cerevisiae*, intergenic transcription has been reported to be required to repress the synthesis of serine on rich media (Martens *et al.,* 2004; Schmitt and Paro, 2004). 4) In a novel paradigm of metabolic regulation, metabolic pathways and metabolites (glycolysis and glucose) have been recently reported to be associated with histone ubiquitination and gene silencing (Dong and Xu, 2004). 5) Evidence for the participation of external signalling mechanisms in a wide variety of organisms including human. Thus, endogenous metabolites excreted to the bloodstream (TCA cycle intermediates, e.g. succinate) have been found acting as signalling molecules (i.e. ligands) for G-protein-coupled receptors, linking the metabolism and injury of tissues with blood pressure (He *et al.,* 2004; Hebert, 2004).

A significant effort of metabolomic studies in the post-genomic era will have to be dedicated to intensive research, to unveil the mechanisms underlying these processes. Together with this, and of no less importance, metabolomics will need to develop new high-throughput methods and refined strategies for the qualitative and quantitative determination of an increasing number of metabolites and their sub-cellular localization in different cell systems (e.g. cells, tissues, body fluids). The final objective will be to combine this information together with studies from all other genomic levels (genome, transcriptome and proteome) in an integrative Systems Biology approach (Kitano, 2002), in order to understand the global behaviour of the cell. Thus, integration in the form of mathematical models based on, for example strategies of top-down control analysis (Quant, 1993; Krauss and Quant, 1996) and metabolic control analysis (MCA) (Fell, 1997; Peletier *et al.,* 2003) can incorporate the new discoveries from the different levels of analysis. Due to the rediscovered high complexity of biological systems, integrative studies in simple touchstone model organisms (see Castrillo and Oliver, 2004) are necessary in order to derive adequate conclusions.

4. METABOLOMIC ANALYSIS IN NEW INTEGRATIVE FUNCTIONAL GENOMICS: YEAST AS A REFERENCE MODEL

4.1 Integrative studies in functional genomics: Systems biology.

From the perspective of the functional genomic levels and relationships shown in Figure 1 it is clear that the metabolome exerts its role in a global integrated cell system, more complex than that usually considered in individual investigations, with relevant contributions to regulation at the post-transcriptional, post-translational, and metabolic levels (Fafournoux *et al.*, 2000; Hansen and Johannesen, 2000; Muratani and Tansey, 2003; Choudhuri, 2004). This reality is clearly being shown in new post-genomic studies in which the lack of a direct correlation between levels of gene expression (mRNA abundance) and protein content has been demonstrated (Lee *et al.*, 2003; Yoon *et al.*, 2003). This fact, first carefully studied in exponential-phase batch cultures of yeast (Gygi *et al.*, 1999) and in integrated microarray-proteome studies of the yeast galactose assimilation pathway (Fell, 2001; Ideker *et al.*, 2001) has been certified in a variety of organisms and culture conditions (Gygi *et al.*, 1999; ter Kuile and Westerhoff, 2001; Glanemann *et al.*, 2003; Lee *et al.*, 2003; Mehra *et al.*, 2003). This intrinsic complexity has also been proved at the metabolomic level, where there is no simple correlation between transcript or protein levels for relevant enzymes and measured metabolic fluxes (Fell, 2001; Ideker *et al.*, 2001; Yoon and Lee, 2002; Bro *et al.*, 2003; Daran-Lapujade *et al.*, 2004). All these results demonstrate the need for more exhaustive and comprehensive integrative studies in the post-genomic era (Delneri *et al.*, 2001; Oliver *et al.*, 2002; Phelps *et al.*, 2002; Urbanczyk-Wochniak *et al.*, 2003; Castrillo and Oliver, 2004; Weckwerth and Fiehn, 2003).

Systems Biology focuses on the importance of a global integrative view of biological processes, including new holistic approaches to elucidate cell complexity by combining global analysis of data sets obtained from systematic genome, transcriptome, proteome and metabolome studies. The objective is to construct mathematical models of complex biological systems by which to interrogate and iteratively refine our knowledge of the cell (Kitano, 2002; Ideker, 2004).

As stated previously, most relevant efforts have focused on strategies combining two functional genomic levels or strategies and, usually have directed to the discovery of the function of unknown genes (e.g. Kumar *et al.*, 2002; Chen *et al.*, 2003). Together with this, the new frontier in the post-

genomic era will focus on new integrative methods and strategies for elucidating complex regulatory networks at each specific level of analysis (genome, transcriptome, proteome and metabolome), and the exploration of the intricate interrelationships between them. For these purposes, new tools and methods to link information from different parallel analyses, algorithms, and advanced tools for *in silico* analysis of specific patterns are being developed (Kell and King, 2000; Fiehn, 2001; de la Fuente *et al.,* 2002; Mendes, 2002; Yao, 2002; Cornell *et al.,* 2003; Fiehn and Weckwerth, 2003; Weckwerth, 2003). These studies on gene expression, proteome and metabolic networks can provide crucial information, but are critically dependent on the accuracy and reliability of the experiments and the raw data generated from them. Thus, proper rigor in comprehensive integrative studies and the use of simple touchstone model organisms under well-defined conditions are essential to the early stages of systems biology (Castrillo and Oliver, 2004).

4.2 Metabolomics in new integrative studies: Yeast as a reference model.

Saccharomyces cerevisiae exhibits a number of favourable characteristics that recommend it as a reference model organism in post-genomic studies, particularly in integrative studies that include metabolomics. Thus: 1) Many cellular mechanisms and metabolic pathways were first elucidated in yeast, and a wide knowledge of the genetics, biochemistry and physiology of yeast is currently available (Lehninger, 1975; Rose and Harrison, 1987-1995; Brown and Tuite, 1998; Burke *et al.,* 2000; Sambrook and Russell, 2000). 2) The existence of simple methods of cultivation and a well-characterized genetics with simple techniques of genetic manipulation. 3) *S. cerevisiae* was the first eukaryotic organism for which the whole genome sequence was completed (Goffeau *et al.,* 1996). This fact, combined with the existence of a comprehensive collection of gene deletion mutants (Giaever *et al.,* 2002; http://www.uni-frankfurt.de/fb15/mikro/euroscarf/complete.html), and high-throughput technologies for global analyses at a genome-wide scale provides a wide range of possibilities for integrative strategies. Yeast is regularly used as a reference model system for the study of eukaryotic cell biology and regulatory mechanisms (Castrillo and Oliver, 2004 and references therein). All these favourable characteristics make it a perfect touchstone model and an optimum platform for integrative studies in the post-genomics era (Oliver, 1997; Oliver *et al.,* 1998; Delneri *et al.,* 2001; Castrillo and Oliver, 2004).

In an example of combining genomic and metabolomic strategies, comprehensive analyses of metabolite profiles from yeast deletion mutants

can be applied to ascribe function to unknown genes. This has been successfully demonstrated, particularly for the case of 'silent' genes (genes whose mutation causes no obvious phenotype) in an approach called functional analysis by co-responses in yeast (FANCY). Based on the fact that mutations involved in same functional responses can lead to similar changes in intracellular metabolite concentrations, matching the metabolic profiles of genes of unknown function with those associated with specific mutations can reveal the function of unknown genes (Raamsdonk *et al.,* 2001). Also, for the case of mutations resulting in characteristic external metabolic signatures, a complementary approach using comparative metabolomics of extracellular profiles has shown the validity of external metabolic footprinting as a high-throughput method for classification of yeast mutants (Allen *et al.,* 2003).

Integrative studies using yeast have demonstrated the lack of a simple direct correlation between transcript or protein levels and metabolic fluxes (Fell, 2001; Ideker *et al.,* 2001; Bro *et al.,* 2003). Hence, more extensive studies are required to unveil the relevant role of metabolites in regulation and to generate the information needed for a global systems biology approach. In these studies again, yeast appears as the preferred model organism. Relevant examples are the investigations on glucose sensing and signalling mechanisms through the Rgt2 sensor (Moriya and Johnston, 2004) and studies on the **tor** signal transduction pathway, linking nutrient sensing with histone acetylation to control the expression of ribosomal protein genes and, thereby, cell growth (Rohde and Cardenas, 2003).

The new knowledge generated in basic studies and the large sets of data generated at the different functional levels have to be processed efficiently. Appropriate bioinformatic tools which integrate metabolome information with data coming from other genomic levels are of central importance. In this respect, effort is being directed at the development of new clustering and machine learning methods appropriate for the analysis of transcriptome, proteome and metabolome data and the study of their interrelationships in complex regulatory networks (Kell and King, 2000; Fiehn, 2001; Kell *et al.,* 2001; ter Kuile and Westerhoff, 2001; de la Fuente *et al.,* 2002; Mendes, 2002; Fiehn and Weckwerth, 2003; Goodacre *et al.,* 2004).

The final objective of obtaining information in systems biology studies is to incorporate these data into mathematical models, descriptive of the cell system. Depending on the specific purposes, these can be simple unstructured models at first, including minimum information of internal genomic levels (e.g. central metabolic pathways only; metabolic steady-state flux models based on top-down control theory or metabolic control analysis; Bailey and Ollis, 1986; Fell, 1997; Segre *et al.,* 2003), whose complexity can be progressively increased. In this respect, yeast models have long been

developed for use in basic and applied studies, which can serve as a reference for the implementation of new models of higher complexity (Bailey and Ollis, 1986 and references therein; Castrillo and Ugalde, 1994 and references therein; Cortassa and Aon, 1994). In these models, one of the main goals is usually the identification of key targets (e.g. enzymatic steps, proteins) whose manipulation via genetic modification or drug treatment would result in a significant change in the flux through the entire pathway (in metabolic control analysis theory, those ones exhibiting a high flux control coefficient; Fell, 1998). At present, many efforts on drug discovery are focusing on targeting specific signalling pathways and protein kinases (Cascante *et al.*, 2002; Gough *et al.*, 2004; Noble *et al.*, 2004) but it remains a difficult task. The latest studies, unveiling the new complexity of the cell referred in this chapter only serve to illustrate the new difficulties and challenges that lie ahead. With the different genomic levels acting coordinately in response to the environment, the objective will be to understand the hierarchical organization of regulatory and metabolic networks within the cell (ter Kuile and Westerhoff, 2001; Ihmels *et al.*, 2004) and their interrelationships, to identify the main processes responsible for the cellular response under specific environmental conditions (Fiehn, 2001; Wu *et al.*, 2002; Fiehn and Weckwerth, 2003; Sandelin *et al.*, 2003). These studies can provide crucial information for the development of new drugs and therapeutic strategies, and for direct application in metabolic engineering towards the synthesis of high value products (e.g. heterologous proteins and/or metabolites; Liao, 2001). This crucial information will only be unveiled by means of integrative studies using touchstone models and in this respect, *S. cerevisiae* is in a privileged position as the optimal starting point for post-genomic studies aimed at a systems approach.

5. CONCLUSIONS AND FUTURE PERSPECTIVES

The new complexity that has arisen from post-genomic investigations constitutes a major challenge. In order to approach this reality, comprehensive integrative studies under well-defined controlled conditions are necessary. These will be required, firstly for the elucidation of the still-unknown regulatory mechanisms at the genomic, transcriptional, post-transcriptional and post-translational levels that participate in the response of the cell to specific environmental conditions (e.g. signal transduction pathways, regulatory networks). Secondly, it will be necessary to incorporate this information into progressively more realistic models, for use in Systems Biology research, from which direct applications (e. g. drug discovery and metabolic engineering) can be derived. The relevant role of metabolites as

sensing molecules as well as participants in global intracellular regulatory mechanisms presented in this chapter, illustrates the importance of including metabolomics, together with transcriptome and proteome studies, in future post-genomic studies. These integrative studies can be performed first in simple model organisms under controlled conditions. This knowledge can be related to information from other organisms, towards a better understanding of the cell biology of more complex systems. In this respect, the optimal characteristics of yeast makes it a perfect reference model to provide new knowledge and insights in cell biology, and a relevant touchstone at the forefront of studies in the post-genomic era.

ACKNOWLEDGEMENTS

This work was supported by an EC contract to SGO within the frame of the Garnish Network of FP5 and the BBSRC's Investigating Gene Function Initiative within COGEME (Consortium for the Functional Genomics of Microbial Eukaryotes; http://www.cogeme.man.ac.uk).

REFERENCES

Adams A. Metabolomics: Small-molecule 'omics. *The Scientist,* 17: 38-40 (2003).

Aharoni A. Ric de Vos CH, Verhoeven HA, Maliepaard CA, Kruppa G, Bino R and Goodenowe DB. Nontargeted metabolome analysis by use of Fourier Transform Ion Cyclotron Mass Spectrometry. *OMICS,* 6: 217-234 (2002).

Alberts B, Johnson A, Lewis J, Raff M, Roberts K and Walter P. *Molecular Biology of The Cell,* 4th ed., Garland Science, Taylor and Francis Group, New York (2002).

Allen J, Davey HM, Broadhurst D, Heald JK, Rowland JJ, Oliver SG and Kell DB. High-throughput classification of yeast mutants using metabolic footprinting. *Nat.Biotechnol.,* 21: 692-696 (2003).

Atkinson B and Mavituna F. *Biochemical Engineering and Biotechnology Handbook,* 2nd ed., M. Stockton Press, New York (1991).

Auesukaree C, Homma T, Tochio H, Shirakawa M, Kaneko Y and Harashima S. Intracellular phosphate serves as a signal for the regulation of the PHO pathway in *Saccharomyces cerevisiae. J. Biol. Chem.,* 279: 17289-17294 (2004).

Bailey JE and Ollis DF. *Biochemical Engineering Fundamentals,* 2nd ed., McGraw Hill, New York (1986).

Brindle KM, Fulton SM, Gillham H and Williams SP. Studies of metabolic control using NMR and molecular genetics. *J. Mol. Recognit.,* 10: 182-187 (1997).

Bro C, Regenberg B, Lagniel G, Labarre J, Montero-Lomeli M and Nielsen J. Transcriptional, proteomic, and metabolic responses to lithium in galactose-grown yeast cells. *J. Biol. Chem.,* 278: 32141-323149 (2003).

Brown AJP and Tuite MF. Yeast Gene Analysis. *Methods in Microbiol.,* 26. Academic Press. San Diego (1998).

Burke D, Dawson D and Stearns T. *Methods in Yeast Genetics*, 2000 Edition: A Cold Spring Harbor Laboratory Course Manual. Cold Spring Harbor Laboratory Press. New York (2000).

Cascante M, Boros LG, Comin-Anduix B, de Atauri P, Centelles JJ and Lee PW. Metabolic control analysis in drug discovery and disease. *Nat. Biotechnol.*, 20: 243-249 (2002).

Castrillo JI and Oliver SG. Yeast as a touchstone in post-genomic research. Strategies for integrative analysis in functional genomics. *J. Biochem. Mol. Biol.*, 37: 93-106 (2004).

Castrillo JI and Ugalde UO. A general model of yeast energy metabolism in aerobic chemostat culture. *Yeast*, 10: 185-197 (1994).

Castrillo JI, Hayes A, Mohammed S, Gaskell SJ and Oliver SG. An optimised protocol for metabolome analysis in yeast using direct infusion electrospray mass spectrometry. *Phytochemistry*, 62: 929-937 (2003).

Cech TR. RNA finds a simpler way. *Nature*, 428: 263-264 (2004).

Chen CN, Porubleva L, Shearer G, Svrakic M, Holden LG, Dover JL, Johnston M, Chitnis PR and Kohl DH. Associating protein activities with their genes: rapid identification of a gene encoding a methylglyoxal reductase in the yeast *Saccharomyces cerevisiae*. *Yeast*, 20: 545-554 (2003).

Choudhuri S. The nature of gene regulation. *Int. Arch. Biosci.*, 1001-1015 (2004).

Cornell M, Paton NW, Hedeler C, Kirby P, Delneri D, Hayes A and Oliver SG. GIMS: An integrated data storage and analysis environment for genomic and functional data. *Yeast*, 20, 1291-1306 (2003).

Cortassa S and Aon MA. Metabolic control analysis of glycolysis and branching to ethanol production in chemostat cultures of *Saccharomyces cerevisiae* under carbon, nitrogen, or phosphate limitations. *Enzyme Microb. Technol.*, 16: 761-770 (1994).

Daran-Lapujade P, Jansen ML, Daran JM, van Gulik W, de Winde JH and Pronk JT. Role of transcriptional regulation in controlling fluxes in central carbon metabolism of *Saccharomyces cerevisiae*. A chemostat culture study. *J. Biol. Chem.*, 279: 9125-9138 (2004).

De Koning W and van Dam K. A method for the determination of changes in glycolytic metabolites in yeast on a subsecond time scale using extraction at neutral pH. *Anal. Biochem.*, 204: 118-123 (1992).

De la Fuente A, Snoep JL, Westerhoff HV and Mendes P. Metabolic control in integrated biochemical systems. *Eur. J. Biochem.*, 269: 4399-4408 (2002).

Delneri D, Brancia FL and Oliver SG. Towards a truly integrative biology through the functional genomics of yeast. *Curr. Opin. Biotechnol.*, 12: 87-91 (2001).

Demain AL. Induction of microbial secondary metabolism. *Int. Microbiol.*, 1: 259-264 (1998).

Dong L and Xu CW. Carbohydrates induce mono-ubiquitination of H2B in yeast. *J. Biol. Chem.*, 279: 1577-1580 (2004).

Fafournoux P, Bruhat A and Jousse C. Amino acid regulation of gene expression. *Biochem. J.*, 351: 1-12 (2000).

Farre EM, Tiessen A, Roessner U, Geigenberger P, Trethewey RN and Willmitzer L. Analysis of the compartmentation of glycolytic intermediates, nucleotides, sugars, organic acids, amino acids, and sugar alcohols in potato tubers using a nonaqueous fractionation method. *Plant Physiol.*, 127: 685-700 (2001).

Fell DA. *Understanding the Control of Metabolism*, Portland Press Ltd., London (1997).

Fell DA. Increasing the flux in metabolic pathways: A metabolic control analysis perspective. *Biotechnol. Bioeng.*, 58: 121-124 (1998).

Fell DA. Beyond genomics. *Trends Genet.*, 17: 680-682 (2001).

Fiehn O. Combining genomics, metabolome analysis and biochemical modelling to understand metabolic networks. *Comp. Funct. Genomics*, 2: 155-168 (2001).

Fiehn O and Spranger J. Use of metabolomics to discover metabolic patterns associated with human diseases; in: *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*, G. G. Harrigan, and R. Goodacre, eds., Kluwer Academic Publishers, Boston, pp. 199-216 (2003).

Fiehn O and Weckwerth W. Deciphering metabolic networks. *Eur. J. Biochem.*, 270: 579-588 (2003).

Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN and Willmitzer L. Metabolite profiling for plant functional genomics. *Nat. Biotechnol.*, 18: 1157-1161 (2000).

Gancedo JM. Yeast carbon catabolite repression. *Microbiol. Mol. Biol. Rev.*, 62: 334-361 (1998).

Gancedo JM and Gancedo C. Concentrations of intermediary metabolites in yeast. *Biochimie*, 55: 205-211 (1973).

Giaever G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, 418, 387-391 (2002).

Glanemann C, Loos A, Gorret N, Willis LB, O'Brien XM, Lessard PA and Sinskey AJ. Disparity between changes in mRNA abundance and enzyme activity in *Corynebacterium glutamicum* and implications for DNA microarray analysis. *Appl. Microbiol. Biotechnol.*, 61: 61-68 (2003).

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H and Oliver SG. Life with 6000 genes. *Science*, 274: 546-567 (1996).

Gonzalez B, François J and Renaud M. A rapid and reliable method for metabolite extraction in yeast using boiling buffered ethanol. *Yeast*, 13: 1347-1355 (1997).

Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG and Kell DB. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.*, 22: 245-252 (2004).

Gough NR, Adler EM and Ray LB. Focus Issue: Targeting signalling pathways for drug discovery. *Sci STKE 225*: eg5, March (2004).

Griffin JL. Metabolic profiles to define the genome: can we hear the phenotypes? *Phil. Trans. Biol. Sciences. R. Soc. Lond. B.*, 359: 857-571 (2004).

Griffin JL and Shockcor JP. Metabolic profiles of cancer cells. *Nat. Rev. Cancer*, 4: 551-561 (2004).

Griffin JL, Williams HJ, Sang E, Clarke K, Rae C and Nicholson JK. Metabolic profiling of genetic disorders: a multitissue ^{1}H nuclear magnetic resonance spectroscopic and pattern recognition study into dystrophic tissue. *Anal. Biochem.*, 293: 16-21 (2001).

Gygi SP, Rochon Y, Franza BR and Aebersold R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.*, 19: 1720-1730 (1999).

Hajjaj H, Blanc PJ, Goma J and François J. Sampling techniques and comparative extraction procedures for quantitative determination of intra- and extracellular metabolites in filamentous fungi. *FEMS Microbiol. Lett.*, 164; 195-200 (1998).

Hancock JT. *Cell signalling*, Prentice Hall, Harlow (1997).

Hansen J and Johannesen PF. Cysteine is essential for transcriptional regulation of the sulfur assimilation genes in *Saccharomyces cerevisiae*. *Mol. Gen. Genet.*, 263; 535-542 (2000).

Harrigan GG and Goodacre R. *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*, Kluwer Academic Publishers, Boston (2003).

He W, Miao FJ, Lin DC, Schwandner RT, Wang Z, Gao J, Chen JL, Tian H and Ling L. Citric acid cycle intermediates as ligands for orphan G-protein-coupled receptors. *Nature*, 429: 188-193 (2004).

Heaton JPW, Brien SE, Adams MA and Graham CH. Method for diagnosing a vascular condition. World Intellectual Property Organisation, *WO Patent*, 9957306 (1999).

Hebert SC. Physiology: orphan detectors of metabolism. *Nature*, 429: 143-145 (2004).

Horinouchi S and Beppu T. Autoregulators. *Biotechnol.*, 28, 103-119 (1995).

Ideker T. Systems biology 101- what you need to know. *Nat. Biotechnol.*, 22: 473-475 (2004).

Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R and Hood L. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292: 929-934 (2001).

Ihmels J, Levy R and Barkai N. Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*. *Nat. Biotechnol.*, 22: 86-92 (2004).

Kaddurah-Daouk R and Kristal BS. Methods for drug discovery, disease treatment and diagnosis using metabolomics. World Intellectual Property Organisation, *WO Patent*, 0178652 (2001).

Kafatos FC and Eisner T. Unification in the century of biology. *Science*, 303: 1257 (2004).

Kell DB and King RD. On the optimization of classes for the assignment of unidentified reading frames in functional genomics programmes: the need for machine learning. *Trends Biotechnol.*, 18: 93-98 (2000).

Kell DB and Mendes P. Snapshots of systems: metabolic control analysis and biotechnology in the post-genomic era. In: *Technological and Medical Implications of Metabolic Control Analysis*, A. Cornish-Bowden, and M. L. Cardenas, eds., Kluwer Academic Publishers, Dordrecht, pp. 3-25 (2000).

Kell DB, Darby RM and Draper J. Genomic computing: explanatory analysis of plant expression profiling data using machine learning. *Plant Physiol.*, 126: 943-951 (2001).

Kitano H. Systems biology: a brief overview. *Science*, 295: 1662-1664 (2002).

Krauss S and Quant PA. Regulation and control in complex, dynamic metabolic systems: experimental application of the top-down approaches of metabolic control analysis to fatty acid oxidation and ketogenesis. *J. Theor. Biol.*, 182: 381-388 (1996).

Kumar A, Harrison PM, Cheung K-H, Lan N, Echols N, Bertone P, Miller P, Gerstein MB and Snyder M. An integrated approach for finding overlooked genes in yeast. *Nat. Biotechnol.*, 20: 58-63 (2002).

Lange HC, Eman M, van Zuijlen G, Visser D, van Dam JC, Frank J, Teixeira de Mattos MJ, and Heijnen JJ. Improved rapid sampling for *in vivo* kinetics of intracellular metabolites in *Saccharomyces cerevisiae*. *Biotechnol. Bioeng.*, 75: 406-415 (2001).

Lee W-NP and Boros LG. Stable isotope based dynamic metabolic profiling of living organisms for characterization of metabolic diseases, drug testing and drug development. US Patent Office, *US Patent*, 2003180800 (2003).

Lee PS, Shaw LB, Choe LH, Mehra A, Hatzimanikatis V and Lee KH. Insights into the relation between mRNA and protein expression patterns: II. Experimental observations in *Escherichia coli*. *Biotechnol. Bioeng.*, 84: 834-841 (2003).

Lehninger AL. *Biochemistry*, 2nd ed., Worth Publishers Inc, New York (1975).

Liao JC. Engineering of metabolic control. World Intellectual Property Organisation, *WO Patent*, 0107567 (2001).

Martens JA, Laprade L, and Winston F. Intergenic transcription is required to repress the *Saccharomyces cerevisiae* SER3 gene. *Nature*, 429: 571-574 (2004).

Martinez-Force E and Benitez T. Separation of *o*-phtalaldehyde derivatives of amino acids of the internal pool of yeast by reverse-phase liquid chromatography. *Biotechnol. Tech.*, 5: 209-214 (1991).

Mashego MR, van Gulik WM, Vinke JL and Heijnen JJ. Critical evaluation of sampling techniques for residual glucose determination in carbon-limited chemostat culture of *Saccharomyces cerevisiae*. *Biotechnol. Bioeng.*, 83: 395-399 (2003).

Mashego MR, Wu L, Van Dam JC, Ras C, Vinke JL, Van Winden WA, Van Gulik WM and Heijnen JJ. MIRACLE: mass isotopomer ratio analysis of U-^{13}C-labeled extracts. A new method for accurate quantification of changes in concentrations of intracellular metabolites. *Biotechnol. Bioeng.*, 85: 620-628 (2004).

Mehra A, Lee KH and Hatzimanikatis V. Insights into the relation between mRNA and protein expression patterns: I. Theoretical considerations. *Biotechnol. Bioeng.*, 84: 822-833 (2003).

Mendes P. Emerging bioinformatics for the metabolome. *Brief. Bioinformatics*, 3: 134-145 (2002).

Monod J, Changeux, J-P., and Jacob, F. Allosteric proteins and cellular control systems. *J. Mol. Biol.*, 6: 306-329 (1963).

Moriya H and Johnston M. Glucose sensing and signalling in *Saccharomyces cerevisiae* through the Rgt2 glucose sensor and casein kinase I. *Proc. Natl. Acad. Sci. USA.*, 101: 1572-1577 (2004).

Mosley AL, Lakshmanan J, Aryal BK and Ozcan S. Glucose-mediated phosphorylation converts the transcription factor Rgt1 from a repressor to an activator. *J. Biol. Chem.*, 278: 10322-10327 (2003).

Muller D, Exler S, Aguilera-Vazquez L, Guerrero-Martin E and Reuss M. Cyclic AMP mediates the cell cycle dynamics of energy metabolism in *Saccharomyces cerevisiae*. *Yeast*, 20: 351-367 (2003).

Muratani M and Tansey WP. How the ubiquitin-proteasome system controls transcription. *Nat. Rev. Mol. Cell. Biol.*, 4: 192-201 (2003).

Noble ME, Endicott JA and Johnson LN. Protein kinase inhibitors: insights into drug design from structure. *Science*, 303: 1800-1805 (2004).

Oliver DJ, Nikolau B and Wurtele ES. Functional Genomics: high-throughput mRNA, protein, and metabolite analyses. *Metab. Eng.*, 4: 98-106 (2002).

Oliver SG. Yeast as a navigational aid in genome analysis. *Microbiology*, 143: 1483-1487 (1997).

Oliver SG, Winson MK, Kell DB., and Baganz, F. Systematic functional analysis of the yeast genome. *Trends Biotechnol.*, 16: 373-378 (1998).

Peletier MA, Westerhoff HV, Kholodenko BN. Control of spatially heterogeneous and time-varying cellular reaction networks: a new summation law. *J. Theor. Biol.*, 225: 477-487 (2003).

Petroski RJ and McCormick SP. *Secondary-metabolite biosynthesis and metabolism*, Kluwer Academic/Plenum Publishers, New York (1992).

Phelps TJ, Palumbo AV and Beliaev AS. Metabolomics and microarrays for improved understanding of phenotypic characteristics controlled by both genomics and environmental constraints. *Curr. Opin. Biotechnol.*, 13: 20-24 (2002).

Plaxton WC. Principles of metabolic control, in: *Functional Metabolism of Cells: Control, Regulation, and Adaptation*, K B. Storey, ed., John Wiley and Sons, Inc., New York, pp. 1- 23 (2004).

Quant PA. Experimental application of top-down control analysis to metabolic systems. *Trends Biochem. Sci.*, 18: 26-30 (1993).

Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ, Westerhoff HV, van Dam K and Oliver SG. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.*, 19: 45–50 (2001).

Rohde JR and Cardenas ME. The tor pathway regulates gene expression by linking nutrient sensing to histone acetylation. *Mol. Cell Biol.*, 23: 629-635 (2003).

Roncal T and Ugalde U. Conidiation induction in Penicillium. *Res. Microbiol.*, 54: 539-546 (2003).

Rose AH and Harrison JS. *The Yeasts*. Vol. 1-6. Academic Press, London (1987-1995).

Saez MJ and Lagunas R. Determination of intermediary metabolites in yeast. Critical examination of the effect of sampling conditions and recommendations for obtaining true levels. *Mol. Cell. Biochem.*, 13: 73-78 (1976).

Sambrook J and Russell D. *Molecular Cloning: a laboratory manual*, 3rd edition. Cold Spring Harbor Laboratory Press. Cold Spring Harbor. New York (2000).

Sandelin A, Hoglund A, Lenhard B and Wasserman WW. Integrated analysis of yeast regulatory sequences for biologically linked clusters of genes. *Funct. Integr. Genomics*, 3: 125-134 (2003).

Schilter B and Constable A. Regulatory control of genetically modified (GM) foods: likely developments. *Toxicol. Lett.*, 127: 341-349 (2002).

Schmitt S and Paro R. A reason for reading nonsense. *Nature*, 429: 510-511 (2004).

Segre D, Zucker J, Katz J, Lin X, D'Haeseleer P, Rindone WP, Kharchenko P, Nguyen DH, Wright MA and Church GM. From annotated genomes to metabolic flux models and kinetic parameter fitting. *OMICS*, 7: 301-316 (2003).

Sellick CA and Reece RJ. Modulation of transcription factor function by an amino acid: activation of Put3p by praline. *EMBO J.*, 22: 5147-5153 (2003).

Sprague GF Jr, Cullen PJ and Goehring AS. Yeast signal transduction: Regulation and interface with cell biology, in: *Advances in Experimental Medicine and Biology,* Vol. 547, *Advances in Systems Biology*, L. K. Opresko, J. M. Gephart, and M. B. Mann, eds. Kluwer Academic/Plenum Publishers, New York, pp. 91-105 (2004).

Stockton GW, Aranibar N and Ott K-H. Metabolome profiling methods using chromatographic and spectroscopic data in pattern recognition analysis. World Intellectual Property Organisation, *WO Patent*, 02057989 (2002).

Sudarsan N, Barrick JE and Breaker RR. Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA*, 9: 644-647 (2003).

Ter Kuile BH and Westerhoff HV. Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett.*, 500: 169-171 (2001).

Teusink B, Baganz F, Westerhoff HV and Oliver SG. Metabolic control analysis as a tool in the elucidation of the function of novel genes. In: *Methods in Microbiology*, 26. A. J. Brown and M. F. Tuite, eds., Academic Press, London, pp. 297-336 (1998).

Theobald U, Mailinger W, Reuss M and Rizzi M. *In vivo* analysis of glucose-induced fast changes in yeast adenine nucleotide pool applying a rapid sampling technique. *Anal. Biochem.*, 214: 31-37 (1993).

Trethewey RN. Gene discovery via metabolic profiling. *Curr. Opin. Biotechnol.*, 12: 135-138 (2001).

Trethewey RN, Krotzky AJ and Willmitzer L. Metabolic profiling: a Rosetta Stone for genomics? *Curr. Opin. Plant Biol.*, 2: 83-85 (1999).

Urbanczyk-Wochniak E, Luedemann A, Kopka J, Selbig J, Roessner-Tunali U, Willmitzer L and Fernie AR. Parallel analysis of transcript and metabolic profiles: A new approach in systems biology. *EMBO Rep.*, 4: 989–993 (2003).

Vaidyanathan S, Rowland JJ, Kell DB and Goodacre R. Discrimination of aerobic endospore-forming bacteria via electrospray ionization mass spectrometry of whole cell suspensions. *Anal. Chem.*, 73: 4134-4144 (2001).

Villas-Boas SG, Delicado DG, Akesson M and Nielsen J. Simultaneous analysis of amino and nonamino organic acids as methyl chloroformate derivatives using gas chromatography-mass spectrometry. *Anal. Biochem.*, 322: 134-138 (2003).

Watkins SM and German JB. Toward the implementation of metabolomic assessments of human health and nutrition. *Curr. Opin. Biotechnol.*, 13: 512-516 (2002).

Weckwerth W. Metabolomics in systems biology. *Annu. Rev. Plant Biol.*, 54: 669-689 (2003).

Weckwerth W and Fiehn O. Can we discover novel pathways using metabolomic analysis? *Curr. Opin. Biotechnol.*, 13: 156-160 (2002).

Weckwerth W and Fiehn O. Combined metabolomic, proteomic and transcriptomic analysis from one, single sample and suitable statistical evaluation data. World Intellectual Property Organisation, *WO Patent*, 03058238 (2003).

Winkler WC, Nahvi A, Roth A, Collins JA and Breaker RR. Control of gene expression by a natural metabolite-responsive ribozyme. *Nature*, 428: 281-286 (2004).

Wu LF, Hughes TR, Davierwala AP, Robinson MD, Stoughton R and Altschuler SJ. Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.*, 31: 255-265 (2002).

Yao T. Bioinformatics for the genomic sciences and towards systems biology. Japanese activities in the post-genome era. *Prog. Biophys. Mol. Biol.*, 80: 23-42 (2002).

Yoon SH and Lee SY. Comparison of transcript levels by DNA microarray and metabolic flux based on flux analysis for the production of poly-γ-glutamic acid in recombinant *Escherichia coli*. *Genome Informatics*, 13: 587-588 (2002).

Yoon SH, Han MJ, Lee SY, Jeong KJ and Yoo JS. Combined transcriptome and proteome analysis of *Escherichia coli* during the high cell density culture. *Biotechnol. Bioeng.*, 81: 753-767 (2003).

Zaragoza O, Lindley C and Gancedo JM. Cyclic AMP can decrease expression of genes subject to catabolite repression in *Saccharomyces cerevisiae*. *J. Bacteriol.*, 181: 2640-2642 (1999).

Chapter 3

METABOLOMICS FOR THE ASSESSMENT OF FUNCTIONAL DIVERSITY AND QUALITY TRAITS IN PLANTS

Robert D. Hall, C.H.Ric de Vos, Harrie A. Verhoeven, Raoul J. Bino.
Plant Research International, Business Unit Bioscience, P.O. Box 16, 6700 AA Wageningen, The Netherlands

1. INTRODUCTION

From the outset there has been tremendous interest in the potential of metabolomics technologies to expand our fundamental knowledge of biological systems and no more so than in the field of plant science. The number of reviews written in the early years of metabolomics significantly outnumbered the number of true, research-driven scientific papers. With other functional genomics technologies paving the way to bigger and better things, scientists' appetites have been whetted for holistic approaches to the study of bio-molecular organisation in living organisms. Metabolomics is not only complementary to the other 'omics' technologies but also is considered to have clear additional advantages (Goodacre *et al.*, 2004). As metabolites are the most distant products downstream from gene expression, changes in the metabolome should be amplified with respect to those for the transcriptome and proteome. Indeed, the metabolome should most closely reflect the activities of a cell at the functional level (Goodacre *et al.*, 2004).

Particularly in plants, where richness and diversity in metabolic composition is unsurpassed among all groups of living organisms (Hall *et al.*, 2002), a metabolomics approach offers a new complementary addition to already-existing functional genomics techniques. In addition, because of its relatively unbiased nature, metabolomics is appropriate for complex analyses of often poorly-predefined systems. While the technology is still in its infancy, expectations are considerable and multiple applications in widely

diverse fields of interest are now evident and further envisaged. The plant world in particular is poised to drive the technology forward. A key task of plant-oriented research groups is now to establish a multidisciplinary approach essential for successful future initiatives. Only with correct, co-ordinated and complementary input from biochemists, technologists, physiologists, bioinformaticists and statisticians, applied within a well defined research framework and driven by the right biological questions, will we reach the stage where metabolomics can truly become an essential tool in biological research.

In this chapter we detail the current aims and achievements of metabolomics technology and indicate how metabolomics is and will continue to be applied to generate information needed to yield a better understanding of the molecular organisation of plants. With this information we can then develop novel, dedicated strategies to direct metabolism to the improvement of plants and plant products.

2. NOVEL STRATEGIES AND CHALLENGES FOR NON-TARGETED BIOCHEMICAL ANALYSES OF PLANT MATERIAL

Metabolomics can be regarded as the non-targeted comprehensive analysis of the composition of complex biochemical mixtures such as plant extracts (Fiehn, 2002, 2003; Hall *et al.*, 2002). The primary challenge is therefore to generate a technology which is robust and which covers the broadest possible qualitative and quantitative range of metabolites. This switch from a traditional reductionist approach to a novel, holistic approach implies a number of inevitable consequences. The metabolomics challenge relates to difficulties which arise due to the broad spectrum of metabolic structures which should be analysed as well as the broad dynamic range of the metabolic components involved. While some metabolites in a plant extract may approach molar concentrations, others, of potentially equal biological and phenotypic importance, may only occur in the micro to nanomolar range. The combination of chemical complexity, metabolic heterogeneity, dynamic range and ease of extraction therefore represent the most significant challenges facing us today in the quest for an effective functional metabolomics technology platform (Goodacre *et al.*, 2004).

Many different extraction and detection techniques have been applied and with a considerable degree of success. Excellent reviews of the technologies available, overviews of the different strategies and comparative analyses of their advantages and limitations can be recommended (e.g. Fiehn, 2002, 2003; Fernie, 2003; Goodacre *et al.*, 2004; Mendes, 2002;

Niessen, 2003; Roessner *et al.*, 2002; Sumner *et al.*, 2003; Weckwerth, 2003). Currently, the most widely implemented approaches are based upon GC-MS and HPLC-MS techniques which offer the most optimal reproducibility, comprehensiveness, sensitivity and dynamic range (see Chapter 7). In some cases, in the search for a high throughput fast (pre)screening approach, the chromatographic component has even been removed and direct infusion has been employed to produce an initial general metabolic fingerprint (Aharoni *et al.*, 2002; Castrillo *et al.*, 2003; Goodacre *et al.*, 2002; Verhoeven *et al.*, 2003). Run times of as short as 30 seconds have been used and despite the potentially low resolving power, reliable comparative analyses have proven possible (Goodacre *et al.*, 2003). Other approaches such as NMR (Defernez and Colquhoun, 2003; Ward *et al.*, 2003), FT-IR spectroscopy (Johnston *et al.*, 2003) and FT-ICRMS (Aharoni *et al.*, 2002) are also receiving attention.

A primary message must nevertheless be emphasised – all current methodologies and detection techniques, irrespective of their high level of sophistication, have unavoidable intrinsic bias against certain metabolite groups. No single extraction or detection technique therefore suffices and multiparallel technologies (Roessner *et al.*, 2002) will continue to be necessary to gain the desired comprehensive assessment of the metabolic composition of biological material. Even then, it will likely remain the case that 'metabolomics' will continue to be more about defining an aim than ever achieving reality (Fiehn, 2003).

The development of dedicated bioinformatics tools is also essential for realising the full potential of any metabolomics strategy. When complex spectral patterns are produced, as are typical of MS technologies, tools are needed to perform automated, comparative *in silico* analyses. Only by effectively eliminating those mass peaks incidental to an observed phenotype can we recognise and focus on those peaks representing the main differences between test and control samples. For this, both analytical and statistical software tools are required. Chemometric approaches together with unsupervised techniques such as hierarchical clustering and principal component analysis (PCA) are already widely applied (Fiehn *et al.*, 2000; Fernie, 2003; Sumner *et al.*, 2003; von Roepenack-Lahaye *et al.*, 2004). However, more advanced techniques such as genetic programming in combination with suitable visualisation tools are still required (Kell, 2002, 2004; Mendes, 2002; Goodacre *et al.*, 2004). Without these tools it will remain difficult to discriminate reliably between samples on the scale required to enable us to extract biologically meaningful information from multivariate datasets (Kose *et al.*, 2001).

3. METABOLOMICS, PLANT PHYSIOLOGY AND PLANT BREEDING

Two of the major areas where metabolomics will prove an invaluable research tool are plant physiology and plant breeding. Metabolomics may indeed prove to be the best and most direct measure of plant physiology and it is already clear that a metabolomics perspective gives us a clear and unambiguous picture of what is going on at the level of the cell (Beecher, 2002). The non-targeted nature of metabolomics leads to an understanding of connections and relationships between metabolites which are not intuitive and provides us, for the first time, with a unique insight into the complexity of these interactions. Through enhancing our understanding of the fundamental molecular basis of the physiology of plants and by following the manner in which this is influenced by biotic and abiotic factors within and beyond our control (genetics, cultivation, treatment applications, environment etc), we gain a greater insight into how plants function and into how plants exploit their metabolic plasticity in an ever-changing and often hostile environment. With this information we shall take up a more effective position from which to develop novel targeted strategies to improve plants in terms of their productivity, suitability for specific ecological conditions, product quality, resistance / tolerance to environmental factors etc.

Research by Roessner *et al.* (2000, 2001a, 2001b) on potato physiology and tuber development not only represented a watershed in the establishment of metabolomics as an extra weapon in the functional genomics arsenal but also provided the first detailed pictures of metabolic profiles from single extracts for comparative, synchronous biochemical analyses of plant materials. Developing tubers grown in the greenhouse as well as *in vitro*-grown microtubers were analysed and compared. Approximately 150 compounds of diverse biochemical origin were detected and quantified. The methodology was demonstrated to be robust and the simultaneous analysis of groups of, generally primary, metabolites revealed clear differences between the tuber systems. Subsequently, combining this approach with reverse genetics proved a powerful tool with which to phenotype, metabolically, potato tubers which had been modified either environmentally or genetically (Roessner *et al.*, 2001a, 2001b). Concurrently, the groundwork was laid both for the concept of metabolic networking and, through the exploitation of statistical and bioinformatic tools, for detailed correlation analyses demonstrating the interactive and interdependent nature of metabolic profiles in the context of plant physiology.

Since the pioneering work of Roessner and colleagues, metabolomics has been applied to interrogate the permutations in metabolic composition of a whole range of systems with regard to response to genetical and physico-chemical modifications to the environment. Using a novel FTMS approach,

Aharoni revealed the enormous complexity of changes which occur during strawberry fruit ripening despite the remarkably short time scale involved of just a few days (Aharoni *et al.*, 2002). The influence of diurnal rhythms on *Cucurbita* and *Pharbitis* phloem and leaf sap composition (involving essentially, primary metabolites (Fiehn 2003; Goodacre *et al.*, 2003)), of circadian rhythms on the release of head space volatiles from *Petunia hybrida* flowers (covering primarily, secondary metabolites; Verdonk *et al.*, 2003) and of a short day regime on the cessation of growth in poplar shoots (Kusano *et al.*, 2003) has also revealed how transient and ever-changing metabolic composition can be. This further emphasises the biochemical flexibility of plants and how rapidly changes in response to environmental perturbations can occur. In addition, this indicates the scale of temporal and spatial resolution required to produce reliable and meaningful metabolomic analyses. In the cucurbit study, for example, not only did the light / day regime result in many metabolites changing in concentration by several orders of magnitude but also, each *individual* leaf was shown to have its own unique metabolic profile. This has particular implications concerning the fundamental way in which we must perform metabolomics experiments so that the relevance of the results obtained can be correlated with possible changes in biological variation.

With the global human population set to double within just a few decades, one of the key issues which must be addressed by plant breeders concerns the development of crop varieties capable of growing beyond the borders of the environment presently suited to their cultivation. Aspects of stress tolerance in relation to salinity, temperature, water etc., need to be better understood before we can design dedicated, novel and improved breeding strategies to produce the ecotypes required. Using FT-IR and chemometrics in an inductive reasoning approach, Johnson *et al.*, (2003) were able to use metabolic profiling to discriminate between wild-type and salt-stressed tomato plants. Further classification of the differences observed will give a better understanding of plant responses to salt stress and will assist in the defining of novel hypotheses to be addressed in the search for a directed breeding strategy for salt tolerance. Quality trait assessments shall also benefit greatly from a metabolomics approach to characterise complex plant features better. Through this and similar examples, metabolomics is anticipated to play a key role in future research activities geared towards overcoming some of the key limitations to global crop production. Biochemical markers will also be mapped in a similar manner and used as a complement to the more traditional, genetic markers. Both can then be applied towards improved progeny selection in dedicated breeding strategies to match crop varieties better to local environmental, cultural and social needs.

4. THE POTENTIAL OF METABOLOMICS APPLICATIONS FOR BIODIVERSITY ASSESSMENT

It is fundamental to metabolomics technology that we are provided with a detailed and broad snap-shot of the complexity of the metabolic composition of plant materials at the time of extraction. Provided that instrumental and biological variation is accounted for, this information, initially in the form of a simple output from an analytical instrument, such as a spectrum from a mass spectrometry or NMR, can be directly exploited as a metabolic fingerprint (Fiehn 2001, 2002). As such, even without recourse to the identity of the compounds present, these spectral or chromatographic outputs can be used very effectively in "fast-track" comparative analysis. Indeed, many metabolomics approaches are geared not to performing a detailed analysis of (all) individual components but rather are initially aimed at discriminating a number of differential peaks against a highly complex background of unchanging ones (Hall *et al.*, 2002; Sumner *et al.*, 2003; Ward *et al.*, 2003).

Bioinformatics and statistical tools are being developed specifically to aid and automate this process (Goodacre *et al.*, 2004; Kell, 2004; Roepenack-Lahaye *et al.*, 2004; Tolstikov *et al.*, 2003; Verhoeven *et al.*, in preparation). The rationale is that when the aim is to compare e.g. genetic mutants (Roepenack-Lahaye *et al.*, 2004), ecotypes (Fiehn *et al.*, 2000; Ward *et al.*, 2003; Schaneberg *et al.*, 2003), genetically modified or molecularly-engineered plants (Roessner *et al.*, 2000, 2001a,b; Le Gall *et al.*, 2003), varieties (Verhoeven *et al.*, 2003), or eventually even the collected progeny from a breeding cross, it can be anticipated that the majority of compounds present will be qualitatively and quantitatively similar if not identical. Consequently, a pre-screening/filtering method to eliminate non-discriminatory mass peaks and biochemical components is required to simplify the multivariate analysis and to allow for a more concentrated effort, dedicated to those differences which are detected and which can be postulated to be causally related to any phenotypic changes observed. Time-consuming and costly confirmation of the identity of key components can then be restricted to only those peaks of potential interest. Correct use of alignment software, baseline correction and reliable noise reduction is essential.

When applied properly, such an approach can prove very effective. We have shown that applying non targeted GC-MS and LC-MS analyses followed by spectral subtraction and supported by appropriate tools, such as PCA and hierarchical clustering, is highly valuable for screening so-called silent (biochemical) mutants with no overt phenotype in a large population of expressed sequence tagged *Arabidopsis* lines. Furthermore, on assessing

the variation in natural fragrance volatiles of different varieties of cultivated roses and of some of their wild relatives using SPME-GCMS, degrees of similarity could be determined and used to predict the pedigree of the lines analysed and to form the basis of a phylogenetic tree (Verhoeven *et al.*, 2003). Detailed statistical analysis, followed by assessment of the discriminatory components also revealed the potential biochemical basis of the differences. Consequently, this information can be exploited, in the future, in a dedicated breeding strategy to return a strong fragrance to modern cultivated rose varieties, a feature lost through intensive breeding in the last century.

Based on metabolic fingerprinting using NMR combined with multivariate statistics as a pre-screening method, Ward *et al.* (2003) also demonstrated that *Arabidopsis* ecotypes could be readily and reproducibly discriminated. The authors could extract residual NMR spectra of those components contributing significantly to the ecotypic differences by applying PCA. In mice, Plumb *et al.* (2003) demonstrated with LC-MS/PCA analysis of urine that not only gender and strain could be distinguished on the basis of a metabolic fingerprint, but also differences due to diurnal variation could be identified. Metabolic fingerprinting as a rapid and simple discriminatory tool for the initial assessment of metabolic biodiversity would therefore appear to be an efficient starting point when the goal is to identify potentially small numbers of lines among e.g. extensive breeding progenies or mutant populations and for studying genetic drift in ecological studies, identifying changes arising due to genetic modification, altered food processing strategies etc.

5. METABOLOMICS AND QUALITY ASSESSMENT IN THE PRODUCTION CHAIN

The quality of plant materials is a complex issue involving a multitude of related and wholly unrelated factors. What is meant by quality is fully dependent upon the type of product and its use. However, generally speaking the quality of a plant product can most readily be defined in terms of its biochemical composition. Nutritional value is dependent on the types and amounts of key components present, such as vitamins, sugars, and proteins, which are of primary importance in our daily diet. Quality, in terms of market value, can also be determined by fundamental, metabolically-definable factors such as flavour, fragrance, colour and texture. Furthermore, many parameters related to quality such as shelf life, suitability for transportation, storage depreciation and freshness also have a tangible link to biochemical composition. Consequently, the application of metabolomics

technologies in the assessment of quality aspects of plant materials is already under detailed consideration.

In an overview on the composition of tomato fruits, van Tuinen *et al.* (2004) described the influence of certain metabolic gene mutations on the tomato metabolome and related the potential importance of the observed changes to their health-promoting potential. Burns *et al.* (2003) used a metabolic profiling approach to determine the levels of key micronutrients in fruits and vegetables with the aim of generating information useful in dietary advice. Furthermore, the authors anticipate that this information will also provide a useful starting point for both the rational engineering of health-promoting phytochemicals in fruit and vegetables and for varietal screening. In relation to the topic of nutrigenomics, Muller and Kersten (2003) predict that metabolomics will play a key role relating nutritional quality to human health. In addition, monitoring the influence of the composition of food metabolites on, for instance, human gene expression will assist in assessing the effects of dietary constituents on our health and well-being. Integrated within an epidemiological context, Detailed biochemical profiling of foodstuffs shall further assist in defining a link between diet and health, when integrated with aspects of human physiology, genetic predisposition to disease, single nucleotide polymorphisms etc., within an epidemiological context, and could ultimately result in the realisation of the concept of personal diets for consumers in high-risk categories (German *et al.*, 2002; Miller and Kersten, 2003; Watkins *et al.*, 2001).

Complex developmental processes such as ripening and organ maturation are also attractive targets for a metabolomic approach. Information generated from studies on the ripening of strawberry fruit (Aharoni *et al.*, 2002) not only provided us with a detailed insight into the changes occurring and the timing involved but also such information can subsequently be extrapolated to yield potential biochemical markers for quality monitoring. Similar studies involving volatiles could be considered for the development of a fully non-invasive quality monitoring system where use could be made of , for example, an 'electronic nose', for a semi-automated decision tool in a real time, fully integrated, quality controlled production chain.

The growing demand for safety monitoring by and for the consumer has stimulated the development of metabolomics strategies in the area of food safety and food adulteration. Consequently, the food industry is using increasingly sophisticated technologies to detect e.g. anti-nutritional components in our food. Metabolomics is also being applied to test for lower quality products which are fraudulently being used to bulk up higher value materials. In a recent study, Goodacre *et al.* (2003) described how a rapid 60 sec direct infusion MS analysis can effectively be used to define contaminants and adulterants in samples of olive oil. Reid *et al.* (2004) used SPME-GC-MS in a chemometric approach to assess the adulteration of

strawberry products with cheaper apple material. A similar application of chemical fingerprinting of botanical medicines has been described for the authentication of *Ephedra* products of varying quality derived from different global sources (Schaneberg *et al.*, 2003). In the area of food safety, metabolomics can play a key role in food monitoring in relation to undesirable changes in plant components resulting from sub-optimal cultivation conditions, modified processing strategies or as a consequence of unexpected changes resulting from classical breeding strategies, genetic modification and genetic engineering (Kuiper *et al.*, 2002; Noteborn *et al.*, 1988, 2000).

6. THE ROLE OF METABOLOMICS TOWARDS A SYSTEMS LEVEL UNDERSTANDING

Undoubtedly, the most significant consequence of entering the metabolomics era is that this will lead to the most complete understanding of plant function by providing an unprecedented insight into the integral complexity and highly interactive nature of the biochemical composition of plants. In combination with mutant screening and the use of reverse genetics, approaches to achieve systematic perturbation of gene expression, we shall gain a much better position from which to elucidate the organisational complexity of complete genomes. By essentially beginning blind, without preconceptions, we will be able to distinguish, in the coming years, those compounds exhibiting greatest variation between genetically diverse lines and those resulting from a range of physico/chemical treatments, enabling us to propose causative, hitherto unknown, relationships between genes, metabolites and phenotypes. The power of unsupervised correlative analyses, when applied to metabolomic datasets, has laid the groundwork for true metabolic networking to give us a more realistic dynamic view of interactive pathway regulation (Roessner *et al.*, 2001a). Enhanced knowledge of the extent of the interactive nature of metabolic networks and metabolic co-dependency (Fernie, 2003) will place us in a better position to assess the biological and, ultimately, the commercial implications of metabolite synthesis, accumulation, turnover, etc.

Following the work of Roessner *et al.* (2000), a change in philosophy resulted directing us to view metabolites not in terms of linear pathways but, more substantially, in terms of highly regulated and integrated networks (Trethewey, 2004, Chapter 7). The consequences of aspects such as pleiotropic effects, feedback inhibition and other internal compensatory mechanisms on biological systems can now be systematically and rigorously assessed in the context of the complete metabolome. Metabolomics provides us with a better insight into the dynamic interactions typifying plant

metabolic networks, while enabling us to define and dissect chemical correlations between and within pathways. In so doing, this will allow us to identify pathways not yet characterised or even recognised. Previously unconsidered relationships (causal connectivity, Weckwerth, 2003) between seemingly unrelated pathways may then come to light (Carrari *et al.*, 2003). The comprehensive metabolic profiling of large numbers of metabolites can be used to query holistic responses of biological systems to external stimuli and will further extend our capacity to harness the biochemical diversity of nature to the benefit of mankind (Dixon and Sumner, 2003). Exploiting this, and by augmenting metabolomics approaches with other functional genomics and physiological strategies, the degree of predictability is greatly enhanced and we are then more able than ever to design dedicated, traditional or genetic modification strategies for crop improvement.

7. SUMMARY AND CONCLUSIONS

Metabolomics provides us with the dedicated tools required to expose and dissect the controlled chaos that is plant metabolism. A better understanding of the molecular complexities of plants will assist in developing novel, targeted strategies for plant improvement and it is evident that metabolomics technologies will continue to provide us with an unprecedented source of valuable information. There are many areas of biology where metabolomics can be very effectively applied in widening our knowledge. In the field of gene function analysis alone, metabolomics, as a complement to other multidisciplinary approaches, will provide us with manifold new opportunities to link functions to many of those thousands of genes which to date have not yet been assigned even a putative function (Schwab, 2003; Weckwerth, 2003). Metabolomics enables the formation of a conceptual basis from which we can elucidate the mechanisms underlying plant phenotype and allows us to query phenotypic responses to internal and external environmental perturbations in the most holistic manner. Upon making the logical assumption that the emerging patterns bear a relationship to the underlying molecular framework (Stauer, 2003), novel approaches can be designed for modifying the biochemical composition of plants and plant materials better, in accordance with requirements.

Many improvements are still necessary and co-operation and collaboration is essential for future development (Dixon and Stack, 2003). The continued success of metabolomics will provide a new driving force for additional, more sophisticated tools for analysis. An ultimate goal will be the development of tools required to perform metabolomics analyses on single cells or organelles in order to enable us to dissect out and clarify the contribution of the spatial element. In this way we shall gain a more detailed

insight into the key differences between those cell types constituting a plant organ and how this differentiation comes about. However, metabolomics remains solely a tool and not a goal. Metabolomics only provides us with a starting point and it is the interpretation of the information obtained and the confirmation of its true biological relevance which justifies the attention the technology receives. It is all very well initiating a non-targeted metabolite profiling strategy, followed by unsupervised data analysis, in the context of a holistic approach to plant physiology research, but if this is not, from the outset, driven and directed by properly-formulated and focussed biological questions then the outcome will be meaningless. In this regard, further development of chemometric, statistical and bioinformatic tools will prove critical. The major challenge remaining is the functional integration of the information obtained from metabolic profiling into an accessible body of knowledge (Stitt and Fernie, 2003).

ACKNOWLEDGEMENTS

Plant Research International, The Dutch Ministry of Agriculture, Nature and Food and The Centre for Biosystems Genomics are acknowledged for financial support.

REFERENCES

Aharoni A, de Vos CHR, Verhoeven HA, Maliepaard CA, Kruppa G, Bino RJ, Goodenowe DB. Nontargeted metabolome analysis by use of Fourier Transform Ion Cyclon Mass Spectrometry. *OMICS*, 6: 217 – 234 (2002).

Beecher C. Metabolomics: a new 'omics' technology. *Am. Genomics / Proteomics technol.*, May-June (2002)

Burns J, Fraser PD, Bramley PM. Identification and quantification of carotenoids, tocopherols and chlorophylls in commonly-consumed fruits and vegetables. *Phytochemistry*, 62: 939-947 (2003).

Carrari F, Urbanczyk-Wochniak E, Willmitzer L, Frenie AR. *Metabol. Eng.*, 3: 191-200 (2003).

Castrillo JI, Hayes A, Mohammed S, Gaskell SJ, Oliver SG. An optimized protocol for metabolome analysis in yeast using direct infusion electrospray mass spectrometry. *Phytochemistry*, 62: 929-937 (2003).

Defernez M, Coloquhoun IJ. Factors affecting the robustness of metabolic fingerprinting using ^{1}H NMR spectra. *Phytochemistry*, 62: 1009-1017 (2003).

Dixon RA, Strack D. Phytochemistry meets genome analysis, and beyond. *Phytochemistry*, 62: 815-816 (2003).

Dixon RA, Sumner LW. Legume natural products: understanding and manipulating complex pathways for human and animal health. *Plant Physiol.*, 131: 88-885 (2003).

Fernie AR. Metabolome characterisation in plant systems analysis. *Func. Plant Biol.*, 30: 111-120 (2003).

Fiehn O. Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comp. Func. Genomics*, 2: 155-168 (2001).

Fiehn O. Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.*, 48: 155-171 (2002).

Fiehn O. Metabolic networks of *Cucurbita maxima* phloem. *Phytochemistry*, 62: 875-886 (2003).

Fiehn O, Kloska S, Altmann T. Integrated studies on plant biology using multiparallel techniques. *Curr. Opin. Biotechnol.*, 12: 82-86 (2001).

Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN, Willmitzer L. Metabolic profiling for plant functional genomics. *Nat. Biotechnol.*, 18: 1157-1161 (2000).

Fiehn O, Weckwerth W. Deciphering metabolic networks. *Eur. J. Biochem.*, 270: 579-588 (2003).

German JB, Roberts MA, Fay L, Watkins SM. Metabolomics and individual metabolic assessment: the next challenge for nutrition. *J. Nutrition*, 132: 2486-2487 (2002).

Goodacre R, Vaidyanathan S, Bianchi G, Kell DB. Metabolic profiling using direct infusion electrospray ionisation mass spectrometry for the characterisation of olive oils. *Analyst*, 127: 1457-1462 (2002).

Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB. Metabolomics by numbers: acquiring and understanding global metabolomics data. *Trends Biotechnol.*, 22: 245-252 (2004).

Goodacre R, York EV, Heald JK, Scott IM. Chemometric discrimination of unfractionated plant extracts analysed by electrospray mass spectrometry. *Phytochemistry*, 62: 859-863 (2003).

Hall RD, Beale M, Fiehn O, Hardy N, Sumner L, Bino R. Plant metabolomics: the missing link in functional genomics strategies. *The Plant Cell*, 14: 1437-1440 (2002).

Johnston HE, Broadhurst D, Goodacre R, Smith AR. Metabolic fingerprinting of salt-stressed tomatoes. *Phytochemistry*, 62: 919-928 (2003).

Kell DB. Metabolomics and machine learning: explanatory analysis of complex metabolome data using genetic programming to produce simple, robust rules. *Mol. Biol. Reports*, 29: 237-241 (2002).

Kell DB. Metabolomics and systems biology: making sense of the soup. *Curr. Opin. Microbiol.*, 7: 296-307 (2004).

Kose F, Weckwerth W, Linke T, Fiehn O. Visualizing plant metabolomic correlation networks using clique-metabolite matrices. *Bioinformatics*, 17: 1198-1208 (2001)

Kuiper HA, Noteboorn HPJM, Kok EJ, Kleter GA. Safety aspects of novel foods. *Food Res. Int.*, 35: 267-271 (2002).

Kusano M, Öberg K, Jonsson P, Gullberg J, Sjöstrom, Moritz T. Identification of metabolic changes during short-day induced cessation of elongation growth in Poplar. Poster *2nd International Plant Metabolomics Congress*, Potsdam, 2003

Le Gall G, DuPont MS, Mellon FA, Davies AL, Collins GJ, Verhoeyen ME, Colquhoun IJ. Characterisation and content of flavonoid glycosides in genetically modified tomato (*Lycopersicon esculentum*) fruits. *J. Agri. Food Chem.*, 51: 2438-2446 (2003).

Mendes P. Emerging bioinformatics for the metabolome. *Brief. Bioinformatics*, 3: 134-145 (2002).

Muller M, Kersten S. Nutrigenomcs: goals and strategies. *Nat. Rev. Genetics*, 4: 315-322 (2003).

Niessen WMA. Progress in liquid chromatography-mass spectrometry instrumentation and its impact on high-throughput screening. *J. Chromat. A*, 1000: 413-436 (2003).

Noteboorn HPJM, Lommen A, van der Jagt RCM, Weseman JM. Chemical fingerprinting for the evaluation of unintended secondary metabolic changes in transgenic food crops. *J. Biotechnol.*, 77: 103-114 (2000).

Noteboorn HPJM, Lommen A, Weseman JM, van der Jagt RCM, Groenendijk, FPJ. Chemical fingerprinting and *in vitro* toxicological profiling for the safety evaluation of transgenic food crops. In: Horning M (Ed), *Food safety evaluation of genetically modified foods as a basis for market introduction*, pp 51-79. Report, Ministry of Economic Affairs, The Hague (1998).

Plumb R, Granger J, Stumpf C, Wilson ID, Evans JA, Lenz EM. Metabonomic analysis of mouse urine by liquid chromatography time of flight mass spectrometry (LC-TOFMS): detection of strain, diurnal and gender differences. *The Analyst*, 128: 819-823 (2003).

Reid LM, O'Donnell CP, Downey G. Potential of SPME-GC and chemometrics to detect adulteration of soft fruit purées. *J. Agri. Food Chem.*, 52: 421-427 (2004).

Roessner, U, Luedemann A, Brust D, Fiehn O, Linke T, Willmitzer L, Fernie A. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *The Plant Cell*, 13: 11-29 (2001b)

Roessner U, Wagner C, Kopa J, Trethewey RN, Willmitzer L. Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *The Plant J.*, 23: 131-142 (2000).

Roessner U, Willmitzer L, Fernie AR. High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. *Plant Physiol.*, 127: 746-764 (2001a).

Roessner U, Willmitzer L, Fernie AR. Metabolic profiling and biochemical phenotyping of plant systems. *Plant Cell Reports*, 21: 189-196 (2002).

Roessner-Tunali U, Hegeman B, Lytovchenko A, Carrari F, Bruedigam C, Granot D, Fernie AR. Metabolic profiling of transgenic tomato plants overexpressing hexokinase reveals that the influence of hexose phosphorylation diminishes during fruit development. *Plant Physiol.*, 133: 84-99 (2003).

Schaneberg BT, Crockett S, Bedir E, Khan IA. The role of chemical fingerprinting: application to Ephedra. *Phytochemistry*, 62: 911-918 (2003).

Schwab W. Metabolome diversity: too few genes, too many metabolites? *Phytochemistry*, 62: 837-849 (2003).

Stauer R, Kurths J, Fiehn O, Weckwerth W. Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, 19: 1019-1026 (2003).

Stitt M, Fernie AR. From measurements of metabolites to metabolomics: an 'on the fly' perspective illustrated by recent studies of carbon-nitrogen interactions. *Curr. Opin. Biotechnol.*, 14: 136-144 (2003).

Sumner LW, Mendes P, Dixon RA. Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry*, 62: 817-836 (2003).

Tolstikov W, Lommen A, Nakanishi K, Tanaka N, Fiehn O. Monolithic silica-based, reversed-phase, liquid-chromatography/electrospray mass spectrometry for plant metabolomics. *Anal. Chem.*, 75: 6737-6740 (2003).

Trethewey RN. Metabolite profiling as an aid to metabolic engineering in plants. *Curr. Opin. Plant Biol.*, 7: 196-201 (2004).

van Tuinen A, de Vos CHR, Hall RD, van der Plas LHW, Bino RJ. Use of metabolomics for development of tomato mutants with enhanced nutritional value by exploiting natural non-GMO light-hyperresponsive mutants. In Jaiwal PK (Ed.), Improving the nutritional and therapeutic qualities of plants, *Plant Genetic Engineering* Vol. 7, SciTech Publishers, Houston, USA (*in press*)

Verdonk JC, de Vos CHR, Verhoeven HA, Haring MA, van Tunen AJ, Schuurink RC. Regulation of floral scent production in *Petunia* revealed by targeted metabolomics. *Phytochemistry*, 62: 997-1008 (2003).

Verhoeven HA, Blaas J, Brandenburg WA. Fragrance profiles of wild and cultivated roses. In: Roberts AV, Debener T, Gudin S (Eds). *Encyclopedia of Rose Science*, Vol. 1, pp 240-248, Elsevier Academic Press, Amsterdam, The Netherlands (2003)

von Roepenack-Lahaye E, Degenkolb T, Zerjeski M, Franz M, Roth U, Wessjohann L, Schmidt J, Scheel D, Clemens S. Profiling of Arabidopsis secondary metabolites by capillary liquid chromatography coupled to electrospray ionisation quadrupole time-of-flight mass spectrometry. *Plant Physiol.*, 134: 548-559 (2004).

Ward JL, Harris C, Lewis J, Beale MH. Assessment of ^{1}H NMR spectroscopy and the multivariate analysis as a technique for metabolite fingerprinting of *Arabidopsis thaliana*. *Phytochemistry*, 62: 949-957 (2003).

Watkins SM, Hammock BD, Newman JW, German JB. Individual metabolism should guide agriculture towards foods for improved health and nutrition. *Am. J. Clin. Nut.*, 74: 283-286. (2001).

Weckwerth W. Metabolomics in systems biology. *Ann. Rev. Plant Physiol.*, 54: 669-689 (2003).

Chapter 4

METABOLOMICS: A NEW APPROACH TOWARDS IDENTIFYING BIOMARKERS AND THERAPEUTIC TARGETS IN CNS DISORDERS

Rima Kaddurah-Daouk [1,*], Bruce S. Kristal [2], Mikhail Bogdanov [3], Wayne R. Matson [4], M. Flint Beal [3]

[1]Metabolon Inc. 800 Capitola Dr., Suite 1, Durham NC 27713, USA; [2]Departments of Biochemistry and Neuroscience, Weill Medical College of Cornell University, 1300 York Ave, NY, NY 10021, USA; and Dementia Research Service, Burke Medical Research Institute, 785 Mamaroneck Ave, White Plains, NY 10605, USA; [3]Weill Medical College of Cornell University, 525 East 68 St., NY 10021, USA; [4]ESA, Inc., 22 Alpha Road, Chelmsford, MA 01824, USA

**Current address: Duke University Medical Center, Department of Psychiatry, Box 3950, Durham NC 27710.*

1. INTRODUCTION

Neurodegenerative diseases, including Alzheimer's disease (AD), Parkinson's disease (PD), Huntington's disease (HD) and Amyotrophic Lateral Sclerosis (ALS) are poorly understood disorders for which there are no effective therapies. Both genetic and environmental factors are thought to contribute to these disease states, which involve a different subset of neurons in each case. Many of these conditions manifest themselves late in life and are therefore considered to be diseases of aging. Thus, as life expectancy increases, the prevalence of these diseases will increase as well. The current patient population of around 15 million is expected to grow to 20 million by 2010. Diseases of the central nervous system (CNS), which include psychiatric disorders as well as neurodegenerative diseases, have major economic impact.

Although some progress has been made in the treatment of neurodegenerative disorders, there is still a large unmet need for more effective therapies that will slow and possibly halt disease progression. Additionally, there is a pressing need for early disease detection. Extensive research has demonstrated that neuronal degeneration is initiated well before symptoms appear. At the time disease is confirmed and therapy initiated a significant number of neurons will have already been destroyed (DeKosky and Marek, 2003). Hence, early detection is important for successful treatment. This requires the ability to monitor disease progression effectively, and reliable biomarkers could fulfill this function. In principle, biomarkers could be used to identify individuals at risk at the preclinical stage of disease, provide better diagnostic and surrogate markers of disease and its progression, allow clinicians to provide a more accurate prognosis, enable better classification of patients, and provide insights into disease mechanisms. Metabolomics is emerging as a powerful new technology platform that could play a key role in the identification of biomarkers of CNS diseases. Additionally, this technology provides the promise of mapping global biochemical perturbations in individuals with CNS disorders that might suggest new approaches for therapy. In this chapter, we will discuss biomarkers and the use of metabolomics in the study of CNS disorders.

2. BIOMARKERS OF DISEASE: AN OVERVIEW

Biological markers or biomarkers refer to cellular, biochemical, or molecular alterations that occur during disease and that are measurable in biological matrices such as tissue, cells, or fluids (Hulka, 1990; Mayeux, 2003, 2004). Biomarkers can, for example, be indicators of exposure to certain risk factors, or markers of the disease state itself. Such markers of disease state could provide a powerful tool to monitor disease and its progression, gain insights into disease mechanisms, and evaluate responses to therapy. Biomarkers need to be validated carefully at different stages of the disease and experimental design carefully evaluated. If a disease course is slowly progressive, and a lengthy longitudinal study is required, issues of timing, persistence, drug dose, selection of appropriate body fluid for analysis, and appropriate sample storage and handling are all important factors in ensuring rigorous biomarker evaluation.

Biomarkers of exposure or antecedent markers are used in risk prediction and can possibly reveal environmental and other factors that result in a disease state (Mayeux, 2003, 2004). There is a great need to identify environmental factors that contribute to neurodegenerative diseases (Tsang

and Soong, 2003; Le Couteur *et al.*, 2002; Sherer *et al.*, 2002). Relying on history of exposure to a suspected risk factor or trying to quantify exposure to an environmental toxin externally is not reliable. The direct measurement of these toxins in a body tissue or fluid or the measurement of biomarkers that directly reflect exposure to a toxin improves the sensitivity and specificity of measurement of the exposure or risk factors.

The ability to identify biomarkers that indicate the susceptibility of individuals to disease is powerful. The field of molecular genetics has already improved our ability to diagnose certain neurodegenerative diseases. An excellent example is HD, which is caused by expansion of a CAG repeat in the Huntington gene (Myers, 2004). Additional biomarkers or disease signatures could potentially identify subpopulations of HD patients with different degrees of susceptibility (Rohlff 2001; Merikangas 2002; Muller and Graeber, 1996). Another example is provided by the identification of variant *APOE* alleles that are associated with increased risk for AD and provide information regarding the pathogenesis of this condition (Liddell *et al.*, 2001; Irizarry, 2004). This information could help screen for additional environmental or genetic risk factors that contribute to AD.

Biomarkers of disease state are useful as indicators of the stage of the disorder or to monitor its progression, and different body fluids, including blood, urine, or cerebrospinal fluid (CSF), can be used to provide needed information. It is important to identify markers of disease pre-clinically, if possible, to recognize individuals who are destined to become affected or who are at a very early stage of disease. Early treatment improves the chances for a favorable outcome. Additionally, there is a great need to try to identify markers that can indicate heterogeneity in a patient population to determine who will respond better to a particular therapy. Surrogate markers that indicate stages of disease progression are also very useful in clinical trials. These could replace typical clinical endpoints such as survival which can take a long time to assess.

The search for biomarkers that might be useful in drug discovery and development is an active area of research (Frank and Hargreaves, 2003; Rolan *et al.*, 2003; de Gruttola, 2001). Reliable clinical biomarkers of disease progression could affect the pathway of drug development at each stage. The use of these markers could result in increased drug efficacy and reduced toxicity, significantly reducing the risk in drug development. Reliable biomarkers should provide measures of parameters that include the delivery of a drug to its intended targets and should predict pathophysiology and response to drug therapy. Ideally, these biomarkers should be used at the early stages of drug development. Millions of dollars are spent on clinical trials that fail because they extrapolate from animal studies to humans. We know that animal models do not reflect all aspects of the human disease and

we also know that patients are not all one and the same. Many clinical trials fail because they do not adequately take these factors into consideration. The combination of genetic diversity between individuals with a given disease and complexity of drug responses, it has become clear that more than one indicator of drug efficacy might be needed. It is believed that a combination of approaches - using data from genetics, transcriptomics, proteomics, metabolomics, clinical epidemiology, and imaging - will turn out to be the most informative way of identifying multiple useful biomarkers.

Some issues and concerns in the development of biomarkers are variability, validity, measurement of errors, bias, confounding cost, and acceptability. Analytical reproducibility is essential. Biological variability is a major concern as there are inter-individual variations that cannot be avoided. The ability of a biomarker to distinguish between two groups (for example, individuals with and without a given disease) is most commonly measured by specificity, sensitivity, and positive and negative predictive power, among other measures. Positive predictive value is the percentage of people with a positive test who actually have the disease. This value provides information about the likelihood of disease being present if a test is positive. Negative predictive value is the percentage of people with a negative test who do not have the disease. These measures are heavily affected by the prevalence and incidence of disease, and low incidence dooms potential markers with even fairly low false positive rates. The gold standard for the identification of useful biomarkers remains identification of potential biomarkers in one set of individuals followed by validation in a second set.

3. BIOMARKERS IN NEURODEGENERATIVE DISEASES

Different types of biomarkers, including genetic, neuroimaging, clinical, and biochemical markers, are used in the detection of neurodegenerative disease (DeKosky and Marek, 2003).

3.1 Genetic markers

As briefly discussed above, one of the triumphs in modern biology has been the use of molecular genetics to identify gene variations associated with disease. The presence or absence of specific alleles identifies individuals who are at risk of developing a given disease, but generally do not predict age of disease onset accurately. HD is an excellent example.

Although the number of CAG repeats in the Huntington gene correlates with disease onset (Myers, 2004), more markers are needed to provide information about when preclinical manifestations of this disorder will start to happen. A series of studies are underway to identify biomarkers that can detect individuals at risk, at early stages of the disease (Gusella *et al.*, 1986; Paulsen *et al.*, 2001; Djousse *et al.*, 2003, Wexler *et al.*, 2004).

A genetic basis has also been identified for certain cases of ALS, the most common form of motor neuron disease in adults (Rowland and Shneider, 2001). Whereas 90% of ALS cases are sporadic (SALS), 10% are familial (FALS). Mutations in the gene encoding cytosolic copper-zinc superoxide dismutase (*SOD1*) have been robustly identified as causing typical FALS (Rosen *et al.*, 1993). Mutations in two additional genes, *ALS2* and the gene encoding dynactin, have also been reported to cause FALS (Yang *et al.*, 2001; Hadano *et al.*, 2001; Puls *et al.*, 2003). Polymorphisms or variations in other genes have also been considered as possible risk factors for ALS, including *APOE* (Al-Chalabi *et al.*, 1996; Mui *et al.*, 1995) and *ALS2* (Al-Chalabi *et al.*, 2003) and the genes encoding ciliary neurotrophic factor (Orrell *et al.*, 1995; Giess *et al.*, 1998), the astrocytic glutamate transporter EAAT2/GLT1 (Lin *et al.*, 1999), and vascular endothelial growth factor (Lambrechts *et al.*, 2003). These genetic finings could help define a new set of biomarkers for specific subsets of patients.

Likewise, mutations in a number of genes have been identified that correlate with or cause either PD or AD with an autosomal dominant pattern of inheritance (Gasser, 2003; Tanzi and Bertram, 2001; Pankratz *et al.*, 2004). Analysis of the proteins encoded by these genes is starting to give insight into disease mechanisms and could provide valuable markers for subtypes of the diseases. For example, AD-associated mutations in the genes encoding the amyloid precursor protein and presenilin 1 and 2 have thus highlighted amyloid related targets for drug design for this disorder. Similarly, PD-associated mutations in the genes encoding α-synuclein and Parkin have indicated the potential involvement of the ubiquitin-proteasome system in the pathogenesis of PD. Other markers seem to associate with disease but are not predictive markers. An increased risk of developing late onset AD occurs in families that carry the *ApoE4* allele (Corder *et al.*, 1993). Other genes that might predispose individuals to a disease state are being investigated (Pankratz *et al.*, 2003). Genetic markers have yet to be identified in the sporadic, apparently non-familial cases of either AD or PD.

3.2 Neuroimaging biomarkers

Data from neuroimaging studies are starting to emerge as powerful supplements to clinical data in the diagnosis of neurodegenerative diseases.

Imaging tests can be done repeatedly from an early stage of the disease and continued throughout progression of the disease. Functional imaging using single photon emission computerized tomography (SPECT) and positron emission tomography (PET) as well as structural imaging (MRI) have been useful research tools to address early disease changes (Rosas *et al.*, 2004; Brooks, 2004; Kamtarci and Jack, 2004; Jagust, 2004; Snow *et al.*, 1993; Bezard *et al.*, 2001; Niznik *et al.*, 1991; Dekker *et al.*, 2003; Khan *et al.* , 2002; Small *et al.*, 1995; Reiman *et al.*, 1996).

Commonly used technologies include ^{13}C-deoxyglucose PET imaging in Alzheimer's disease, which shows a characteristic pattern of reduced glucose metabolism in the temporo-parietal region. In patients with dementia with Lewy bodies, there is also reduced glucose metabolism in the occipital cortex. Recent studies showed the feasibility of imaging β-amyloid plaques using PET. In Parkinson's disease, dopamine terminals can be evaluated by SPECT using β-CIT, and by fluoro-dopa using PET. In Huntington's disease, there is reduced glucose metabolism, as determined by PET in the basal ganglia even in presymptomatic gene carriers.

Volumetric MRI imaging can be used to assess the size of the hippocampus, and to detect progressive cortical atrophy in AD. In HD, there is progressive loss of volume in the basal ganglia, which can be quantified. In ALS one can detect and quantify progressive damage in the corticospinal tract in the posterior limb of the hippocampus using tensor diffusion MRI. In our hands this has been a sensitive marker of ALS, even in patients who do not show upper motor neuron signs (Finsterbusch *et al.*, 2003; Toosy *et al.*, 2003).

Another valuable imaging technique is NMR spectroscopy. In AD there are reductions of N-acetylaspartate (NAA), a neuronal marker in the hippocampus, which can be quantified. In HD, there are reductions in NAA and increases in lactate in the basal ganglia, which correlate with the length of the CAG expression in the Huntington gene. In ALS, there is a reduction in NAA in the motor cortex.

Eventually it will be of great interest to correlate some of these potential surrogate disease markers with metabolomic measurements. All metabolomic markers will also need to be validated against other clinical assessment scales such as the Unified Parkinson's Disease Rating Scale (UPDRS), the Hamilton Depression Rating Scale (HDRS), Alzheimer Disease Assessment Scale-cognitive subscale (ADCRS), and scales of motor function in ALS.

3.3 Clinical biomarkers

There are a broad range of biomarkers that are used clinically to monitor disease and its progression. These markers range from the loss of a certain function to survival end points. Markers of early stages of disease are much needed. There is controversy around the use of mild cognitive impairment (MCI) as a measure of early AD (Steffenburg *et al.*, 1989; Folstein and Rosen-Sheidley, 2001; Pickles *et al.*, 1995). Some people with MCI do progress to full fledged AD whereas others do not. On average about 15%/year of patients diagnosed with MCI convert to definitive AD. In PD research, very early manifestations of motor dysfunction such as tremor, writing abnormalities, and gait disturbance has been evaluated but do have not proven clinical usefulness as early predictive markers. Loss of olfaction has provided a potential marker for early PD (Cohen *et al.*, 2003; Scheiffele P, 2000). More robust markers are needed.

3.4 Biochemical markers

Extensive research has been aimed at the identification of biochemical markers in blood and CSF for diagnostic purposes. The search for these markers is typically based on research hypotheses and findings related to disease pathology. None of the markers identified to date have the desired sensitivity and specificity. Robust biomarkers for AD are still not available. The introduction of new symptomatic treatments has led to an increased push towards the identification of biochemical markers for early stage AD. Tested biomarkers from plasma and serum include pathophysiologic processes such as amyloid plaque formation, inflammation, oxidative stress, and lipid metabolism, as well as apolipoprotein E changes, and vascular disease markers such as homocysteine (Irizarry, 2004). None of these are robust biomarkers for AD, but they correlate to the condition. None seem to have the needed specificity and sensitivity to predict disease or track responses to therapy. Proteomics approaches seem to provide hope for providing characteristic patterns of biomarkers in individuals with AD. For example, CSF concentrations of total tau, phospho-tau, and the 42 amino acid form of β-amyloid have been evaluated as potential biomarkers for AD (Blennow, 2004). CSF protein biomarkers may have clinical utility in distinguishing AD from normal aging and other CNS disorders.

In ALS, initial symptoms and disease progression varies from patient to patient, making monitoring of clinical trials difficult. Some markers of oxidative stress have been found to be elevated in ALS and Friedriech Ataxia (Bogdanov *et al.*, 2000; Schulz *et al.*, 2000). Surrogate markers are much needed for this disorder and could complement the use of clinical

markers. At the moment clinical endpoints involve voluntary strength evaluation and use of functional rating scales. There are no reliable markers that reflect disease state and its progression and that have the acceptable sensitivity and specificity.

4. METABOLOMICS: A NEW APPROACH FOR IDENTIFYING BIOMARKERS AND THERAPEUTIC TARGETS FOR CNS DISORDERS

4.1 Concepts

Over the last several years, researchers have started to explore the new array based technologies to map biomarkers of disease and identify targets for drug design. These technologies include proteomics, transcriptomics, and most recently metabolomics. The use of automated and high throughput approaches combined with sophisticated mathematical tools promises to provide signatures that are characteristic for each disease state. In this section we highlight the approach of metabolomics in biomarker and target identification and give examples with applications in neurodegenerative diseases.

4.1.1 Metabolomics in the stream of information flow

The "Central Dogma" of molecular biology holds that DNA is transcribed into RNA and RNA is translated into protein. This paradigm and its recognized exceptions – such as the reverse transcription of retroviruses – form a framework for much of modern biology. DNA is the blueprint – the information that provides a description of the potential of a system. RNA serves as a messenger – carrying the currently relevant messages from the blueprint that is DNA to the workers that are the proteins. As such, DNA, RNA, and proteins provide tremendous amounts of information about a biological system and give insight into multiple levels. As such, studies at these levels have provided both biomarkers and risk factors for disease. But these approaches are, in fact, limited.

DNA does not always define destiny. As one example, life span and the incidence of disease can be dominantly and beneficially impacted by caloric restriction. As another example, not all women carrying *BRCA 1* mutation develop breast cancer, and not all people carrying the *APOE4* allele develop AD (e.g., Schrag *et al.*, 2000, Mayeux *et al.*, 1993).

RNA does not always define destiny. It is the material of introductory courses that many genes are regulated post-transcriptionally, and that even considerable up-regulation of mRNA expression is not ubiquitously associated with changes in biological properties.

Protein levels are not ultimate destiny either. High levels can be deceptive, for example, when the proteins are inactive, when they are mislocalized within a cell, or when critical partners or substrates are missing. Although these issues can be addressed, they complicate understanding of the overall picture. Measurements of protein activity might be more useful than measurements of protein concentration, but assays for activity require assumptions about proper conditions. Protein levels are also less responsive to some changes in environment and physiological states. Again, this problem can be circumvented in the case of some signaling proteins, but the caveats and complexity mount as one requires progressively more and more constraints. Thus, neither DNA, nor RNA, nor proteins are, themselves, destiny in all cases. That said, there are clearly cases where they provide sufficient information to act upon. Examples include pre-natal screening by analysis of DNA (Down syndrome, Tay-Sachs disease, sickle-cell anemia) and protein (e.g., assessing neural tube defects through the measurement of alpha-fetoprotein levels), and the wave of cancer diagnostics being developed based on microarray classifiers. Thus, it is important to consider that each form of analysis provides a piece of the puzzle.

Metabolomics is also not destiny – but it is an important piece of the puzzle. The primary human metabolome encoded by the genome may be smaller (perhaps <3000 metabolites), and thus easier to address comprehensively, than either the genome (~35,000 genes), transcriptome (perhaps 100,000 transcripts), or proteome (perhaps 100,000-1,000,000 proteins). Yet added complexity could arise from the small molecules that we get from the environment and from those that are produced by the gut flora. The metabolome is the most responsive to sudden change (potentially both a positive and a negative factor in these studies) and the one that most reflects the current functional status of the organism. As such, the metabolome offers potential for understanding both normal and aberrant physiology.

4.1.2 Metabolomics: Identifying disease signatures

Metabolomics is, essentially, the extension of traditional biochemistry into the -omics domain. As such, the goal is to take a snapshot of metabolism – and thus identify a metabolic signature of an individual or a condition. The four major tools for analyzing metabolites have been NMR, Mass Spectroscopy, HPLC, and GC, with hybrids such as LC-MS and GC-

MS also playing important roles. Each of these tools generates large databases, which can then be parsed using the tools of modern informatics, data mining, computational biology, and chemometrics. As in any biological experiment, signatures are typically compared between different groups of interest, e.g., patients and matched controls. The complexity of the data collected make validation using an external set a key component in these studies.

4.1.3 Metabolomics: Identifying therapeutic targets

Current approaches to modern drug discovery have focused on what has been termed "Target Identification." In other words, drug discovery has been driven by the identification of proteins whose expression appears intrinsic to pathways involved in disease. The development of COX-2 inhibitors serves as a direct example (Flower, 2003). Historically, such linkages have been based on the accumulated wisdom of "the field" or "the company," and have thus been dependent on the synthesis of complex and disparate pieces of information, or years of directed research. Computerized knowledge bases and computational modeling has added powerful tools and predictive components to this picture, but have not changed the overall view. More recently, the advent of 'omics technologies has expanded the ability of researchers to address the physiological status associated with specific diseases and resulted in a new wave of potential targets being identified.

The inherent nature of metabolomics to reflect an organism's current metabolic status, and to link these changes with known metabolic pathways - or demonstrate that such changes are not associated with these pathways - offers an important synergistic complement to the knowledge available from probing other 'omics domains. Metabolomics also offers a means of observing the effects of pharmaceutical interventions.

The study of transgenic mouse models of neurodegenerative diseases and the effects of therapeutic interventions will also help to validate metabolomic profiles. At present, excellent transgenic mouse models exist for AD, HD, and ALS and a number of therapeutic agents show efficacy in these models (see for example, Beal and Ferrante, 2004)

4.2 Issues to consider in studying CNS disorders

The selection of samples for the study of neurodegenerative diseases is an important issue to consider. Brain tissue is complex and heterogeneous and as such, requires that one focus on tissue derived from a specific area of interest. CSF closely represents some aspects of brain metabolism but is very difficult to obtain and is not very practical for the development of

biomarkers. Plasma and urine are easily obtainable but are somewhat distal to brain function and thus one has to sort out what components are relevant to brain metabolism and its dysfunction. It is therefore important to determine whether CSF alterations are reflected in plasma changes. This may not be a critical issue since many neurodegenerative diseases show alterations in peripheral tissues.

One should certainly take into consideration the heterogeneity of the patient population and the effect of therapies on biochemical signatures. A full knowledge of all clinical aspects of each patient sample is valuable as it could enable a correlation between clinical endpoints and biochemical markers. To confirm the specificity of biochemical signatures one should compare signatures of a particular disease with those of other CNS disorders. It is also important to determine whether changes are due to non-specific changes such as muscle breakdown in ALS. An important control in these patients is therefore patients with myopathies. If other organs are involved in the pathogenesis of the disease, the biochemical signature needs to be refined to reflect contributions of these distal organs from that derived from the brain. Longitudinal studies over the course of the disease could be very informative in mapping surrogate markers for the disease. Once a signature for the disease is established one can use such signatures to potentially identify environmental and genetic factors that contribute to disease process.

Biomarkers for neurological diseases should ideally reflect changes early in the disease cycle. Currently, neurological diseases are poorly diagnosed and by the time disease is confirmed a significant number of neurons have already been compromised, making response to therapy difficult. Solid surrogate markers are almost nonexistent with the exception of one imaging test for MS (Bakshi, 2004). In MS the number of new T_2 lesions and lesions which enhance with gadolinium on MRI has been accepted as a reliable surrogate marker. Clinical trials in the absence of robust biomarkers depend on clinical endpoints that are difficult to measure or take a long time to reach (such as survival times). Many trials fail to develop CNS therapies because of the combined lack of good disease markers and the inability to detect the disease in early stages.

4.3 Early findings in studying metabolomics of CNS disorders

We have started to explore fully metabolic perturbations in neurodegenerative diseases. We attempt to understand unique and common perturbed pathways that contribute to the death of neurons in these degenerative diseases. Identifying biomarkers, stratifying patients, and

gaining insights into disease mechanism will accelerate the process of finding more effective therapies.

4.3.1 Metabolomic analysis of motor neuron diseases

Motor neuron diseases (MNDs) are a heterogeneous group of rare disorders that affect motor neurons and cause diverse signs and symptoms. Treatments are mostly non-existent or supportive at best. There is an extension of survival of about 3 months with the drug Riluzole (used to treat patients with ALS). Some MNDs have known causes (e.g., genetic, autoimmune, HIV-related) but for the majority, the cause is unknown.

ALS affects both anterior horn cells and corticospinal tracts (Rowland and Shneider, 2001). In contrast, other categories of MNDs include disorders that selectively affect either the lower or the upper motor neurons (LMNDs and UMNDs respectively) (Rowland and Shneider, 2001). Patients with UMND or primary lateral sclerosis have a much more benign course than patients with ALS and may not become disabled for 10-12 years. Whether ALS is a single disease entity or a syndrome caused by different conditions remains unknown. Additionally, it is not clear whether MNDs are closely related but distinct disorders or whether they represent different points on the spectrum of a single disease. For basic research and clinical trials, strict definitions of ALS and each subset of MND become critical.

We hypothesized that diseases such as ALS might produce characteristic perturbations of the metabolome. Using an HPLC-electrochemical detection platform (EC array – see Chapter 8) we have been able to distinguish ALS patients from controls using three measures of class association: the *t*-statistic, Pearson's correlation coefficient, and the relative class association as well as Partial least squares discriminant analysis (PLS-DA) (Rozen, *in press*). A subsignature for patients with LMND and slow disease progression has been identified. Additionally, a signature for Riluzole that includes endogenously induced metabolites has been defined. These results indicate that metabolomic studies can be used to ascertain metabolic signatures of disease in a non-invasive fashion. Elucidation of the structures of signature molecules in MNDs should provide insight into aberrant biochemical pathways and might provide diagnostic markers and targets for drug design. A program has been initiated to collect samples nationally to dissect motor neuron diseases better. We are defining metabolic signatures for UMNDs, LMNDs, and ALS to determine the relatedness of these diseases at the metabolic level and to define biochemical markers. Although clinicians can classify these diseases in a rough way, we believe that metabolomics can provide a more objective classification.

4.3.2 Metabolomic analysis in other CNS disorders

EC-Array detection systems, such as ESA's CoulArray®, allow targeted and very sensitive (pg) analysis of numerous redox active compounds in complex biological samples. Metabolic signatures for diseases such as PD and AD are starting to unravel. Additionally, a comparison of metabolic profiles from children with cerebral palsy with those of healthy controls resulted in our ability to separate the two groups based on a subset of 62 plasma variables (Kaddurah-Daouk 2004). Similarly, autistic children can be distinguished biochemically from their control counterparts (Kaddurah-Daouk 2004). Further analysis suggests that autistic children are not one class but could be two to three classes based on biochemical variations. Patients with stroke seem to have characteristic biochemical profiles. All this provides hope that metabolomics is going to play a central role in understanding and diagnosing CNS disorders.

4.3.3 Metabolomics in psychiatric disorders

Psychiatric disorders provide very rich areas to explore and include depressive disorders, schizophrenia, and addictive disorders, among others. There is evidence for metabolic disturbances in individuals with learning disabilities and hyper-reactivity syndromes. The metabolomic analysis of these conditions has great promise for the identification of new targets for drug intervention.

5. A PERSONALIZED APPROACH TO CNS THERAPIES

The completion of a high quality sequence of the human genome opened the genomic era of human biology. The human genome project provides valuable information on genes and proteins that control the different facets of cell function. Technological advances in genomics, proteomics, and metabolomics technologies have opened the door for a more personalized approach to medicine. The clinical use of such technologies could (i) provide insights about an individual's predisposition to disease; (ii) allow detection of disease at an early stage; (iii) predict severity of disease; (iv) aid in the selection of the most appropriate therapy; (v) be used to monitor disease progression and outcome; and (vi) facilitate continued optimization of care for each patient. The integration of information from these of the -omics technologies, combined with clinical information and imaging data, is opening new avenues for better clinical care and management.

Metabolomics promises to provide a very valuable niche in this new vision of clinical care.

6. CONCLUSIONS

Metabolomics is emerging as a powerful technology for the development of biochemical signatures for neurodegenerative diseases. These signatures could reveal perturbed biochemical pathways, provide better understanding of disease mechanisms, and potentially identify new targets for drug design.

The early and effective detection CNS disorders will probably depend on the use of several biomarkers in parallel. These markers could include alleles that predispose an individual to a disease as well as markers derived from imaging data and biochemical and clinical studies. Rapid, affordable, and reliable diagnostic markers that depend on easily accessible samples are needed. Surrogate makers that indicate disease course and response to therapy are also much needed and will accelerate clinical trials and drug development.

REFERENCES

Al-Chalabi A *et al.* Variants in the ALS2 gene are not associated with sporadic amyotrophic lateral sclerosis. *Neurogenetics*; 4: 221-222 (2003).

Al-Chalabi A *et al.* Association of apolipoprotein E E4 allele with bulbar-onset motor neuron disease. *Lancet*, 347: 159-160 (1996).

Bogdanov M *et al.* "Increased oxidative damage to DNA in ALS patients." *Free Radic. Biol. Med.*, 29: 652-8 (2000).

Bakshi R, Hutton GJ, Miller JR, Radue EW. The use of magnetic resonance imaging in the diagnosis and long-term management of multiple sclerosis. *Neurology*, 63: S3-11 (2004).

Beal MF, Ferrante RJ. Experimental therapeutics in transgenic mouse models of Huntington's disease, *Nat. Rev. Neurosci.*, 5: 373-84. (2004).

Bezard E, Dovero S, Prunier C, Ravenscroft P, Chalon S, Guilloteau D, Crossman AR, Bioulac B, Brotchie JM, Gross CE. Relationship between the appearance of symptoms and the level of nigrostriatal degeneration in a progressive 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine-lesioned macaque model of Parkinson's disease. *J. Neurosci.*, 21: 6853-61 (2001).

Blennow K.Cerebrospinal fluid protein biomarkers for Alzheimer's disease: *Neuro Rx.*, 1: 213-226 (2004).

Brooks DJ. Neuroimaging in Parkinson's disease: *Neuro Rx.*, 1: 243-254 (2004).

Cohen DR, Matarazzo V, Palmer AM, Tu Y, Jeon OH, Pevsner J, Ronnett GV. Expression of MeCP2 in olfactory receptor neurons is developmentally regulated and occurs before synaptogenesis. *Mol. Cell. Neurosci.*, 22: 417-29 (2003).

Corder EH *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, 261: 921-923 (1993).

de Gruttola VG *et al.* Consideration in the evaluation of surrogate end points in clinical trials. Summary of a national institute of health workshop. *Control.Clin.Trials.*, 22: 485-502 (2001).

Djousse L *et al.* Interaction of normal and expanded CAG repeat sizes influences age at onset of Huntington disease. *Am. J. Med. Gen.*, 119A: 279-82 (2003).

Dekker M, Bonifati V, van Swieten J, Leenders N, Galjaard RJ, Snijders P, Horstink M, Heutink P, Oostra B, van Duijn C. Clinical features and neuroimaging of PARK7-linked parkinsonism. *Movement Disorders*, 18: 751-7 (2003).

DeKosky S,Marek K. Looking backward to move forward: Early detection of neuro digenctive disorders. *Science*, 302: 830-834 (2003).

Finsterbusch J S, Weishaupt JH, Khorram-Sefat D, Frahm J, Ehrenreich H. Diffusion tensor imaging for long-term follow-up of corticospinal tract degeneration in amyotrophic lateral sclerosis. *Neuroradiology*, 45: 598-600 (2003).

Flower RJ. The development of COX2 inhibitors. *Nat. Rev. Drug Discov.*, 2: 179-91 (2003).

Folstein SE. Rosen-Sheidley B. Genetics of autism: complex aetiology for a heterogeneous disorder. *Nat. Rev. Gen.*, 2: 943-55 (2001.)

Frank R, Hargreaves R. Clinical Biomarkers in drug discovery and development. *Nat. Rev. Drug Discov.*, 2: 566-580 (2003).

Gasser T. Overview of the genetics of Parkinsonism. *Adv. Neurology*, 91: 143-52 (2003).

Giess R *et al.* Potential implications of a ciliary neurotrophic factor gene mutation in a German population of patients with motor neuron disease. *Muscle and Nerve*, 21: 236-238 (1998).

Gusella JF, Gilliam TC, Tanzi RE, MacDonald ME, Cheng SV, Wallace M, Haines J, Conneally PM, Wexler NS. Molecular genetics of Huntington's disease: *Cold Spring Harb Symp Quant Biol.*, 51: 359-64 (1986).

Hadano S *et al.* A gene encoding a putative GTPase regulator is mutated in familial amyotrophic lateral sclerosis 2. *Nat. Gen.*, 29: 166-173 (2001).

Harrigan GH, and Goodacre R. *Metabolic profiling: Its role in Biomarker discovery and Gene Function Analysis*; Kluwer Academic Publishers Group, NY, 171-198, 199-216; 311-320 (2003).

Hulka BS. Overview of biological markers. In: *Bilogical markers in epidemiology* (Hlka BS, Griffit GD, Wilcosky TC, eds), 3-15.New York:Oxford university press (1990).

Irizarry MC. Biomarkers of Alzheimer's disease in plasma. *Neuro Rx.*, 1: 226-234 (2004).

Jagust W. Molecular neuroimaging in Alzheimer's disease. *Neuro Rx.*, 1: 206-212 (2004).

Kaddurah-Daouk R, Beecher C, Kristal BS, Matson WR, Bogdanov M, Asa DJ. Bioanalytical advances for metabolomics and metabolic profiling. *PharmaGenomics*, January (2004).

Kamtarci K, Jack CR. Quantitative magnetic resonance techniques as surrogate markers of Alzheimer's disease. *Neuro Rx.*, 1: 196-205 (2004).

Khan NL. Valente EM. Bentivoglio AR. Wood NW. Albanese A. Brooks DJ. Piccini P. Clinical and subclinical dopaminergic dysfunction in PARK6-linked Parkinsonism: an 18F-dopa PET study. *Annals Neurology*. 52: 849-5 (2002).

Lambrechts D *et al.* VEGF is a modifier of amyotrophic lateral sclerosis in mice and humans and protects motoneurons against ischemic death. *Nat. Gen.*, 34: 383-394 (2003).

Le Couteur DG. Muller M. Yang MC. Mellick GD. McLean AJ. Age-environment and gene-environment interactions in the pathogenesis of Parkinson's disease. *Rev. Environ. Health*, 17: 51-64 (2002).

Liddell MB, Lovestone S, Owen MJ. Genetic rich of Alzheimer's disease: Advising relatives *The Brit. J. Psychiatry*, 178: 7-11 (2001).

Lin C *et al.* Aberrant RNA processing in a neurodegenerative disease: the cause for absent EAAT2, a glutamate transporter, in amyotrophic lateral sclerosis. *Neuron*, 20: 589-602 (1998).

Mayeux R. Biomarkers: Potential uses and limitations, *Neuro Rx.*, 1: 182-188 (2004).

Mayeux R. Epidemiology of neurodegeneration. *Annu. Rev. Neurosci.*, 26: 81-104 (2003).

Mayeux R *et al.*, The apolipoprotein epsilon 4 allele in patients with Alzheimer's disease. *Ann. Neurol.*, 34: 752-4 (1993).

Merikangas K. Genetic epidemiology: bringing genetics to the population. The NAPE Lecture 2001. *Acta Psychiatr. Scand.*, 105: 3–13 (2002).

Muller U, Graeber MB. Neurogenetic diseases: molecular diagnosis and therapeutic approaches. *J. Mol. Med.*, 74: 71–84 (1996).

Mui S, Rebeck G, McKenna-Yasek D, Hyman B, Brown R. Apolipoprotein E E4 allele is not associated with earlier age at onset in amyotrophic lateral sclerosis. *Ann. Neurol.*, 38: 460-463 (1995).

Myers RH. Huntington's disease genetics. *Neuro Rx.*, 1: 255-264 (2004).

Niznik HB, Fogel EF, Fassos FF, Seeman P. The dopamine transporter is absent in parkinsonian putamen and reduced in the caudate nucleus. *J. Neurochemistry*, 56: 192-8 (1991).

Orrell R, King A, Lane R, de Belleroche J. Investigation of a null mutation of the CNTF gene in familial amyotrophic lateral sclerosis. *J. Neurol. Sci.*, 132: 126-128 (1995).

Pankratz N *et al.* Significant linkage of Parkinson disease to chromosome 2q36-37. *Am.J.Hum.Genet.*, 72: 1053-1057 (2003).

Paulsen JS, Zhao H, Stout JC, Brinkman RR, Guttman M, Ross CA, Como P, Manning C, Hayden MR. Shoulson I. Huntington Study Group. Clinical markers of early disease in persons near onset of Huntington's disease. *Neurology.*, 57: 658-62 (2001).

Pickles A, Bolton P, Macdonald H, Bailey A, Le Couteur A, Sim CH. Rutter M. Latent-class analysis of recurrence risks for complex phenotypes with selection and measurement error: a twin and family history study of autism. *Am J Hum Gen.*, 57: 717-26 (1995).

Puls I *et al.* Mutant dynactin in motor neuron disease. *Nat. Genet.*, 33: 455-456 (2003).

Reiman E, Caselli R, Yun L. Preclinical evidence of Alzheimer's disease in persons homozygous for the epsilon 4 allele for apolipoprotein E. *N. Engl. J. Med.*, 334: 752-8 (1996).

Rohlff C. Proteomics in neuropsychiatric disorders. *Int. J. Neuropsychopharmacol.*, 4: 93–102 (2001).

Rolan P, Atknson A, Lesko LJ. Use of biomarkers from drug discovery through clinical practice. Report from the 9th European federation of Pharmaceutical sciences conference on optimizing drug development. *Clin. Pharm. Ther.*, 73: 284-291 (2003).

Rowland L, Shneider N. Amyotrophic lateral sclerosis. *N. Eng. J. Med.*, 344: 1688-1700 (2001).

Rosas HD, Feighn AS, Hersch MS. Using advances in neuroimaging to detect, understand, and monitor disease progression in Huntington's disease. *Neuro Rx.*, 1: 263-272 (2004).

Rosen DR, *et al.* Mutations in Cu/Zn superoxide dismutase are associated with familial amyotrophic lateral sclerosis. *Nature*, 362: 59-62 (1993).

Rozen S, Cudkowicz ME, Bogdanov M, Matson WR, Kristal BS, Beecher C, Harrison S, Vouros P, Flarakos, Vigneau-Callahan, Matson TD, Newhall KM, Beal MF, Brown R H, Kaddurah-Daouk R. Metabolomic Analysis and Signatures in Motor Neuron Disease. *Proc. Nat. Acad. Sci.USA (in press).*

Schulz, J. B., *et al.* "Oxidative stress in patients with Friedreich ataxia." *Neurology*, 55: 1719-21 (2000).

Sherer TB. Betarbet R. Greenamyre JT. Environment, mitochondria, and Parkinson's disease. *Neuroscientist*, 8: 192-7 (2002.)

Scheiffele P, Fan J, Choih J, Fetter R, Serafini T. Neuroligin expressed in nonneuronal cells triggers presynaptic development in contacting axons. *Cell*, 101: 657-69 (2000).

Schrag D *et al.* Life expectancy gains from cancer prevention strategies for women with breast cancer and BRCA1 or BRCA2 mutations. *JAMA*, 283: 617-624 (2000)

Small GW *et al.* Apolipoprotein E type 4 allele and cerebral glucose metabolism in relatives at risk for familial Alzheimer disease. *JAMA*, 22-29, 273: 942-7 (1995)

Snow BJ, Tooyama I, McGeer EG, Yamada T, Calne DB, Takahashi H, Kimura H.Human positron emission tomographic [^{18}F]fluorodopa studies correlate with dopamine cell counts and levels. *Ann Neurol.*, 34: 324-30 (1993).

Tanzi RE, Bertram L, New frontiers in Alzheimer's disease genetics. *Neuron*, 32: 181-4 (2001).

Toosy AT, Werring DJ, Orrell RW, Howard RS, King MD, Barker GJ, Miller DH, Thompson AJ. Diffusion tensor imaging detects corticospinal tract involvement at multiple levels in amyotrophic lateral sclerosis. *J. Neurol. Neurosurg. Psychiatry*, 74: 1250-7 (2003).

Tsang F. Soong TW. Interactions between environmental and genetic factors in the pathophysiology of Parkinson's disease. *IUBMB Life*, 55: 323-7 (2003).

Wexler NS *et al.* U.S.-Venezuela Collaborative Research Project. Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. *Proc. Natl. Acad. Sci. USA*, 101: 3498-503 (2004).

Yang Y *et al.* The gene encoding alsin, a protein with three guanine-nucleotide exchange factor domains, is mutated in a form of recessive amyotrophic lateral sclerosis. *Nat. Genet.*, 29: 160-5 (2001).

Conflict of Interest Statement:

BSK has interests including licensing, consulting and equity with Metabolon

Chapter 5

COMPARATIVE METABOLOME PROFILING USING TWO DIMENSIONAL THIN LAYER CHROMATOGRAPHY (2DTLC)

Applications to bacterial metabolomics

Thomas Ferenci and Ram Maharjan
School of Molecular and Microbial Biosciences, University of Sydney G08, N.S.W. 2006, Australia

1. INTRODUCTION

A tremendous amount of DNA sequence information is now available for both pathogenic and non-pathogenic bacteria and provides the foundation for studying how the genome of these organisms functions. To complete the full phenotypic characteristics of bacteria, it is essential to define the metabolite components (the metabolome) as well (Figure 1). Since the first paper on the metabolome of *Escherichia coli* only six years ago (Tweeddale *et al.*, 1998), there is now an exponential growth phase in metabolome research, as occurred earlier with genome, proteome and transcript analyses.

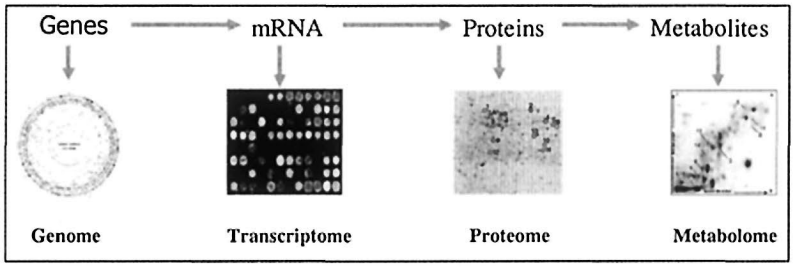


Figure 1. The –omic content of the cell.

In contrast to DNA sequencing, microarrays and proteomes, the lack of a generally adoptable and accessible "global" technology for metabolome analysis is a considerable stumbling block to implementing studies of the metabolome. Ideally, metabolome analysis should allow all of the hundreds of metabolites in a cell to be quantitated under any given condition or in any given strain or mutant. It has been estimated that a bacterial metabolome such as that of *E. coli* comprises 791 metabolites (EcoCyc data base (Karp *et al.*, 2002)). As is also true for proteomes, it is unlikely that every component of the metabolome will be detectable in a single analysis. Most single metabolite assay techniques are not global enough to analyze all cellular metabolites. Also, the concentration of metabolites in a cell differs over a greater than 10^4-fold range, from micromolar (for signal molecules like cAMP (Notley-McRobb *et al.*, 1997)), to above 10 mM for metabolites like glutamate in many bacteria (Tempest *et al.*, 1970). Sensitivity of detection for low abundance metabolites and/or masking by high-abundance metabolites are potential difficulties in analysis of complex mixtures such as proteomes or metabolomes. As a historical parallel, it is instructive that the technique of 2-D acrylamide electrophoresis (Wilkins *et al.*, 1996) took many years to be refined and increased in reproducibility to now provide reproducible proteome profiles. At this stage of its development, the methods discussed in this review do not claim to provide a perfect metabolome. But, as discussed below, the 2D-thin-layer chromatography (2DTLC) approach is still capable of extensive development and improvement whilst already being mature enough to yield novel, biologically significant information. The current aim in 2DTLC is not necessarily to detect every single metabolite, but to maximize the *comparative* monitoring of many areas of metabolism.

Current multiple metabolite assay methods for microbial samples include HPLC separations (Bhattacharya *et al.*, 1995) as well as mass-spectrometry (MS) (Sweetman *et al.*, 1996). GC-MS methods in metabolome analyses were applied to study plant (Fiehn *et al.*, 2000) and *Corynebacterium glutamicum* metabolites (Klapa and Stephanopoulos, 2000). LC-MS is also useful and used for phosphorylated intermediates (Buchholz *et al.*, 2001). Potentially powerful *in situ* nuclear magnetic resonance (NMR) methods have also been used to look at metabolic processes in bacteria (reviewed in Weuster-Botz and de Graaf (1996)). These techniques detect multiple metabolites and, given sophisticated equipment, can provide instantaneous assays on growing intact cultures. However even sophisticated technology such as NMR or capillary electrophoresis resolve only a limited number of metabolites at a time (Dillon and Sears, 1998). Even with the best HPLC or NMR technology, not all classes of metabolites can be included in an analysis. Also, NMR was not sufficient in one recent instance to quantify

individual metabolite levels and enzymatic analysis had to be applied additionally (Raamsdonk *et al.*, 2001). It is impractical to consider NMR techniques in mass screening of knockout mutants, multiple strains or environmental studies where the culture needs to be manipulated. Even the GC-MS approaches typically suffer from the need for chemical derivatization, which excludes some metabolites (Fiehn *et al.*, 2000). As a cost-effective and high-throughput contrast to the above, 2D-TLC can display all extractable metabolites in a single 2-D space, much as proteome studies display proteins after a 2D separation. The practical advantages and limitations of the 2D-thin-layer chromatography (2DTLC) method are discussed in section 4 below.

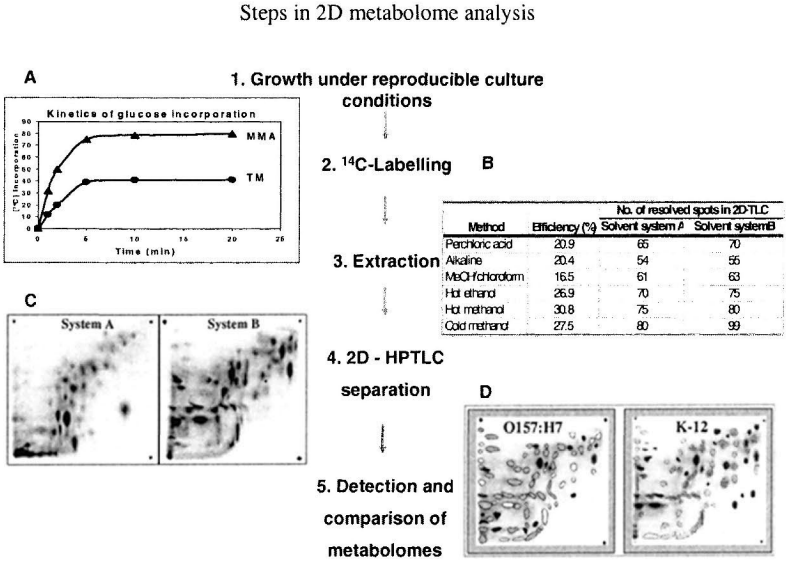


Figure 2. The essential steps in 2D-metabolome analysis using TLC. The numbered steps are illustrated by data illustrating each step. (A) shows glucose incorporation in rich (T.M.) and minimal (MMA) media, (B) shows the efficiency of extractions using various extraction methods in Maharjan and Ferenci (2003), (C) shows elution patterns after 2DTLC with commonly used pairs of solvent systems A and B respectively and (D) shows a Phoretix comparison of patterns between two *E. coli* strains. Differences in spot intensities between the two plates and the absence or presence of spots in one or the other can be seen in (D).

Historically, TLC in all its forms has proved important in chemical and biochemical analysis (Poole and Poole, 1995). Despite its long history, TLC is still an evolving science (Poole, 2003) and its simplicity and cheapness makes TLC an attractive methodology for metabolomics. In view of the multi-dimensional development possibilities in TLC separations, we consider 2DTLC in comparative studies of metabolomes.

2. THE METHODOLOGY OF METABOLOME PROFILING BY 2DTLC

As described in Figure 2, the 2DTLC approach involves five steps.

1. Bacterial growth under reproducible conditions, usually with chemostats and minimal salts media.
2. Labeling metabolites during growth using ^{14}C-glucose.
3. Extraction of (labeled) metabolites from cells.
4. Thin layer chromatography (TLC) in two dimensions using selected pairs of solvent systems
5. Detection of labeled metabolites on TLC plates, spot quantitation and differential comparisons between controls and stressed bacteria.

Each of these steps is critical to the overall methodology and all would still benefit from further advances, as discussed below.

2.1 Culture conditions

The influence of growth rate on metabolite pools is very evident and some metabolites are seen only in slow-growing bacteria (Tweeddale *et al.*, 1998). Metabolome analyses must therefore take growth conditions into account. For comparisons of different strains or mutants, chemostat cultures are more stable and better controlled for growth rate (Jannasch and Egli, 1993). However, it is technically limiting to set up a large number of complicated continuous cultures. Simple batch cultures still allow reproducible metabolome profiles as long as repeatable growth to early exponential phase is achieved (Tweeddale *et al.*, 1998). The effect of genotypic differences on growth rate presents a more serious problem for comparisons. Many natural isolates of *E. coli* vary in their growth rates in lab media (Wada *et al.*, 2000), so comparisons need matching strains.

2.2 Metabolite labeling conditions

The labeling in Figure 2 involved incorporation of ^{14}C -label into cellular metabolites, supplied in (U) ^{14}C -glucose over a period dependent on growth rate during steady-state growth. Combined with the high sensitivity of the phosphorimager detection (section 2.5 below), and the separations shown in Figure 2D, this method permits detection of about 15% of all supposed database metabolites of *E. coli* (Karp *et al.*, 2002). Of course, many metabolites in the database are intermediates in pathways not induced under particular growth conditions. Low-abundance metabolites (we estimated to be below approx. 0.5 mM in the cell) were not identifiable using the original method. For example, several glycolytic phosphorylated intermediates are below this concentration in the cell (Schaefer *et al.*, 1999). The sensitivity of the approach can nevertheless be improved in a simple but expensive way, by using higher specific activity glucose in the labeling. In all experiments so far, we have not used the highest available activity of ^{14}C label. The time course of glucose incorporation is also a useful indicator of the behaviour of each comparison strain. Sampling times are standardized, so that similar, saturating levels of label are incorporated in each strain, and metabolome differences are not the result of, for example, mutational or strain-specific differences in glucose uptake.

A possible avenue for expanding 2DTLC data interpretation is to include the use of isotopes other than ^{14}C in metabolome experiments. Alternative isotopes would be helpful in providing a short-cut to the identification of spots in metabolome maps. By superimposing ^{14}C spot patterns with those labeled with ^{32}P or ^{35}S, it should be possible to pinpoint subsets of phosphorylated or S-containing metabolites as was done by Bochner and Ames (1982) for nucleotides. This approach may yield useful data not only in identification, but also if these subsets of metabolites are differentially affected in strain comparisons.

2.3 Culture extraction

The first reported metabolome study with *E. coli* used stable but simple extraction conditions. After steady-state labeling, bacteria were harvested by centrifugation and boiling of pelleted cells in the presence of ethanol to extract metabolites, which gave reproducible results (Tweeddale *et al.*, 1998, 1999, Liu *et al.*, 2000). This method did not instantaneously stop metabolism and is open to the criticism of possible turnover during the centrifugation step. This limitation is one reason why our current methodology has limited usefulness in the kinetic analysis of metabolism. To overcome problems with extraction was one of the reasons why we investigated alternative lysis

methods (Figure 2.2, (Maharjan and Ferenci, 2003)). The stoppage of metabolism by rapid quenching with methanol-dry ice at –40°C has been used for determinations of pool kinetics (de Koning and van Dam, 1992, Gonzalez *et al.*, 1997, Buchholz *et al.*, 2001). We need to assess the advantages of simplicity (as with the current method) in relation to the advantage of having an instantaneous profile.

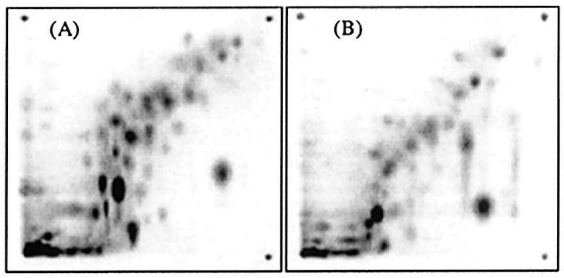
Any sample preparation protocol must necessarily remain a compromise between several factors including rapidity, complete recovery of different compound classes and avoidance of chemical or physical breakdown of more labile metabolites. The study by Maharjan and Ferenci (2003) has already led to the awareness that extraction with cold methanol results in 20% more spots on 2D plates than the traditional hot ethanol extraction used previously (Tweeddale *et al.*, 1998) or with acid extraction (Figure 2). These kinds of technical advances, as well as those below, are expected to increase the numbers of spots analysed and hence the global reach of the analysis.

We are aware that highly lipophilic metabolites are not represented in the standard metabolome if we use polar solvents for extraction. For example, chloroform-methanol extraction was originally designed for the extraction of non-polar metabolites (Folch *et al.*, 1957). We can also investigate whether the increased range of metabolites covered is worthwhile, given the cost of doubling the analysis load and data to be compared. As noted from Figure 2, it is already evident that chloroform-methanol will not give a comprehensive global coverage (2DTLC studies (Maharjan and Ferenci, 2003) show a lack of polar, phosphorylated compounds) so it would have to be used in addition to polar extractions. Nevertheless, it would be worthwhile to include lipophilic analysis to test the extent of variation, especially with a variety of organisms. One reason for this is the documented change in membrane lipids such as cyclopropane fatty acids in *Mycobacterium tuberculosis* pathogenesis and in *E. coli* in response to stress (Cronan, 2002).

2.4 TLC methods

In Figure 2D, analysis involved 2-dimensional separation of extracted ^{14}C -glucose-labelled metabolites on standard Merck silica gel 60 TLC plates. The primary advantage of this form of analysis is that *all water-soluble carbon-containing compounds of a cell are applied and present in the 2-D space* to be analysed. The amount of radioactivity loaded can be standardized and the proportion of label in each spot estimated as a proportion of the total metabolome. This advantage is not available in analyses relying on several separate types of separation and quantitation, where comparison between extracts is more complicated and absolute quantitation requires added internal standards. Significantly, another

advantage over piecemeal HPLC and GC-MS strategies is that the 2-D approach permits use of parallel technology to proteome studies, such that the graphical analysis software developed to match spots in 2-dimensional arrays is also usable in metabolome studies (see Section 2.5 below).



Strain	System A		System B	
	TLC	HPTLC	TLC	HPTLC
K-12	74 ± 5	84 ± 2	94 ± 2	99 ± 3
M534 (0157:H7)	75 ± 3	93 ± 3	91 ± 3	100 ± 4

Figure 3. Metabolome of *E. coli* (K-12) growing on glucose. Extracts were obtained from chemostat growing strain BW2952 at dilution rate of 0.1 h-1 with limiting glucose (0.02% wt/vol) at 37°C. ^{14}C-labeled extracts were separated on HPTLC, Merck silica gel-60, (A) and TLC Merck silica gel-60 (B) using solvent pair A. The table shows the counted number of resolved spots on HPTLC and TLC after separation with pairs of solvent systems A and B.

Recently, the significance of the thin layer medium was also investigated. An improvement in reducing spot diameters, reducing spot overlap, concentrating weak, diffuse spots and hence increasing the number of individual spots resolved was achieved using "high resolution" high-performance TLC (HPTLC) systems available from Merck. The results in Figure 3 compare TLC and HPTLC results on identical extracts with two *E. coli* strains. As seen both visually and by spot counting using Phoretix software, the HPTLC resulted in improved resolution with either of the commonly used elution methods A or B (Tweeddale *et al.*, 1998; R. Maharjan, unpublished results).

The proportion of label unresolved at the origin in our standard TLC plates represents an appreciable proportion of potential metabolites (up to 40%, depending on solvents used for development (Tweeddale *et al.*, 1998)). Optimization of TLC media and solvent properties could also reduce this unanalyzed portion of the metabolome, although the physical properties of

particular metabolites may limit what elutes from the origin. It should also be noted that the extraction method is partly responsible for the origin-bound nature of some material; the cold methanol extraction shown in Figure 2 has an added advantage in that it results in lower (<10%) of the label sticking near the origin (Maharjan and Ferenci, 2003).

Further advances could be made in another aspect of methodology, in the chromatographic development process of the TLC itself. As discussed in (Tweeddale *et al.*, 1998), no single elution system is ideal for resolving all metabolites and we initially adopted up to 4 different solvent systems for separating metabolites in each sample. Approximately 60-90% of the total metabolome label is resolved in each development system (i.e. is not left at the origin or a smear close to the origin) and some major metabolites were better resolved in some elution conditions than others. The metabolites in the unseparated region near the origin in solvent system B such as sugar phosphates and polyamines were better resolved in solvent mix A. On the other hand, amino acids (other than basic ones) and sugars were better resolved in system B. The other two remaining elution systems were used to confirm identities and quantities of metabolites separated. However, the 2-dimensional separation of metabolite spots suffers from the same potential drawbacks as proteome analyses, with overlaps, smears and irregular spots difficult to quantify.

As noted above, highly lipophilic metabolites are not represented in the standard metabolome; if lipids are also to be considered in the analysis, different TLC elution solvents used in lipid analysis will be required here as well. The increased range of metabolites covered is of course at the cost of doubling the analysis load and data to be compared.

2.5 Detection of labelled metabolites on TLC plates, spot quantitation and differential comparisons between controls and stressed bacteria

High-sensitivity phosphorimaging is used for the detection of ^{14}C label in metabolite spots, which has the additional advantage of the availability of Molecular Dynamics ImageQuant software to quantitate the amount of each spot. Each metabolite is calculated as a proportion of total metabolites (i.e. total pixels detected on the TLC plate), as used in recent metabolome publications (Tweeddale *et al.*, 1998, 1999; Liu *et al.*, 2000). Quantitation in this way avoids problems of standardizing different samples with different loadings on separate plates. Phosphorimaging data files are also directly accessible to analysis with Phoretix software, used in proteome analysis (Mahon and Dupree, 2001) for spot-matching and comparison of multiple

changes in intensities of spots under different growth conditions. The data in Figure 2D provide an example of "proteome" software to compare metabolome 2-D maps and visualize differences as a result of strain differences. The same approach can of course be used to compare methodology improvements, stress effects or mutant comparisons for functional genomics.

The next stage of analysis, spot identification, is distinctly deficient in our current procedure. Each of the dozen or so metabolites quantified in (Tweeddale *et al.*, 1998, 1999; Liu *et al.*, 2000) was resolved and identified (by comigration with standards) in at least 2 elution systems. This simple approach is fine for a few individual metabolites but there is a definite need to be able to identify all unknown spots. Converting 2DTLC into a mature metabolome analysis requires identification of ALL the spots on plates as in Figure 2D. The most obvious way of identifying resolved spots is through mass-spectrometry (Poole, 2003) and a combined TLC-MS approach remains an option.

Sensitive chemical profiling of amino acid content of metabolomes is an easy way of cross-checking whether the ratio of radioactive amino acid spots quantitated in ImageQuant is not due to biased ^{14}C incorporation into particular branches of metabolism and has proved useful in confirming results evident from 2DTLC. For example, the presence of elevated valine levels in metabolome profiling by the 2-D analysis under oxidative stress was verified by HPLC amino acid analysis (Tweeddale *et al.*, 1999). Indeed in all published studies, the amino acid pool changes in 2DTLC were consistent with HPLC amino acid analyses on identical samples (Tweeddale *et al.*, 1998, 1999). The consistency of results between 2DTLC and chemical analysis was good evidence that the 2DTLC detects real pool shifts. Likewise for sugars, the trends for trehalose reproduce previous findings with other analytical techniques (Tweeddale *et al.*, 1998, 1999; Liu *et al.*, 2000).

The overall reproducibility of the metabolome patterns in 2DTLC is potentially affected by many of the previously discussed experimental factors. Nevertheless, the 2DTLC technique is robust and the data obtained (Tweeddale *et al.*, 1998, 1999; Liu *et al.*, 2000), suggests the patterns and ImageQuant comparisons are reproducible in the three or more cultures and extracts of each condition analysed, with each sample subject to several TLC separation runs. Reproducibility was also found in the study on extraction optimization (Figure 2). So the methods used for later examples are already functional enough to make statements about metabolome shifts between strains and stresses.

3. APPLICATIONS TO THE FUNCTIONAL GENOMICS OF BACTERIA

The characteristics of the metabolome, as with all cellular properties, are determined by a combination of genetic potential and environmental signals. Examples of metabolome changes resulting from both genomic and environmental differences are now available and reveal the potential of 2DTLC analysis in studying global metabolic differences.

3.1 Stress effects on the metabolome

Stress effects on gene expression and cellular protein patterns in bacteria (Storz and Hengge-Aronis, 2000) are far better appreciated than environmental influences on metabolism. Identifying the extent of metabolic changes in response to stress would be of great benefit to fermentation technology and metabolic engineering. To investigate metabolome changes in stressed *E. coli* cells, bacteria were subject to six different types of stress, as described in Table 1.

The stresses applied were heat (42°C instead of 37°C), oxidation stress (paraquat, t-butylhydroperoxide or hydrogen peroxide), heavy metal (copper sulfate), ethanol, osmotic stress (high sucrose) and acid stress (with the medium buffered at pH 6 instead of pH 7).

Table 1. Metabolite pool size (% of total metabolome)[a] of *E. coli* growing in the presence of indicated stress stimulants as determined by metabolome analysis.

Metabolites	Stress stimulants						
	Control	42°C	Paraquat[b]	$CuSO_4$	Ethanol	Sucrose	pH6
Glutamate	20	8	12	**2**	11	**3**	9
Trehalose	<0.05	<0.05	<0.05	<0.05	<0.05	**11**	<0.05
Glucose	<0.05	<0.05	<0.05	<0.05	<0.05	**1.4**	<0.05
UDP-Glc/Gal	1.1	1.1	1.26	1.95	2.05	0.95	0.95
Aspartate	0.35	**0.05**	0.42	0.4	0.39	0.1	**0.75**
Lysine	0.4	0.2	0.4	**1.45**	**1.05**	0.4	0.5
UDP-GlcNAc	1.5	1.05	1.4	0.95	1.9	**0.4**	1.05
Glutathione	1.25	0.95	**2.5**	1.05	1.75	**0.05**	1.8
Putrescine	1.8	1.1	1.5	1.5	0.9	**0.25**	**3.6**
Valine	0.05	0.05	**2.26**	0.05	0.1	0.1	0.25

[a] The metabolite pool size as a percentage of the total ^{14}C-labelled metabolite pool in the extract was calculated as previously described (Tweeddale *et al.*, 1998). [b] The data are adopted from reference (Tweeddale *et al.*, 1999). The values with bold numbers are significantly up- or down-regulated as an effect of stress.

The metabolome 2DTLC patterns were obtained with extracts of bacteria grown at steady state in continuous culture with a dilution rate of 0.6 h^{-1} (doubling time of approx 70 min). This was done so the changes observed were not the result of changes in growth rate (see Section 2.1 above).

The smallest qualitative differences were between bacteria growing at 37°C and 42°C, but even here there were quantitative shifts in several spots, including a marked decrease in aspartate. The major metabolite within control bacteria at D=0.6 h^{-1} was glutamate (~20% of total metabolome), as was expected from earlier studies (Tempest *et al.*, 1970). In some stress states however, the glutamate spot was reduced 10-fold, as in bacteria growing in high osmolarity or in the presence of copper sulfate. Trehalose replaced glutamate under steady-state high osmolarity conditions as the major metabolite (~10% of total metabolome) but not under other stresses. Also consistent with earlier findings, the level of the polyamine putrescine was greatly reduced with osmotic stress (Tkachenko *et al.*, 1997). Also expected was the elevated level of glutathione under conditions of oxidative stress, particularly paraquat, consistent with a role of glutathione in minimising oxidative damage (Penninckx and Elskens, 1993). Less predictably, glutathione pools were strongly reduced during growth with high osmolarity. Although this was possibly a consequence of lesser precursor (glutamate) availability, there may be a connection between osmoregulation and glutathione levels in that glutathione controls some K^+ efflux systems (Elmore *et al.*, 1990).

The most striking change in the presence of paraquat was the elevation of valine pools, with valine becoming one of the more abundant metabolites, which suggested a role for the valine pool in mopping up reactive oxygen radicals (Tweeddale *et al.*, 1999).

Other less predictable changes were evident with various stresses. For example, glucose increased with high osmolarity, adenosine increased in ethanol (but with a large variation in different cultures) and putrescine increased at pH 6. With copper sulfate or ethanol, there was a significant increase in the lysine pool, which was not evident with other stresses. A large number of other metabolites changed significantly under one or more stresses. For example, the oxidative stress samples (paraquat, hydrogen peroxide and t-butylhydroperoxide) as well as ethanol had a greater proportion of the metabolome in the unresolved region in Figure 3a to the right of the origin which contains a large proportion of negatively-charged compounds.

The results in Table 1 illustrate the scale of changes in the environmental control of metabolism and how little we really know about this. Metabolome research of this kind is of direct relevance to industrial fermentation processes, all of which apply some stress to cultures. Global analysis offers a

means of optimising nutrient throughput and identifies targets for elimination of unproductive metabolism such as trehalose accumulation if growth yield is to be maximised.

3.2 Mutational changes in 2DTLC profiles – gene - metabolome interactions

A common approach to bacterial functional genomics is the use of gene-disruption mutants to identify bacterial orphan genes (Kobayashi *et al.*, 2003; Mori *et al.*, 2000). At least 30% of genes in most genomes are unidentified in function. In principle, metabolomics can be used to identify genes with roles in metabolism in any organism (Fiehn, 2002; Raamsdonk *et al.*, 2001). To what extent are genomic changes resulting from mutation interpretable in bacterial metabolism? Flux measurements and metabolic outcomes have been followed in *E. coli* with individual mutations in genes such as *pykA* (Emmerling *et al.*, 2002) *pgi*, *gnd* and *zwf* (Hua *et al.*, 2003; Zhao *et al.*, 2004) as well as other genes (Fischer and Sauer, 2003), but the global metabolome changes are not known. How useful is the 2DTLC approach to this question of interpreting gene knockouts?

To probe the usefulness of 2DTLC analysis, metabolome comparisons have been made with model disruption mutants of *E. coli*. The examples shown in Figure 4 involve a blockage of the tricarboxylic acid cycle before succinate; mutants lacking *sucC* are still able to grow on glucose but do not fully oxidise it. Another source of succinate in cells is via the glyoxylate cycle, and this was also inactivated through an *aceA* mutation (Figure 4).

The global patterns in these mutants are certainly changed compared to wild-type bacteria, but in complex ways. The intuitive result that succinate levels would drop significantly due to the *sucC* mutation was not found; the proportion of succinate in the global metabolome slightly increased from 3.8% in wild-type to 4.1% in the mutant. Most likely, compensatory changes in metabolism removed the impact on TCA cycle intermediates, consistent with the elasticity of metabolic processes (Edwards and Palsson, 2000). However, in asking whether metabolome assays can reveal a direct hint as to the site of the blockage, the answer was not very encouraging. Also unexpectedly, the block in the glyoxylate cycle produced a greater drop in succinate pool size, to 1.4%. Looking at the results of Figure 4, it is evident that complex flux analysis and modelling would be required to interpret such complex data comparisons (Price *et al.*, 2003). Comparative metabolomics does not provide a simple means of discerning gene-metabolome links.

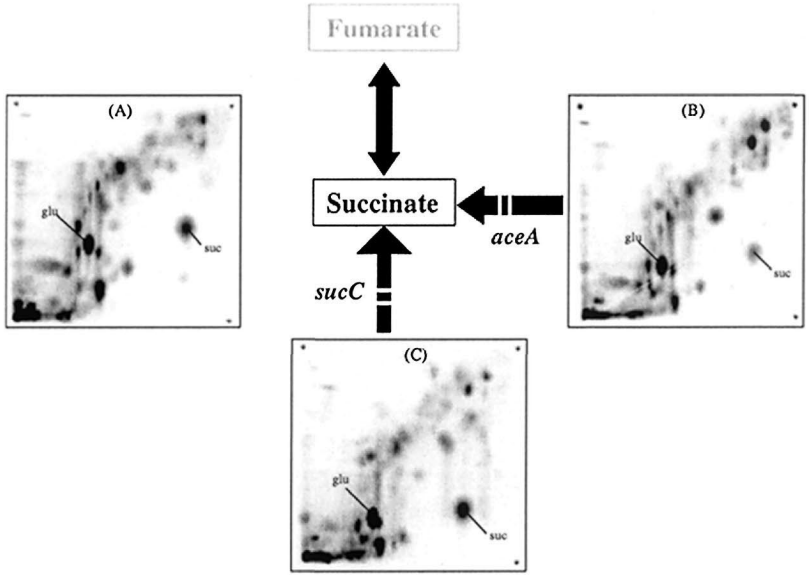


Figure 4. Mutation effect on the metabolome of *E. coli* K12. 2DTLC metabolome comparisons of *E. coli* K12 gene knockout mutants grown under identical conditions with strains (A); BW2952 (wild-type), (B); BW2952 aceA and (C); BW2952 sucC. The extracts collected after CM extraction were separated in TLC solvent system A. The spots glu and suc represent glutamate and succinate respectively.

3.3 Bacterial taxonomy through metabolome comparisons

The above results suggested that there was a disappointingly complex relationship between the effect of a single gene disruption and metabolome patterns. This of course raises the interesting question of whether there is any obvious informational link between the metabolome and the genome. Another way to approach this question is to ask whether there are discernible links between metabolome patterns and phylogenetic groupings, and whether metabolome relationships are the same as genomic relationships between different isolates of the same species. *E. coli* isolates were used to test whether metabolome data reveals the same linkages as genomic relationships between bacteria. The species *E. coli* is relatively diverse and the genetic phylogenetic tree for many isolates is already available (Pupo *et al.*, 1997).

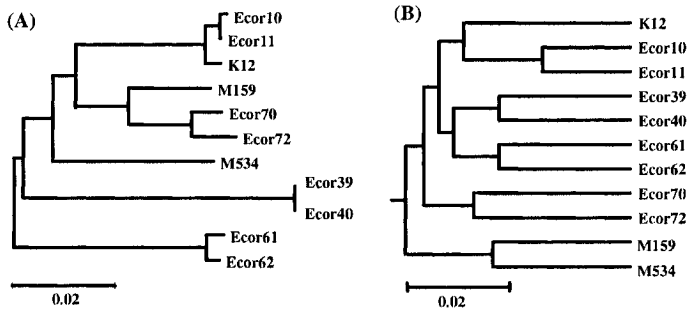


Figure 5. Genomic and metabolomic relationship in *E. coli* isolates. The tree based on metabolome data (panel B) was generated, using MLEECOMP program (Pupo et al., 2000) from the Australian Genomic Information Service. Scoring was based on the presence, absence and quantities of ^{14}C-labeled metabolite spots after separation by 2D-HPTLC. The tree generated on the basis of genomic similarity (A) used *purA, adk, gyrB, rcd, fumC, recA* and *mdh* gene sequences in strains ECOR10, ECOR39, ECOR61, ECOR70 and M159 (all commensal strains) and ECOR 11, ECOR40, ECOR62 and ECOR72 (urinary tract infectious strains). Strain M534 is a pathogenic EHEC (0157:H7) strain (Pupo et al., 2000).

Ten isolates from various locations and the laboratory K-12 strain of *E. coli* were compared through metabolome analysis using HPTLC and methanol extraction (Miller and Maharjan, in preparation). Four of the isolates had genetic elements associated with urinary tract infections (Lai *et al.*, 1999), one was associated with enterohaemorrhagic disease and five were commensal strains from diverse sources (the ECOR collection, (Selander *et al.*, 1987). The plan was to test if metabolism is also associated with a disease lifestyle, or more related to genetic similarity between the subtypes of *E. coli*.

Indeed, differences in the metabolome patterns of the ten strains were evident. Differences were seen in the presence and absence of spots as well as the quantity of individual spots present in more than one isolate (Miller and Maharjan, unpublished results). Firstly, dendrograms based solely on the presence and absences of spots were produced, ignoring quantity differences. An improvement in data analysis was evident when the presence and absence of spots and the quantities of spots were both included to produce dendrograms (using program MLEECOMP, (Pupo *et al.*, 2000)). Furthermore, the spot information from the same extract but separated in solvent systems A and B were combined. The metabolome relationship tree shown in Figure 5B was based on the combined data.

Most interestingly, the tree based on metabolome data resembles closely the tree in Figure 5A based on genetic sequence. However, a difference between the genetic and metabolic trees is in the position of isolates M159

and M534. In metabolomic trees, a major division occurs between these so-called group B1 isolates and the other groups, because the metabolome patterns of these B1 strains differed most to the other strains. This differing metabolism is of interest as B1 strains have also been shown to be able to utilize a wider range of sugars (Pupo *et al.*, 2000; Souza *et al.*, 1999).

Furthermore, the pathogenic strains were not shown to group together, however pair members of taxonomic groups chosen on the basis of genetic similarity between a pathogen and commensal, did group together. The tree branch distance joining the pathogen and commensal is longer than the distance between the groups. This suggests that there are relatively few spots in common in the metabolomes of all strains and no metabolism additional and common in all pathogens. Still, the pathogenic strains had spots in the metabolome not present in taxonomically close commensal bacteria. The finding of pathogen-specific spots could lead to the identification of metabolism unique to pathogens. In the future such metabolome approaches could result in a deeper understanding of pathogen-specific cell function and provide leads for antibiotic studies.

The finding in dendrograms, that metabolism is more closely linked to phylogenetic grouping than pathotype, shows the possibility that metabolome data could have applications in taxonomic groupings. However, the long branch distances to the first node in the metabolome trees are longer than the branch lengths between clusters, and isolates differ more in metabolism than they do genetically. This shows that there are many differences between the strains within clusters consisting of pathogenic and non-pathogenic isolates. These results complement analyses of UTI strains using ultra-violet resonance Raman spectroscopy reflecting metabolite contents of bacteria (Jarvis and Goodacre, 2004). They also strengthen the case for the development of taxonomic approaches based on metabolite analyses of bacteria (Goodacre *et al.*, 2004).

4. CONCLUSIONS- ADVANTAGES AND LIMITATIONS OF THE 2DTLC APPROACH

In the work we described, our strategy was not *comprehensive* or descriptive of every metabolite in the cell. Rather, the examples we gave highlight a *comparative* approach that has many applications in functional genomics. A further advantage of our approach is that the procedure can be used for any cell type that can be grown under reproducible conditions. Although not generally applied so far, there is certainly scope for 2DTLC approaches to eukaryotic microorganisms like yeast or different cells amenable to cell culture.

In the future, there will be a place for rapid comparative metabolomics as well as the comprehensive analysis. Technically, the most beneficial future development would be the combination of 2DTLC with instant *in situ* mass-spectrometry for identification of all the spots. The technology for TLC-MS is being developed (Poole, 2003) and this will be a realistic metabolome approach in the future.

The significance of making metabolome analysis global and easily accessible is that it will enable many fundamental questions in biology to be answered in the future. For example, 2DTLC metabolome analysis can be used in numerous applications, such as to permit the channeling of metabolism to specific products in metabolic engineering (Koffas *et al.*, 1999) by metabolic screening of metabolomes in strains under development. Another microbial application will be to identify new or characteristic metabolites in the search for new drugs or targets of antibiotic action. If new compounds unique to pathogens will be indeed identified through metabolome analysis, we would have far more fundamental information to be used to plot antibacterial strategies. The data in Figure 5 shows how metabolomics can also reveal taxonomic properties of organisms. But finally, the biggest fundamental application of 2DTLC will be in understanding cellular physiology and metabolism in organisms. The examples we described illustrate the potential of this approach.

ACKNOWLEDGEMENTS

We thank our fellow lab members, past and present, for their invaluable contributions to the studies we discussed.

REFERENCES

Bhattacharya M, Fuhrman L, Ingram A, Nickerson KW and Conway T. Single-run separation and detection of multiple metabolic intermediates by anion-exchange high-performance liquid chromatography and application to cell pool extracts prepared from *Escherichia coli*. *Anal. Biochem.*, 232: 98-106 (1995).

Bochner BR and Ames BN. Complete analysis of cellular nucleotides by two-dimensional thin layer chromatography. *J. Biol. Chem.*, 257: 9759-9769 (1982).

Buchholz A, Takors R and Wandrey C. Quantification of intracellular metabolites in *Escherichia coli* K12 using liquid chromatographic-electrospray ionization tandem mass spectrometric techniques. *Anal. Biochem.*, 295: 129-137 (2001).

Cronan JE. Phospholipid modifications in bacteria. *Curr. Opinion Microbiol.*, 5: 202-205 (2002).

de Koning W and van Dam K. A method for the determination of changes of glycolytic metabolites in yeast on a subsecond time scale using extraction at neutral pH. *Anal. Biochem.*, 204: 118-23 (1992).

Dillon PF and Sears PR. Capillary electrophoretic measurement of tissue metabolites. *Am. J. Physiol. - Cell Physiol.*, 274: C840-C845 (1998).

Edwards JS and Palsson BO. Robustness analysis of the *Escherichia coli* metabolic network. *Biotech. Prog.*, 16: 927-939 (2000).

Elmore MJ, Lamb AJ, Ritchie GY, Douglas RM, Munro A, Gajewska A and Booth IR. Activation of potassium efflux from *Escherichia coli* by glutathione metabolites. *Mol. Microbiol.*, 4: 405-412 (1990).

Emmerling M, Dauner M, Ponti A, Fiaux J, Hochuli M, Szyperski T, Wuthrich K, Bailey JE and SauerU. Metabolic flux responses to pyruvate kinase knockout in *Escherichia coli*. *J. Bacteriol.*, 184: 152-164 (2002).

Fiehn O. Metabolomics - the link between genotypes and phenotypes. *Plant Mol. Biol.*, 48: 155-171 (2002).

Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN and Willmitzer L. Metabolite profiling for plant functional genomics. *Nat. Biotechnol.*, 18: 1157-1161 (2000).

Fischer E and Sauer U. Metabolic flux profiling of *Escherichia coli* mutants in central carbon metabolism using GC-MS. *Eur. J. Biochem.*, 270: 880-891 (2003).

Folch J, Lees M. and Stanley SGH. A simple method for the isolation and purification of total lipids from animal tissues. *Biol. Chem.*, 226: 497-509 (1957).

Gonzalez B, Francois J and Renaud M. A rapid and reliable method for metabolite extraction in yeast using boiling buffered ethanol. *Yeast*, 13: 1347-1355 (1997).

Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG and Kell DB. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends in Biotechnology* 22: 245-252 (2004).

Hua Q, Yang C, Baba T, Mori H and Shimizu K. Responses of the central metabolism in *Escherichia coli* to phosphoglucose isomerase and glucose-6-phosphate dehydrogenase knockouts. *J. Bacteriol.*, 185: 7053-7067 (2003).

Jannasch HW and Egli T. Microbial growth kinetics: a historical perspective. *Ant. Van Leeuwenhoek Int J. Gen.Mol. Microbiol.*, 63: 213-224 (1993).

Jarvis RM and Goodacre R. Ultra-violet resonance Raman spectroscopy for the rapid discrimination of urinary tract infection bacteria. *FEMS Microbiology Letters*, 232: 127-132 (2004).

Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C and Gama-Castro S. The EcoCyc database. *Nuc. Acids Res.*, 30: 56-58 (2002).

Klapa MI and Stephanopoulos G. Observability and redundancy analysis of complex metabolic networks. *Faseb J.*, 14: A1313-A1313 (2000).

Kobayashi K *et al.* Essential *Bacillus subtilis* genes. *Proc.Nat. Acad. Sc. USA*, 100: 4678-4683 (2003).

Koffas M, Roberge C, Lee K and Stephanopoulos G. Metabolic engineering. *Ann. Rev. Biomed. Eng.*, 1: 535-557 (1999).

Lai X, Wang S and Uhlin B. Expression of cytotoxicity by potential pathogens in the standard *Escherichia coli* collection of reference (ECOR) strains. *Microbiol.*, 145: 3295-3303 (1999).

Liu XQ, Ng C and Ferenci T. Global adaptations resulting from high population densities in *Escherichia coli* cultures. *J. Bacteriol.*, 182: 4158-4164 (2000).

Maharjan RP and Ferenci T. Global metabolite analysis: the influence of extraction methodology on metabolome profiles of *Escherichia coli*. *Anal. Biochem.*, 313: 145-154 (2003).

Mahon P and Dupree P. Quantitative and reproducible two-dimensional gel analysis using Phoretix 2D Full. *Electrophor.*, 22: 2075-2085 (2001).

Mori H, Isono K, Horiuchi T and Miki T. Functional genomics of *Escherichia coli* in Japan. *Res. Microbiol.*, 151: 121-128 (2000).

Notley-McRobb L, Death A and Ferenci T. The relationship between external glucose concentration and cAMP levels inside *Escherichia coli* - implications for models of phosphotransferase-mediated regulation of adenylate cyclase. *Microbiol.*, 143: 1909-1918 (1997).

Penninckx MJ. and Elskens MT. Metabolism and functions of glutathione in micro-organisms. *Adv.Microb. Physiol.*, 34: 239-301 (1993).

Poole CF. Thin-layer chromatography: Challenges and opportunities. *J. Chromat. A*, 1000: 963-984 (2003).

Poole CF and Poole SK. Multidimensionality in planar chromatography. *J. Chromat. A*, 703: 573-612 (1995).

Price ND, Papin JA, Schilling CH and Palsson BO. Genome-scale microbial *in silico* models: the constraints-based approach. *Trends Biotechnol.*, 21: 162-169 (2003).

Pupo GM, Karaolis DK, Lan RT and Reeves PR. Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and mdh sequence studies. *Inf. Imm.*, 65: 2685-2692 (1997).

Pupo GM, Lan RT, Reeves PR and Baverstock PR. Population genetics of *Escherichia coli* in a natural population of native Australian rats. *Env. Microbiol.*, 2: 594-610 (2000).

Raamsdonk LM, Teusink B, Broadhurst D, Zhang NS, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ, Westerhoff HV, van Dam K and Oliver SG. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat Biotechnol.*, 19: 45-50 (2001).

Schaefer U, Boos W, Takors R and Weuster-Botz D. Automated sampling device for monitoring intracellular metabolite dynamics. *Anal. Biochem.*, 270: 88-96 (1999).

Selander RK, Caugant DA and Whittam TS. In *Escherichia coli and Salmonella typhimurium. Cellular and molecular biology* (Ed, Neidhardt, F. C.) ASM Press, Washington DC, pp. 1625-1648 (1987).

Souza V, Rocha M, Valera A and Eguiarte LE. Genetic structure of natural populations of *Escherichia coli* in wild hosts on different continents. *Appl.Env. Microbiol.*, 65: 3373-3385 (1999).

Storz G and Hengge-Aronis R. *Bacterial stress responses,* ASM Press, Washington D.C (2000).

Sweetman G, Trinei M, Modha J, Kusel J, Freestone P, Fishov I, Joseleaupetit D, Redman C, Farmer P and Norris V. Electrospray ionization mass spectrometric analysis of phospholipids of *Escherichia coli*. *Mol. Microbiol.*, 20: 233-234 (1996).

Tempest DW, Meers JL and Brown CM. Influence of environment on the content and composition of microbial free amino acid pools. *J. Gen. Microbiol.*, 64: 171-185 (1970).

Tkachenko AG, Salakhetdinova OY and Pshenichnov MR. Putrescine/potassium exchange as an adaptive response of *Escherichia coli* to hyperosmotic stress. *Microbiol.*, 66: 274-278 (1997).

Tweeddale H, Notley-McRobb L and Ferenci T. Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool ("metabolome") analysis. *J.Bacteriol.*, 180: 5109-5116 (1998).

Tweeddale H, Notley-McRobb L and Ferenci T. Assessing the effect of reactive oxygen species on *Escherichia coli* using a metabolome approach. *Redox Rep.*, 4: 237-241 (1999).

Wada A, Mikkola R, Kurland CG and Ishihama A. Growth phase-coupled changes of the ribosome profile in natural isolates and laboratory strains of *Escherichia coli*. *J. Bacteriol.*, 182: 2893-2899 (2000).

Weuster-Botz D and de Graaf AA. Reaction engineering methods to study intracellular metabolite concentrations. *Adv. Biochem. Eng. Biotechnol.*, 54: 75-108 (1996).

Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, Sanchez JC, Yan JX, Gooley AA, Hughes G, Humpherysmith I, Williams KL and Hochstrasser DF. From proteins to proteomes - large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnol.*, 14: 61-65 (1996).

Zhao J, Baba T, Mori H and Shimizu K. Global metabolic response of *Escherichia coli* to gnd or zwf gene-knockout, based on ^{13}C-labeling experiments and the measurement of enzyme activities. *Appl. Microbiol. Biotechnol.*, 64: 91-98 (2004).

Chapter 6

CAPILLARY ELECTROPHORESIS AND ITS APPLICATION IN METABOLOME ANALYSIS

Li Jia and Shigeru Terabe
Graduate School of Material Science, University of Hyogo, Kamigori, Hyogo, 678-1297, Japan

1. INTRODUCTION

Capillary electrophoresis (CE) has undergone remarkably rapid developments since its introduction in the early to middle 1980s. The emergence of the first-generation commercially available equipment in 1989 opened up the possibility for analytical chemists to explore applications of CE in greatly diverse fields. With high field strength in narrow capillaries (less than 100 μm I.D.), CE has demonstrated high separation efficiency up to two orders of magnitude greater than high performance liquid chromatography (HPLC). Many advantageous features of CE include rapidity, low sample volume requirements (less than nL), ease of operation, economy, and overall versatility. Separations are based on the differences in electrophoretic mobilities of ions in electrophoretic media. There are several separation modes in CE, including, capillary zone electrophoresis (CZE), capillary gel electrophoresis (CGE), micellar electrokinetic chromatography (MEKC), capillary electrochromatography (CEC), capillary isoelectric focusing (CIEF), and capillary isotachophoresis (CITP). Among these, CZE and MEKC are more popular modes and appropriate for the study of small molecules and therefore well suited for metabolome analyses. Hence, in this chapter we just introduce these two modes of CE and their application within metabolomics.

2. CAPILLARY ELECTROPHORESIS

2.1 Basic principles

2.1.1 CZE

Currently CZE is the most commonly used separation mode in CE. A schematic diagram of the separation principle of CZE is shown in Figure 1. Separation in CZE is based on the differences in the electrophoretic mobilities resulting in different velocities of migration of ionic species in the electrophoretic buffer. Upon the application of a constant electric field, ionic species move with a constant velocity, which is proportional to the applied electric field. The proportionality constant, the electrophoretic mobility, is a characteristic property of a given ion in a given medium and at a given temperature. The separation is mainly based on the differences in charge-to-size ratios of solutes, and therefore, only ionic or charged analytes can be separated by the method. In most cases, uncoated open-tubular fused silica capillaries, which contain surface silanol groups, are employed in CZE. Electroosmotic flow is an important phenomenon in CE. It originates from negative charges on the inner wall of the capillary tube as the silanol groups become ionized in the presence of the electrophoretic medium (at pH > 2). The electroosmotic flow transports the bulk solution in the capillary with a flat velocity profile from the positive to negative electrode. It is stronger than the electrophoretic velocity of the individual ions in the injected sample. Consequently, both anions and cations migrate toward the negative electrode and can be separated in the same run.

2.1.2 MEKC

MEKC was introduced by Terabe and co-workers in 1984 (Terabe *et al.*, 1984). A schematic diagram of the separation principle of MEKC is shown in Figure 2. In MEKC, the main separation mechanism is based on solute partitioning between the micellar phase and solution phase. The technique provides a way to resolve neutral molecules as well as charged molecules.

A capillary is filled with an ionic surfactant solution of a concentration higher than its critical micelle concentration (CMC), above which the micelle is formed by the aggregation of surfactant molecules, as an electrophoretic solution instead of the simple buffer solution used in CZE. The ionic micelle works as the separation solution, and under the capillary electrophoretic condition the ionic micelle migrates at a different velocity

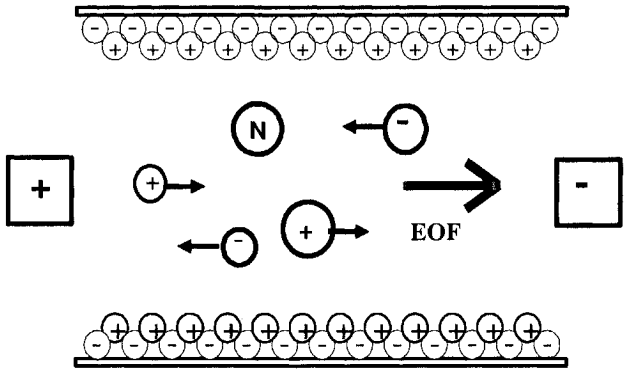


Figure 1. Schematic diagram of the separation principle of CZE. +, cation; -, anion; N, neutral; EOF, electroosmotic flow.

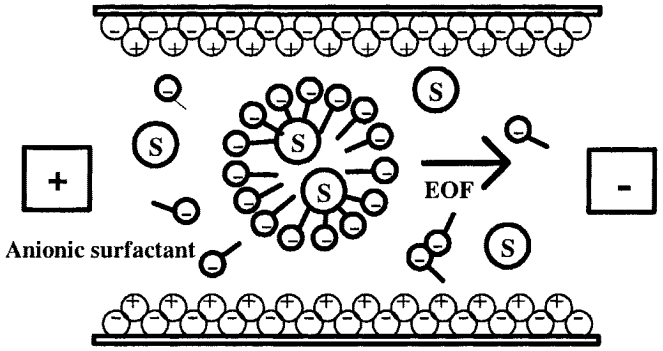


Figure 2. Schematic diagram of the separation principle of MEKC. +, cation; -, anion; S, solute; EOF, electroosmotic flow.

from the bulk solution because the micelle is subjected to the electrophoretic migration. The micelle corresponds to the stationary phase in chromatography, and therefore is called the pseudostationary phase. A fraction of the analyte is incorporated by the micelle in rapid equilibrium, having an effective electrophoretic mobility depending on the ratio of the incorporated analyte to the free analyte. The analyte free from the micelle migrates only by the electroosmotic flow, while the analyte totally incorporated by the micelle migrates at the velocity of the micelle or the sum of the electroosmotic velocity and the electrophoretic velocity of the micelle. Under neutral or alkaline conditions, the electroosmotic velocity is faster than the electrophoretic velocity of the micelle, and hence the micelle also

migrates in the same direction as the electroosmotic flow. When an anionic micelle such as sodium dodecyl sulfate (SDS) is employed, all the neutral analytes migrate toward the cathode due to the strong electroosmotic flow. The less-incorporated analytes migrate faster than the more incorporated analytes by the SDS micelle. The fraction of the analyte incorporated by the micelle increases with increase in hydrophobicity of the analytes. For ionic compounds, charge-to-size ratios, hydrophobicity and charge interactions at the surface of the micelles combine to influence the separation of the analytes.

2.2 Instrumentation

All electrophoretic modes except for CITP can be carried out, in principle, using the same equipment, which consists of an injection system, a high-voltage power supply, two buffer reservoirs, a capillary and a detector. The basic instrumental set-up to accomplish CE is depicted in Figure 3. Commercially available CE instruments are additionally equipped with an autosampler for sample injection allowing series analysis, column thermostating and a computer for instrumental control and data acquisition.

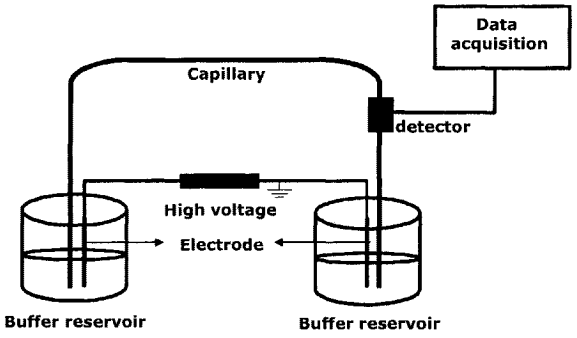


Figure 3. Basic instrumental setup for a CE system

Cylindrical polyimide-coated fused silica capillaries with narrow diameter (10-100 μm) are the most often used today. The narrow capillary diameter facilitates the dissipation of Joule heating generated by the electrical resistance of the electrolyte inside the capillary. During separation, the capillary filled with the buffer solution is placed between two buffer reservoirs. The electric field is applied by means of a high voltage power supply, which can generate voltages up to 30 kV. Injection of the analytes is performed by replacing one buffer reservoir by the sample vial. A defined

sample volume is introduced into the capillary by either hydrodynamic flow or electromigration. An on-column detector is located close to the end of the capillary, which is opposite to the injection site. Since injection and detection systems are the most important and most critical components of the instrumentation, particular emphasis is laid on them in the following discussion.

2.2.1 Injection

There are two fundamental injection systems, hydrodynamic injection and electrokinetic injection. For hydrodynamic injection, the sample is introduced into a capillary by means of differential pressure along the capillary, which is created by three main techniques, hydrodynamic, siphoning, or hydrostatic. The sample volume introduced by hydrodynamic injection can be manipulated by varying the injection time and the pressure difference. The injection volume is temperature dependent since it depends on the viscosity of the solution. A major limitation of the hydrodynamic injection is that it is not suitable for the injection of highly viscous samples.

Electrokinetic injection is also called electromigration injection, and is based on the fact that voltage causes electrophoretic and electroosmotic movement. To perform electrokinetic injection, the capillary and the electrode at the inlet side are removed from the buffer vial and placed into the sample vial. A voltage is then applied for a short interval of time, resulting in the transport of sample into the capillary by electromigration, which includes contributions from both electrophoretic migration of charged sample ions and electroosmotic flow of the sample solution. The sample volume can be controlled by varying the injection time and the applied voltage. It should be mentioned that there are two problems occurring in electrokinetic injection (Huang *et al.*, 1988). Firstly, a discrimination of the injected sample components occurs due to the mobility differences of the analytes. The ions with high mobilities are injected in larger quantities than those with low mobilities. The second problem is that the changes in the absolute amount injected into the capillary would occur due to the difference in the conductivity of the sample solution, which causes the changes in the electrophoretic mobilities and electroosmotic flow.

In view of the above, hydrodynamic injection is preferable over electrokinetic injection. However, there are occasions where the latter mode is to be preferred if discrimination of the component of interest from contaminants or a concentrating of a component from a diluted sample solution is desired.

2.2.2 Detection

A wide range of detection techniques have been studied in CE. Among them, on-column UV adsorption and fluorescence detection have been the most commonly used detection techniques for CE applications. Since mass spectrometry (MS) provides additional structural information of the separated compounds, the hyphenation of MS with CE is very useful for metabolome analysis. Hence, the three detection techniques will be discussed in this section.

In on-column UV absorbance detection, the capillary itself serves as the cylindrical detection cell, which was made by removing the polyimide coating from a short section of the fused silica capillary. UV absorbance detection is the most popular detection due to its relatively universal detection capability, simple adaptation and low cost. However, the detection sensitivity is not very high due to the limitation of the small inside diameter of the capillary and low injection volume. The concentration sensitivity is in the order of μM for most analytes with chromophores. In order to improve the sensitivity, several techniques have been developed by extending optical path length and on-column sample preconcentration. Extended path length absorbance detectors are commercially available, which include Z-shaped (Moring *et al.*, 1993) or bubble (Heiger, 1992) cells. On-column sample preconcentration techniques will be discussed below. Photodiode array (PDA) detection is employed to obtain the multiwavelength spectral information, which can be used to aid in the identification of unknown compounds and examination of peak purity.

On-column fluorescence detection is another very popular detection in CE, whose major advantage is its high detection sensitivity. The light source for fluorescence detection can be either an arc lamp or a laser. In contrast to arc lamps, lasers are particularly useful for sensitive detection on capillaries because of the ability to be focused into smaller volume. For laser-induced fluorescence detection (LIF), the concentration sensitivity is in the order of nM for analytes with fluorophores. The disadvantage is that the excitation wavelengths available from current types of laser sources are rather limited. Since most analytes are non-fluorescent, pre- or post-column derivatization of the sample with some type of fluorophore allows the extension of fluorescence detection to many analytes.

For compounds, which lack chromophores or fluorophores, indirect UV absorbance or fluorescence detection is available, where an electrolyte containing chromophore or fluorophore is used as a visualizing agent and analyte peaks are detected as negative peaks. Indirect detection can be performed using the same instrumentation as for the corresponding direct

detection. The sensitivity for indirect detection is slightly less than that for the direct detection counterpart.

The use of MS for detection not only provides excellent sensitivity and selectivity, but also structural information of unknown compounds. Moreover, it does not require that analytes have native UV absorbance or fluorescence. Hence, the hyphenation of MS with CE offers great potential for metabolome analysis. The detection sensitivity for MS is in the order of nM for most analytes. Unlike on-column UV absorbance and fluorescence detection, MS is an off-column detection method for CE. Therefore, the design of the interface of CE to MS is very important. The interfacing of CE to MS has been accomplished by the most common ionization techniques, namely electrospray ionization (ESI), which provides very mild ionization conditions that ensure molecular weight determination. Compatibility problems between CE and MS may arise from the buffer system used in CE. Non-volatile buffers such as sodium phosphate or borate widely used in CE are less suitable for CE-MS coupling. Volatile CE buffers such as ammonium acetate, triethylamine or trifluoroacetatic acid are compatible with MS.

2.3 Optimizing parameters

2.3.1 Capillary dimensions

Fused silica capillary dimensions used in CE range from 10 to 100 μm inner diameter, 375 μm outer diameter and 10 to 100 cm in length. The typical capillary dimension used in most CE experiments is 50 μm or 75 μm i.d., and 50 cm in length. The selection of the capillary dimensions influence several factors, such as migration time, resolution, detection sensitivity, and heat dissipation. At constant field strength, migration time increases with increase in the capillary length, as do the separation efficiency and the peak resolution. The inner diameter of the capillary affects the separation performance. The separation efficiency decreases with an increase in the inner diameter of the capillary since Joule heating is dissipated much better in small diameter capillaries. On the other hand absorption detection sensitivity decreases with smaller inner diameter capillaries because of the shorter optical path length.

2.3.2 Field strength

The field strength applied across the capillary is the driving force in CE, which is defined as the applied voltage divided by the total capillary length. Since both the electrophoretic migration velocity and the electroosmotic

flow velocity are directly proportional to the electric field, higher field strengths will bring about shorter analysis times. The separation efficiency increases with increase in the applied voltage for low values of field strength. A dramatic loss in resolution is found if the field strength is increased too high due to the influence of excessive heat generation. The optimal field strength can be determined from the plot of the field strength versus the resulting current as the point where deviation from linearity starts since the plot deviates from linearity in the high field strength due to the effect of excessive heat production.

2.3.3 Temperature

Joule heating, resulting from the electric current passing along a capillary is a major problem in CE separation, since it brings about an increase in the temperature within the capillary and a parabolic temperature gradient across the capillary. An increase in the temperature within the capillary can significantly reduce the efficiency in CE. Hence, it is important to dissipate Joule heating efficiently in the capillary by temperature control. Despite its negative effects in terms of Joule heating, electrolyte temperature can be exploited as a selective parameter. Joule heating can increase elelctroophoretic mobilities by about 2% per degree centigrade (Knox, 1988), owing to the decrease in viscosity of the electrophoretic buffer, resulting in the decrease in the migration time. Temperature can also influence the chemical equilibrium, such as metal chelation, micelle partitioning, complex formation and dissociation.

2.3.4 Electrolyte system

The electrolyte system plays a central role in CE performance. Properties like pH, ionic strength and the composition affect both selectivity and efficiency tremendously. The pH value of the electrolyte solution is the most important separation parameter for manipulation of the separation selectivity, since it influences the dissociation of weakly acidic, basic or zwitter-ionic analytes. Besides the pH, the ionic strength is an important tool that we can use to improve efficiency, resolution and sensitivity of the separation system. The ionic strength of the electrolyte system not only determines the degree of Joule heating at constant voltage, but also has a marked influence on both electroosmotic flow and electrophoretic mobility. The buffer composition can also improve efficiency as well as selectivity since the mobility of the buffer ions has effects on electrophoretic dispersion and the resulting current at a given field strength. The buffer capacity must be high enough such that the local pH and conductivity will not change as a

result of sample injection and electrolysis of water at the electrodes. The use of additives such as organic solvents and complexing agents (cyclodextrin, crown ether) is also an effective technique to improve resolution. Many enantiometic pairs are successfully separated by using a cyclodextrin derivative as a chiral additive.

The use of surfactants as micelle-forming modifiers to permit the separation of neutral analytes is a separation mode of CE called MEKC. Since MEKC is a chromatographic technique, the separation selectivity is manipulated by the chromatographic considerations. The choice of the surfactant, pH and composition of the running solution and the use of additives are important factors to manipulate selectivity. The chemical structure of the surfactant, in particular that of the polar group, affects selectivity significantly. Highly hydrophobic analytes tend to be totally incorporated by the micelle and migrate at the velocity of the micelle, being unresolved. To resolve highly hydrophobic compounds by MEKC, several modifiers (cyclodextrin, organic solvents, urea or glucose) are developed to reduce the fraction of analytes incorporated by the micelle.

2.4 On-line sample preconcentration

As mentioned above, the manipulation of the on-line capillary detection window afforded up to 10-fold response improvement with the most common UV detector. A more practical and moderate way to concentrate samples is the on-line preconcentration approach, which has developed into an exciting field of research. Several on-line sample preconcentration methods will be discussed in the following section.

2.4.1 Field-enhanced sample stacking

Field-enhanced sample stacking utilizes a high electric field observed in the sample zone by preparing the sample solution in a low electric conductivity matrix. Since the electrophoretic velocity is proportional to the field strength, analyte ions migrate at much faster velocity in the sample solution zone than in separation solution zone and stack at the boundary between the sample and separation solution zones. Sample stacking can be performed in both the hydrodynamic and electrokinetic injection modes, which includes several modes, such as normal stacking mode, large volume sample stacking (LVSS), LVSS with polarity switching, LVSS without polarity switching, filed-enhanced sample injection (FESI), etc., as reviewed by Quirino *et. al.* (2000). Deterioration of concentration efficiency in the sample stacking is caused by a mismatch of the electroosmotic flow. The electroosmotic velocity is also proportional to the field strength and must be

different between the two zones due to the difference in electric field strength. However, owing to the continuity of the solution, the bulk electroosmotic velocity must be constant throughout the capillary. Therefore, mixing must occur at the boundary of the two zones. This discrepancy is minimized when the electroosmotic flow is suppressed.

2.4.2 Sweeping

Sweeping is a preconcentration technique in MEKC developed by Quirino and Terabe (1998). It utilizes the phenomenon that hydrophobic analytes tend to be incorporated into the micelle. In sweeping, a homogeneous electric field is preferable unlike the sample stacking, that is, the sample solution is prepared as a solution having the same conductivity as that of the separation solution or background solution (BGS). Under a suppressed electroosmotic flow, when an ionic micelle like SDS enters continuously the long plug of the sample zone devoid of the micelle by electrophoresis from the inlet vial upon the application of the voltage, the analyte in the sample zone is picked and accumulated by the micelle at the front end of the micelle zone until the micelle reaches the end of the sample zone or the boundary between the sample zone and BGS zone. The analyte zone is focused into a very narrow zone if the interaction is strong between the analyte and the micelle, and separated by MEKC after the end of sweeping. Sweeping is effective for both charged and uncharged analytes, which interact strongly with the micelle. Sweeping is also powerful even in the presence of a strong electroosmotic flow although concentration efficiency is high under a suppressed electroosmotic flow. An advantage of sweeping is that sample matrix can contain relatively high concentrations of electrolytes since low conductivity is not required for the sample matrix. Unfortunately, sweeping is not efficient for the preconcentration of hydrophilic analytes or weakly interacting analytes with the micelle.

2.4.3 Dynamic pH junction

Dynamic pH junction was first reported by Britz-McKibbin *et. al.* (1998) when developing a specific assay for epinephrine in dental anesthetic solutions. It is an efficient preconcentration technique for the weakly ionic analytes if the difference in pH between the sample matrix and BGS can cause significant changes in their mobilities. Dynamic pH junction is defined when two or more sections of buffer that possess a different pH are loaded into the capillary to form a discrete step pH junction at the interface of the sample and BGS zones. Preconcentration by dynamic pH junction is hypothesized to be caused by the formation of a transient pH gradient (pH

titration) within the sample zone, which results in rapid focusing of analytes that undergo velocity changes in the selected pH range. The sample may consist of the same buffer or different electrolyte type as BGS to optimize the pH junction range for the focusing of weakly acidic, basic or zwitter-ionic analytes (mobility is pH dependent) based on their pK_a and/or pI.

2.4.4 Dynamic pH junction-sweeping

A hyphenated dynamic pH junction-sweeping technique was developed by Britz-McKibbin *et al.* (2003). It is an effective on-line preconcentration method suitable for both hydrophilic (weakly ionic) and hydrophobic (neutral) analytes. Dynamic pH junction-sweeping is defined when the sample is devoid of micelle (sweeping condition) and has a different buffer pH (dynamic pH junction condition) relative to the BGS, which permits efficient focusing of large volumes of analytes directly on-capillary. Compared to either sweeping or dynamic pH junction techniques alone, several fold enhancements in analyte sensitivity was demonstrated by dynamic pH junction-sweeping. Analyte focusing is mediated by three distinct factors: differences in buffer pH, borate complexation, and micelle partitioning. Highly focused analyte bands are important not only for enhanced sensitivity, but also for improved resolution in CE.

2.4.5 Transient-isotachophoresis

Transient-isotachophoresis (t-ITP) is a simple form of ITP, which is easy to couple to CZE. In t-ITP-CZE, high concentrations of leading/terminating co-ions that possess mobilities greater and less than the mobility of the analyte, respectively, are added to the sample and/or BGS. Both ITP preconcentration and CZE separation process are conducted in the same capillary and can be run on commercial instruments. Karger and co-workers described on-column t-ITP preconcentration technique (Foret *et al.*, 1992). In many cases the t-ITP step occurs accidentally in samples containing high concentrations of salts or it can also be induced by addition of an appropriate leading or terminating ions to samples. The preconcentration of a sample ion with an intermediate ion mobility present at a low concentration is due to the need to change its concentration and in turn its field strength to keep up in pace with the velocity of the leading ion. The technique can concentrate both small and large molecules. Careful selection of appropriate leading and terminating co-ions is normally required for specific analytes.

3. APPLICATION IN METABOLOME ANALYSIS

Metabolome analysis is the systematic chemical analysis of metabolites present in a cell. Metabolites represent hundreds of diverse classes of small organic molecules, including amino acids, nucleotides, carbohydrates, carboxylic acids, vitamins and coenzymes. Because of the large number and low concentration of many intracellular metabolites and the changes in their concentrations with environment and cell history, metabolome studies require sensitive, selective, and high throughput separation techniques. Two different approaches to intracellular metabolite analysis can be adopted: comprehensive (complete metabolite profile) and selective (specific class of metabolites or metabolites in common metabolic pathway). Owing to the advantages of CE as mentioned above, it is employed to develop the comprehensive analytical methods of intracellular metabolites. Due to the relatively low concentration sensitivity in CE, on-line preconcentration approaches are utilized in metabolome analysis.

3.1 Target metabolites

The flavins, riboflavin (RF), flavin mononuleotide (FMN), and flavin adenine dinucleotide (FAD) represent an important class of metabolites and are natively fluorescent. CE with LIF detection was applied to analyze trace amounts of flavins from different types of biological samples (including bacterial cell extracts, recombinant protein, pooled human plasma and urine) using dynamic pH junction-sweeping as an on-line preconcentration technique (Britz-McKibbin *et al.*, 2003). Over a 1200-fold improvement in concentration sensitivity was demonstrated compared to conventional injections, resulting in a limit of detection (LOD) of about 4.0 pM for the flavin coenzymes FAD and FMN. Figure 4 shows electropherograms depicting analysis of flavin coenzymes in cell extracts of *Bacillus subtilis* by CE-LIF.

Intracellular nucleotide profiles are vital in studies of cell metabolism and their changes associated with a variety of disease processes. Nucleotide profiles from a mouse lymphoma were analyzed by CE with UV detection using dynamic pH junction as an on-line preconcentration technique (Britz-McKibbin *et al.*, 2000). The method allows the injection of large volumes of sample (~300 nL), resulting in at least 50-fold improvement in concentration sensitivity. The LOD of 40 nM for nucleotides can be achieved in optimum conditions. The elimination of time-consuming preconcentration and desalting procedures for biological samples can be realized using the method.

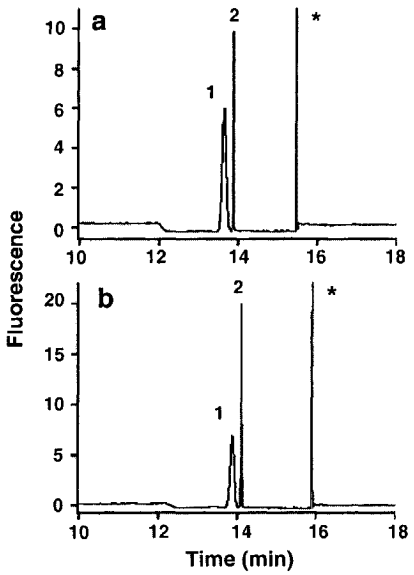


Figure 4. Electropherograms depicting analysis of submicromolar concentrations of flavin coenzymes in cell extracts of *B. subtilis* by CE-LIF with dynamic pH junction-sweeping using (a) glucose and (b) malate as carbon source in culture media. Samples were diluted 25-fold in 75mM phosphate, pH 6.0, prior to injection. Conditions: BGE, 140mM borate, 100mM SDS, 5mM β-CD, pH 8.5; voltage, 15 kV; capillary length, 57 cm; injection, 60 s (15% capillary length). Analyte peak numbering corresponds to 1, FMN and 2, FAD; *represents system peak.

The pyridine nucleotides (NAD, NADP, NADH, and NADPH) represent a class of coenzymes involved in a number of critical catabolic and anabolic pathways in living organisms. The adenine nucleotides (AMP, ADP, and ATP) also play an important role as physiological signaling molecules which bind to membrane purine receptors. Our group developed a sensitive CE method to analyze the pyridine and adenine nucleotide metabolites derived from *B. subtilis* cell extracts using UV as detector and sweeping by borate complexation as an on-line preconcentration approach (Markuszewski *et al.*, 2003). LODs of nucleotide metabolites were less than 20 nM and LOQs were between 50 and 80 nM. Figure 5 presents electropherograms in analysis of *B. subtilis* cell extracts from glucose and malate as culture media.

For some important metabolites in the cell, which have no or weak chromophores for photometric detection, the derivatization of the specific types of metabolites with fluorescent or UV probes, or utilization of indirect detection mode may be the best choice. Amino acids are important metabolites in the cell, but most amino acids do not have strong

chromophores. Therefore, derivatization of amino acids with fluorescent or UV probes is required to enhance detector sensitivity. In our research group, LIF detection with argon ion laser (488 nm) as an excitation source was employed and 4-fluoro-7-nitrobenz-2-oxa-1,3-diazole (NBD-F) was used as a fluorescent reagent to derivatize amino acids. Due to the hydrophobicity of the derivatized amino acids, an MEKC method was developed to analyze amino acids in the cell extract of *B. subtilis* (Terabe *et al.*, 2001).

In the tricarboxylic acid (TCA) cycle, the main metabolites are di- and tricarboxylic acids, most of which have low molar absorptivity and thus are poorly detectable compounds by photometric detection. In our research group, tri- and dicarboxylic acids from TCA cycle as well as carboxylic acid metabolites from other metabolic pathways (e.g., glycolysis, urea cycle and metabolism of amino compounds) in *B. subtilis* cell from two different cultures (glucose and malate) were analyzed by CE with indirect detection mode using 2,6-pyridinedicarboxylic acid as a highly UV absorbing carrier electrolyte (Markuszewski *et al.*, 2003). With an electrokinetic injection mode LODs of the analytes in the range of 11 – 60 μM were achieved. Figure 6 shows electropherograms of *B. subtilis* cell extracts from glucose and malate as culture media by CE with indirect UV detection. Six aromatic

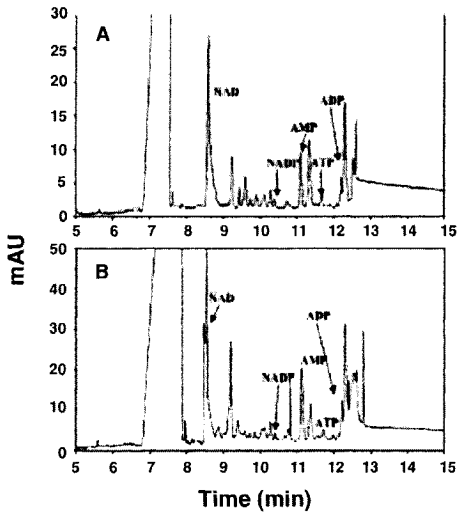


Figure 5. Electropherograms presenting analysis of *B. subtilis* cell extracts from: (A) glucose and (B) malate as culture media by CE with sweeping by borate complexation. Experimental conditions: 150 mM borate, applied potential, 20 kV; injection pressure, 50 mbar, 40 s; capillary temperature, 20 °C; detection, 200 nm; fused-silica capillary, 56.0 cm (48.6 cm effective length) x 50 μm I.D.

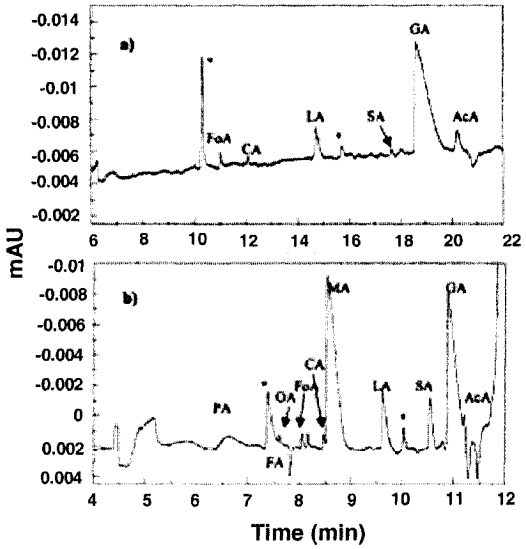


Figure 6. Electropherograms depicting analysis of carboxylic acids in bacteria *B. subtilis* cell extracts from: (a) glucose and (b) malate, as culture medium by CE with indirect UV detection. Experimental conditions: BGS, 4 mM 2,6-pyridinedicarboxylic acid, 0.2 mM CTAB (pH 3.5), 10% ethylene glycol and 10% acetonitrile; untreated fused-silica capillary, 75 μm I.D., 70.6 cm effective length; electrokinetic injection, 10 kV x 10 s; detection, 200 nm; voltage -30 kV; temperature, 15 °C. Samples: PA, pyruvic acid; OA, 2-oxoglutaric acid; FA, fumaric acid; FoA, formic acid; CA, citric acid; MA, malic acid; LA, lactic acid; SA, succinic acid; GA, glutamic acid; AcA, acetic acid; * unidentified peaks.

carboxylic acids (cinnamic acid, 3-(4-hydroxyphenyl)pyruvic acid, 4-hydroxybenzoic acid, 4-hydroxyphenylacetic acid, protocatechuic acid, and 3,4-dihydroxyphenylacetic acid) expected to be found in the cell metabolites were analyzed by CZE using field-enhanced stacking as an on-line preconcentration method. The LODs of the analytes were from 0.4 to 2.0 μM at S/N ratio of 3 (Terabe *et al.*, 2001).

The purines represent an important class of metabolites, which serve as vital precursors for the biosynthesis of DNA and RNA nucleotides in a cell. The analysis of purine levels in biological fluids plays an important role in the diagnosis of a variety of metabolic disorders. On-line focusing of xanthine and other purine derivatives by CE using dynamic pH junction as an on-line preconcentration method was reported. The technique was demonstrated to analyze micromolar concentrations of xanthine in pooled urine (Britz-McKibbin and Chen, 2003). New separation platforms for high

throughput analysis based on multiplexed CE (capillary array format) promise rapid and highly efficient separations, as highlighted by its important role in rapid DNA sequencing used in the Human Genome Project. In our research group, a multiplexed CE system with UV detection in conjunction with dynamic pH junction was demonstrated as a novel method for the sensitive and high-throughput analysis of purine metabolites (Britz-McKibbin *et al.*, 2003). Multiplexed CE can be used for the rapid optimization of the focusing conditions using up to 96 different sample matrixes in a single run. Over a 50-fold enhancement in concentration sensitivity compared to conventional injections is realized by this technique.

Soga *et al.* (2002) developed a method for simultaneous determination of 32 standard anionic metabolites containing carboxylic acids, phosphorylated carboxylic acids, phosphorylated saccharides, nucleotides, and nicotinamide and flavin adenine coenzymes of glycolysis and the TCA cycle pathways based on CE coupled to ESI-MS. A cationic polymer-coated capillary was employed, where EOF was reversed. The method was applied to the comprehensive analysis of metabolic intermediates extracted from *B. subtilis*, and 27 anionic metabolites were detected and quantified. A method for the simultaneous and quantitative analysis of multivalent anions, such as citrate isomers, nucleotides, NAD, FAD and coenzyme compounds extracted from *B. subtilis* cells was also reported by Soga *et. al.* (2002) based on pressure-assisted CE (PACE) coupled to ESI-MS. In the method, a capillary coated with a noncharged polymer (poly-dimethylsiloxane) was used to prevent anionic species from adsorbing onto the capillary wall. Comparing with the CE/ESI-MS method, the PACE/ESI-MS method improved reproducibility and sensitivity of the anions, but the theoretical plates were inferior.

3.2 Metabolome profiling

The comprehensive analysis of intracellular metabolites can reveal the connection of biochemical networks and provide a systematic understanding of the cell. Hence, the metabolome profiling analysis is of crucial importance. Since the metabolome is dynamic and highly variable with cell types, genes, environment and history, and over 1000 different metabolites exist in a cell, it is impossible to analyze the intracellular metabolites profile in one run using a single chromatographic or electrophoretic technique. A key point is that a single dimensional separation method has limited peak capacity. According to the mathematical model introduced by Giddings (1987), the peak capacity of a multi-dimensional separation system is the product of the peak capacity of its components. Therefore, a multi-

dimensional separation system is a more practical choice in order to separate as many metabolites as possible.

Soga *et. al.* (2003) reported that 352 metabolic standards were first separated by CE based on their charge-to-size ratios and then selectively detected using MS by monitoring over a large range of *m/z* values. The method was applied to analyze 1692 metabolites from *B. subtilis* extracts, revealing significant changes in metabolites during *B. subtilis* sporulation. Recently, our research group developed a two-dimensional separation method for profiling *B. subtilis* metabolites, which hyphenated chromatography and electrophoresis (Jia *et al.*, 2004). *B. subtilis* cell extract was first separated based on their hydrophobicity using a monolithic silica-ODS column and a laboratory-assembled micro-LC apparatus operated in the gradient elution mode. The fractions of effluent from the column were collected every minute (2 μL/vial). After collection, the fractions were dried at room temperature under vacuum, and reconstituted with 10 μL of 50 mM phosphoric acid or 75 mM sodium phosphate (pH 6.0) before CE analysis. The early-eluting fractions were separated by dynamic pH junction CZE based on their charge-to-size ratios, while the late-eluting fractions were separated by sweeping MEKC based on their hydrophobicity. The middle fractions were analyzed using both modes of CE. Concentration strategies, namely dynamic pH junction and sweeping, were employed to interface the two dimensions, which proved to be beneficial for the detection of metabolites. Some important metabolites in the *B. subtilis* cell were identified. This method provided great potential for resolving complex biological samples containing compounds having different characteristics.

4. ROLE OF CAPILLARY ELECTROPHORESIS IN METABOLOME ANALYSIS

High efficiency, high sensitivity and high throughput are important and crucial aspects of metabolomics research. CE offers a unique platform for carrying out separations that is useful for metabolomic studies due to its advantages highlighted above. In CE, different on-line sample preconcentration approaches have been used to improve sensitivity. The introduction of commercial multiplexed CE system provides a convenient platform for rapid method development and high throughput analysis of hundreds of metabolites present in a cell. Hence, CE would be advantageous for analysis of specific classes of metabolites in a cell.

So far no single chromatography or electrophoresis procedure in one run is likely to resolve a complex mixture of cell metabolites. A multidimensional system, which employs two or more orthogonal separation

techniques or separation methods with different separation mechanisms, will significantly improve the chances of resolving such a complex mixture of cell metabolites. Multiplexed CE platform would be extremely useful for high-throughput metabolite profiling. Multiplexed CE can be coupled orthogonally to HPLC for multidimensional separations by collecting fractions of complex samples using 96 microtiter plates with the aid of a microfraction collector. UV detection is beneficial for metabolites having chromophores, while LIF detection is advantageous for those having fluorophores. Due to the complexity of the metabolome, it is impossible to find a universal UV-absorbance or fluorescence-labeling reagent. MS is a rational choice owing to its universality, sensitivity, selectivity, and ability to providing structural information of metabolites. Hence, a multidimensional system with MS as detection would be a practical alternative for future metabolomic studies, in which CE would play an active role.

REFERENCES

Britz-McKibbin P *et al.* Quantitative assay for epinephrine in dental anesthetic solutions by capillary electrophoresis. *Analyst*, 123: 1461-1463 (1998).

Britz-McKibbin P, Bebault GM, Chen DY. Velocity-difference induced focusing of nucleotides in capillary electrophoresis with a dynamic pH junction. *Anal. Chem.*, 72: 1729-1735 (2000).

Britz-McKibbin P, Otsuka K, Terabe S. On-line focusing of flavin derivatives using dynamic pH junction-sweeping capillary electrophoresis with laser-induced fluorescence detection. *Anal. Chem.*, 74: 3736-3743 (2003).

Britz-McKibbin P *et al.* Picomolar analysis of flavins in biological samples by dynamic pH juntion-sweeping capillary electrophoresis with laser-induced fluorescence detection. *Anal. Biochem.*, 313: 89-96 (2003).

Britz-McKibbin P, Chen DDY. Velocity-difference induced focusing of xanthine and purine metabolites by capillary electrophoresis using a dynamic pH junction. *Chromatographia*, 57: 87-93 (2003).

Britz-McKibbin P, Nishioka T, Terabe S. Sensitive and high-throughput analyses of purine metabolites by dynamic pH junction multiplexed capillary electrophoresis: a new tool for metabolomic studies. *Anal. Sci.*, 19: 99-104 (2003).

Foret F, Szoko E, Karger BL. On-column transient and coupled column isotachophoretic preconcentration of protein samples in capillary zone electrophoresis. *J. Chromatogr. A*, 608: 3-12 (1992).

Giddings JC. Concepts and comparisons in multidimensional separation. *J. High. Resolut. Chromatogr.*, 10: 319-323 (1987).

Heiger DN. *High performance capillary electrophoresis*. pp. 100-101, Hewlett-Packard (1992).

Huang X, Gordon M, Zare RA. Bias in quantitative capillary zone electrophoresis caused by electrokinetic sample injection. *Anal. Chem.*, 60: 375-377 (1988).

Jia L *et al.* Two-dimensional separation method for analysis of *Bacillus subtilis* metabolites via hyphenation of micro-liquid chromatography and capillary electrophoresis. *Anal. Chem.*, 76: 1419-1428 (2004).

Knox JH. Thermal effects and band spreading in capillary electro-separation. *Chromatographia*, 26: 329-336 (1988).

Markuszewski MJ *et al.* Determination of pyridine and adenine nucleotide metabolites in *Bacillus subtilis* cell extract by sweeping borate complexation capillary electrophoresis. *J. Chromatogr. A*, 989: 293-301 (2003).

Markuszewski MJ *et al.* Analysis of carboxylic acid metabolites from the tricarboxylic acid cycle *in Bacillus subtilis* cell extract by capillary electrophoresis using an indirect photometric detection method. *J. Chromatogr. A*, 1010: 113-121 (2003).

Moring SE, Reel RT, Van Soest RJ. Optical improvements of a Z-shaped cell for high-sensitivity UV absorbance detection in capillary electrophoresis. *Anal. Chem.*, 65: 3454-3459 (1993).

Quirino JP, Terabe S. Exceeding 5000-fold concentration of dilute analytes in micellar electrokinetic chromatography. *Science*, 282: 465-468 (1998).

Quirino JP, Terabe S. Sample stacking of cationic and anionic analytes in capillary electrophoresis. *J. Chromatogr. A.*, 902: 119-135 (2000).

Soga T *et al.* Simutaneous determination of anionic intermediates for *Bacillus subtilis* metabolic pathways by capillary electrophoresis electrospray ionization mass spectrometry. *Anal. Chem.*, 74: 2233-2239 (2002).

Soga T *et al.* Pressure-assisted capillary electrophoresis electrospray ionization mass spectrometry for analysis of multivalent anions. *Anal. Chem.*, 74: 6224-6229 (2002).

Soga T *et al.* Quantitatiove metabolome analysis using capillary electrophoresis mass spectrometry. *J. Proteome. Res.*, 2: 488-494 (2003).

Terabe S *et al.* Electrokinetic separations with micellar solutions and open-tubular capillaries. *Anal. Chem.*, 56: 111-113 (1984).

Terabe S *et al.* Capillary electrophoretic techniques toward the metabolome analysis. *Pure Appl. Chem.*, 73: 1563-1572 (2001).

Chapter 7

METABOLITE PROFILING WITH GC-MS AND LC-MS

A key tool for contemporary biology

Ralf Looser, Arno J. Krotzky, Richard N. Trethewey
metanomics GmbH and Co. KGaA, metanomics Health GmbH, Tegeler Weg 33 10589 Berlin, Germany.

1. INTRODUCTION

Compared to the recent rapid progress in elucidating genomes, our understanding of metabolites, metabolite pathways, as well as the nature and regulation of metabolic networks is still in its infancy. This is widely misunderstood given the often authoritative nature of biological textbooks. Whilst we do have a good understanding of the pathways of primary metabolism our understanding of both secondary metabolism and the regulation of metabolic flux is poor.

Recently, the term "metabolome" has been adopted for the total metabolite complement of a cell, tissue or organism; correspondingly the study of some or all of the metabolome has become known as metabolomics or metabonomics. There is considerable debate about the precise use of these terminologies and the reader is advised to exert caution (Fiehn, 2002; Nicholson *et al.*, 2002; Sumner *et al.*, 2003). Estimates vary greatly as to how large metabolomes are. More than 100,000 secondary metabolites have so far been identified in plants and this probably amounts to less than 10 % of the total in the plant kingdom (Wink, 1988). For an individual plant species, or for humans, estimates of the metabolome size vary in the 3-25,000 range. In addition, metabolites of similar or different chemical nature can occur *in vivo* as conjugates, which significantly adds to the size and complexity of the metabolome. Whilst being highly complex and representing a large range of different chemistries, the metabolome

comprises molecules which are indicative of the genetic or physiological status of an organism. In addition, in contrast to DNA and RNA, the metabolome also reflects the entirety of internal or external influences on an organism or tissue – including environment, behavior and disease.

Whilst there are currently no methods that are close to delivering a comprehensive measurement of the metabolome there are a range of technologies that can generate quantitative metabolite profiles of several hundred metabolites. Mass spectrometry is particularly suited to this task and we will focus on GC-MS and LC-MS. We will introduce the technologies, discuss their relative strengths, review the challenges associated with high throughput operation and summarize some of the key recent applications.

2. GC-MS BASED METABOLITE PROFILING

The origin of metabolite profiling can be traced back to exploratory applications of GC to clinical research in the late 1960s and early 1970s. (Horning and Horning, 1971), in a research article in *Clinical Chemistry*, provided one of the first definitions: "Metabolic profiles are multi-component GC analyses that define or describe metabolic patterns for a group of metabolically or analytically related metabolites". This paper reported on novel methods for the analysis of steroids, acids and drugs or drug metabolites in human and rat urine using GC and GC-MS. Even though these methods were only able to resolve around 20 analytes, the authors realized the potential of the approach: "Profiles may prove to be useful for characterizing both normal and pathological states, for studies of drug metabolism and the effects of drugs on human metabolism, and for human developmental studies". Subsequently, clinical applications were pursued consistently by a small number of groups over the next decades (Niwa, 1986) with most approaches utilizing GC-MS. However, due to the difficulties associated with hardware stability and the computational limitations of dealing with complex profiles these approaches were never widely established and the literature thins out noticeably in the late 1980s.

Emboldened by the medical literature, the group of Sauter at BASF AG began in the late 1970s to develop methods for classifying herbicide mode of action. An example was published in 1991 which describes the classification of metabolite profiles following the application of known and unknown herbicides to barley seedlings (Sauter *et al.*, 1991). The authors were able to resolve between 100 and 200 peaks with a high degree of reproducibility and could determine the structure of around 70 compounds. The report concludes that metabolite profiling is a very powerful tool to support the chemist

working on the design and synthesis of new agrochemicals and that, following the establishment of a library of responses, "one will learn gradually to interpret the strange language of the profile".

In the 1990s rapid developments in both the engineering of GC-MS systems and in the power and affordability of computing systems led to the emergence of robust bench-top systems. Impressive improvements in software enabled biological laboratories to perform GC-MS analysis on a routine basis. Driven by the need to establish rapid and wide approaches to characterizing metabolism in plant metabolite engineering projects (Section 5), the group of Trethewey adapted the method of Sauter to the analysis of potato tuber material. Initial studies revealed exceedingly complex chromatograms with many hundreds of distinct peaks. However, a large number of these peaks were multiple isomers of the abundant sugars present in tubers, formed as part of the silylation derivatization procedure. Therefore the method was modified by separating the extracts into polar and non-polar components and a two-step derivatization procedure was adopted to reduce isomer complexity. A full validation of the method was published showing that it allows the qualitative and quantitative determination of more than 150 tuber metabolites including sugars, sugar alcohols, dimeric and trimeric saccharides, amines, amino acids and organic acids (Roessner *et al.*, 2000). The analytical variability of the method was demonstrated to be an order of magnitude lower than the biological variability. Subsequently the same methodology was adapted to the measurement of *Arabidopsis* leaves and similar performance was observed (Fiehn *et al.*, 2000a). Typical examples of chromatograms generated by this method at the company *metanomics* are shown in Figure 1.

Despite being a relatively mature technology, there is continued innovation in GC-MS that is applicable to metabolite profiling (Santos and Galceran, 2003). Compact commercial systems where GC is coupled to time-of-flight (TOF) mass spectrometry first became available in the late 1990s. In comparison to quadrupole technology TOF provides faster scan times. This enables either reduced run times for complex mixtures or improved peak annotation (deconvolution) and has a higher accuracy in the determination of mass-to-charge (*m/z*) ratios. Currently this technology is much discussed in the community but there are hitherto only a few publications on metabolite profiling applications (Fiehn, 2003; Taylor *et al.*, 2002; Wagner *et al.*, 2003). A further significant recent innovation has been comprehensive two-dimensional GC coupled to MS (GC x GC-MS) (Dallüge *et al.*, 2003). The ability to perform a second dimension of GC separation offers a range of advantages for metabolite profiling: faster run times, increased separation of isomers and improved peak deconvolution. The 2-dimensional nature of the chromatography presents new challenges

for the software required to process metabolite profiles and it will probably be some time before this technology finds routine application.

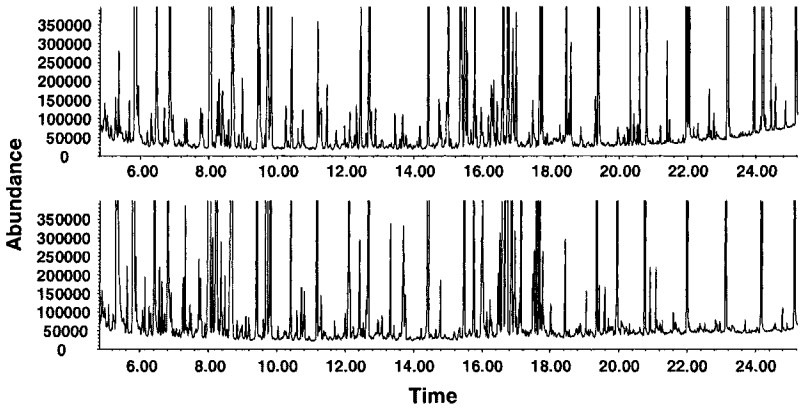


Figure 1. GC-MS chromatograms of a polar extract of Arabidopsis leaves (top) and rat blood plasma (bottom).

GC-MS is and will remain for the foreseeable future a key technology for metabolite profiling. Whilst it is older and has limitations in the range of chemistries that can be analyzed, it has the decisive advantages of being robust, relatively inexpensive, highly reproducible, draws on standardized spectral libraries and can be performed with a range of dedicated software. The recent innovations in GC-TOF and GC x GC-MS will ensure that increased performance can be anticipated from GC-based metabolite profiling.

3. LC-MS BASED METABOLITE PROFILING

Historically, LC separation technologies suffered from limited peak capacity (resolution of individual analytes in complex mixtures) compared to capillary techniques like high-resolution GC or capillary electrophoresis (CE). However, the emergence of the LC-MS coupling, first commercially available in the early 1990s and the iterative improvement in the performance of LC columns has recently made comprehensive metabolite profiling, as described before for GC, possible (Niessen, 1999). This has considerably widened the analytical range to numerous metabolites which are not amenable to GC-MS, either since they cannot be made volatile or because they are unstable during derivatization procedures. Furthermore, the

various ionization techniques available for LC-MS like electrospray (ESI), atmospheric pressure ionization (API), atmospheric pressure chemical ionization (APCI) and atmospheric pressure photoionization (APPI) are soft and versatile and even capable of generating ions from labile analytes (Hayen and Karst, 2003; Hsieh *et al.*, 2003). A very important issue to be considered when applying such MS techniques is ion suppression due to matrix derived compounds passing through the ion source at the same time as the metabolites to be analyzed (Annesley, 2003; Matuszewski *et al.*, 2003). Thus chromatographic separation combined with MS detection is always favorable in quantitative applications in comparison to direct injection MS.

The development of applications for LC-MS has been pioneered by the pharmaceutical industry. In particular, there is a considerable literature on the analysis of the metabolites of particular drugs or xenobiotics, an application often misleadingly called (drug) metabolite profiling (Niessen, 2003). Applications of LC-MS to the study of natural products have concentrated on profiling approaches targeted to distinct substance classes (Huhman and Sumner, 2003; Lange *et al.*, 2001) However, it is only very recently that comprehensive metabolite profiling approaches, equivalent to the application previously described for GC-MS, have been reported for plant (Tolstikov *et al.*, 2003; Tolstikov and Fiehn, 2002) and animal matrices. A non-targeted profiling of extracts of *Arabidopsis* leaves and roots with capillary LC coupled to TOF-MS yielded 2000 different mass signals, originating from both primary and secondary plant metabolites (Roepenack-Lahaye *et al.*, 2004). Some of the mass signals could be assigned either to known chemical structures or at least structural classes. A second study with *Arabidopsis* extracts has reported that the combination of monolithic capillary LC-columns coupled to ion-trap MS detection allows several hundred peaks to be followed (Tolstikov *et al.*, 2003). An application for the metabolite profiling of rat urine following the administration of pharmaceuticals has recently been published (Plumb *et al.*, 2002). The authors determined a peak capacity of 78 from a 10 minute chromatogram. When the metabolite profiles were subject to principal component analysis (PCA) it was possible to differentiate the dosed samples from the control samples. Another report of the same group describes the use of LC-MS profiling of mouse urine combined with subsequent pattern recognition techniques. Here, it was possible to discriminate between mouse strain and gender and, in addition, diurnal variation in the endogenous metabolites could also be distinguished (Plumb *et al.*, 2003a).

As with GC-MS, considerable further innovation in LC-MS technologies can be anticipated in the near future. The combination of liquid chromatography with more sophisticated mass spectrometric techniques like

linear ion traps (Schwartz *et al.*, 2002; Stolker *et al.*, 2004) and Fourier-transform ion cyclotron resonance mass spectrometry (FT-ICR-MS, also FTMS) may be particularly promising. In the case of FTMS, the high mass accuracy may be of great value to support structural elucidation of unknown molecules in the metabolite profiles. However, to our knowledge there are currently no reports on the use of LC-FTMS in metabolomics and currently the potential can only be estimated by extrapolation from applications in peptide research and proteomics (Aharoni *et al.*, 2002; Davidson and Frego, 2002; Martin *et al.*, 2000; Zubritsky, 2000).

4. HIGH THROUGHPUT METABOLITE PROFILING WITH GC-MS AND LC-MS

Technological advances in the fields of analytical instrumentation and informatics have made the high throughput operation of GC-MS and LC-MS possible. High throughput may be necessary for the analysis of large sample numbers where a fast delivery of results is required for time critical studies. Also, high throughput enables the multiple analyses of samples with a larger number of analytical methods in parallel, in order to achieve the widest possible analytical range. There are nonetheless different obstacles to overcome. Before implementation, each step of the analytical process has to be rigorously evaluated and optimized for compatibility with high throughput and possible automation. In particular, automation is a critical issue as there are currently no off-the-shelf solutions for robotic systems and the correct balance needs to be found between hard wiring of processes and maintaining flexibility.

For high throughput operation it is desirable to minimize the complexity of the extraction procedure. Suitable techniques include ball mills, (Fiehn *et al.*, 2000b), extraction systems like pressurized solvent extraction (PSE, also referred to as accelerated solvent extraction ASE (Richter *et al.*, 1995)) or ultrasound assisted extraction (Luque-Garcia and Luque de Castro, 2004; Rostagno *et al.*, 2003). The appropriate extraction procedure has to be selected according to the quality of the measurements that are required for the objectives of the work. Critical for successful high throughput operation is that the chromatographic systems have to be very reliable ("work horses") and constantly maintained at a high performance level; this is one of the decisive advantages of GC-MS and LC-MS technologies in comparison to newer technologies. However, considerable care has to be exerted in choosing the appropriate commercial provider and system. Further, if the results of experiments performed over months or even years have to be comparable, then a high degree of temporal stability is required. In general,

it is mandatory for metabolite profiling that instrument parameters are continuously monitored in order to avoid instrument based artifacts which can compromise analytical results much more than in conventional analytical techniques.

Up-scaling to high throughput operation should only occur after a thorough method validation under real time conditions. In contrast to classical analytical procedures (ISO: International Organization for Standardization, GLP: Good Laboratory Practice, AOAC International, www.aoac.org) there are currently no commonly agreed standards for what is necessary to validate a metabolite profiling method, let alone considering the needs of high throughput. Various initiatives are emerging to target this question (e.g. www.smrs.net; www.metabolomics.nl) and it can be anticipated that widely accepted standards will be established in the near future. Important approaches for method validation include recovery experiments and determinations of limits of detection and quantification (LOD, LOQ) across all peaks that are studied – even if no suitable reference materials are available.

A key bottleneck in the handling of chromatographic data is the reliable integration of chromatographic peaks. If many systems are run in parallel under identical chromatographic conditions differences in retention times of the same analyte on the respective systems have to be minimized. One possibility for GC analysis is the retention time locking system whereby the retention time of a known analyte is monitored from run to run and is adjusted automatically via small changes in carrier gas pressure. Automatic retention time adjustments are also possible for LC runs using relative retention times to either standard analytes or known peaks in the sample. Manual inspection of the chromatograms is not possible in a high throughput environment such as that at *metanomics* where some 300,000 peaks are processed per day. For this purpose software has to be developed which automatically checks peak data for compliance with predefined rules and guides the analyst to those data which have failed the validation process.

Figure 2 shows the schematic sample workflow of a complete high throughput metabolite profiling platform as implemented at the authors company, *metanomics*. The amount of raw data generated from full-scan LC-MS and GC-MS in high throughput sums up to terabytes per year and data storage and management systems capable of handling millions of files are necessary. In addition, to ensure sample tracking, a Laboratory Information Management System (LIMS) is required to underpin all steps in the operation. Maintaining quality, over time, in high throughput is a daunting challenge and the operation must be supported by routine quality control measures: for example, at *metanomics* some 30 % of the

chromatograms are of control samples and samples to monitor critical process steps and instrument parameters.

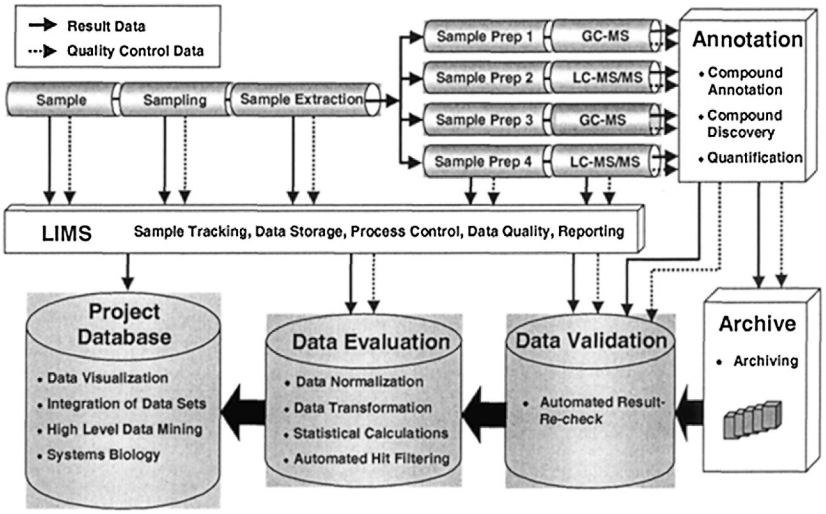


Figure 2. High throughput metabolite profiling at *metanomics*: schematic sample and data flow.

5. CONTEMPORARY APPLICATIONS OF GC-MS AND LC-MS METABOLITE PROFILING

The rapid improvement in the technologies of GC-MS and LC-MS has led to a range of new applications. In this section we introduce examples from some of the key fields that are shaping contemporary biology. It should be noted, that metabolic profiling can be applied in targeted approaches (focused on a large number of metabolites of interest), in non-biased approaches (comprising all metabolites and peaks which can be deconvoluted) or a combination of both.

5.1 Metabolism research and metabolite engineering

The GC-MS methodologies developed at the Max Planck Institute for Molecular Plant Physiology (Fiehn *et al.*, 2000a; Roessner *et al.*, 2000) have been extensively applied to studies in the area of plant metabolic engineering. The now routine ability to transform genes into most plant species of commercial interest have given rise to numerous projects which

aim to establish or alter the synthesis of high value compounds (Fernie *et al.*, 2002; Galili and Hofgen, 2002; Sweetlove *et al.*, 2003). Our limited understanding of metabolic networks means that there are often unpredicted consequences when single pathway steps are engineered. The consequent need to obtain rapid and non-biased information about the extent of changes in metabolic engineering projects has been a major driver in the development of metabolite profiling.

This type of application can be illustrated with respect to the apparently simple pathway of starch biosynthesis (Roessner *et al.*, 2001a). The authors applied the GC-MS profiling method to developing potato tubers of lines that over-express a yeast invertase, specifically in the tuber (Trethewey *et al.*, 1998, 1999b). This transgenic modification was performed in anticipation that enhanced cleavage of sucrose would lead to more starch accumulation. However, the GC-MS profiling revealed an alteration in partitioning away from starch biosynthesis and towards glycolysis through an unknown regulatory mechanism. In addition, the non-biased profiling provided many surprises. For example, there was a significant accumulation of shikimate and a large reduction in the inositol content, observations which were not anticipated given the focus of the project on starch biosynthesis. Thus wider profiling provides the opportunity to discover unexpected perturbations in the metabolic network. This work was extended to a thorough analysis of a wide range of lines altered in starch biosynthesis (Roessner *et al.*, 2001a, 2001b, 2002). Cluster analysis was performed to distinguish groups within the metabolite profiles and the interpretation of the clustering patterns led to a range of new insights into the regulation of primary metabolism in potato tubers. As successful metabolic engineering will require precise and specific alterations in flux, metabolite profiling is emerging as an indispensable tool for metabolic engineering projects (Trethewey, 2004).

To date, LC-MS metabolite profiling has not yet been applied to metabolic engineering projects targeting primary metabolism. However, more targeted forms of LC-MS profiling have seen a range of applications in secondary metabolism research. For example, metabolite profiling has recently contributed to the elucidation of the specificities of some key enzymes involved in the pathway of lignin biosynthesis (Guo *et al.*, 2001; Meyermans *et al.*, 2000). Similarly, targeted LC-MS profiling has recently been established for isoprenoids in peppermint (Lange *et al.*, 2001) and alkaloids in camptothecin producing plants (Yamazaki *et al.*, 2003a, 2003b).

One of the challenges of studying metabolism is that it is highly compartmented within and between cells (Arlt *et al.*, 2001). There is thus considerable interest to develop metabolite profiling protocols for small sample volumes from highly differentiated sources. To date there have been

three examples of this type of GC-MS profiling all in the context of plant research. Farré *et al.* (2001) have coupled GC-MS to a non-aqueous fractionation procedure for the separation of vacuolar, plastidic and cytosolic components from potato tuber tissue. Fiehn (2003) studied phloem exudates (some 400 components in GC-TOF were distinguished with deconvolution algorithms) whilst Morris *et al.* (2004) reported on the analysis of around 60 constituents of loblolly pine xylem tissue.

5.2 Functional genomics

Because metabolites are direct indicators of phenotype and function, metabolite profiling has received considerable attention as a highly valuable approach for functional genomics (Fiehn, 2002; Glassbrook and Ryals, 2001; Trethewey *et al.*, 1999a). The group of Oliver who were studying genetic function and redundancy in yeast were one of the first to realize the potential importance of "metabolomics" in this context (Oliver, 1997). This group proceeded to develop an approach that they termed FANCY (functional analyses by co-response in yeast) which involved disrupting genes with unknown function and then performing paired comparisons of metabolite levels, with particular reference to the glycolytic pathway (Raamsdonk *et al.*, 2001). However, it was plant researchers that first provided a demonstration of the power of wide, non-biased, metabolite profiling in the context of functional genomics. Using GC-MS metabolite profiling Fiehn *et al.* (2000a) analyzed four distinct genotypes of *Arabidopsis* including two known mutants. Using principal component analysis they were able to demonstrate that the profiles of each genotype gave rise to distinct clusters. Although this work did not actually lead to the assignment of gene function it clearly showed the potential for classifying phenotype and for coupling this information to genotypic data.

The next logical step is to utilize metabolite profiling to perform an "annotation" of gene function. The high throughput GC-MS and LC-MS platform at metanomics (Figure 2) has been deployed to do just this. This group has used *Arabidopsis* to create both loss-of-function populations through selecting for single loci homozygous T-DNA lines and gain-of-function populations through the systematic over-expression of all individual genes from a genome (e.g. yeast, *E. coli* genomes). These populations of more than 200,000 individual plant lines have been subjected to GC-MS and LC-MS profiling. Genes which give rise to changes in particular metabolites are selected and studied further. In general, such approaches are able to link genes to individual metabolite changes or shifts in entire pathways as well as characterizing the overall response of the plant metabolic network to perturbation through addition or loss of gene activity.

Parallel to the commercial interest in using metabolite profiling as a functional genomics tool, a range of public programs are being established (Mazur, 2003; Sumner *et al.*, 2003). For example, metabolite profiling is performed within the Genome *Arabidopsis* Resource Network (GARNet) in the UK, whilst the Samuel Roberts Noble Foundation has targeted a model legume, *Medicago truncatula* (May, 2002).

5.3 Medical, nutritional and pharmaceutical research

Whilst the origins of metabolite profiling were in clinical diagnostics and the largest body of literature, particularly for NMR technologies, is related to diagnostic studies (Lindon *et al.*, 2004) there are still relatively few publications on wide comprehensive profiling with GC-MS and LC-MS in medical and pharmaceutical research. Nevertheless there is a widespread conviction that metabolomics, including GC-MS and LC-MS profiling, will play a major role in drug discovery and development in the future (Watkins and German, 2002). From contemporary conference presentations it is evident that a range of public institutions and companies are currently establishing GC-MS and LC-MS profiling in medical and pharmaceutical applications. Our experience at *metanomics* has shown that it is readily possible to adapt profiling platforms which were developed for other applications to sample types relevant for these fields (e.g. Figure 1). A recent paper showed that it is possible to use LC-MS profiling of urine to differentiate control rats from those that had been administered candidate pharmaceuticals thus providing an early indication of this potential (Plumb *et al.*, 2003b).

The possible application of metabolite profiling to the nutritional sciences has raised much interest as it may open up new avenues to investigate the relationship between food, nutrition, health and genotype (German *et al.*, 2003; van Ommen and Stierum, 2002). In particular, the technology has the potential to identify metabolite markers indicative for the early onset of diseases and to develop prevention schemes. One of the attractions of such metabolomics approaches is that they can be applied both to the food material as well as to samples from the human participants in such studies. Knowledge that is generated through such investigations may be used to guide the development of nutritionally optimized crops through breeding or plant biotechnology (Watkins *et al.*, 2001).

6. SUMMARY AND FUTURE PROSPECTS

GC-MS and LC-MS methodologies are rapidly improving and will certainly play a major role in metabolite profiling for the foreseeable future. We anticipate that there will be continued innovation in instrument technology and software and we will probably see the emergence of integrated solutions for metabolomics applications. Study of the metabolome is increasingly realized to be a central component of contemporary biology and we anticipate a rapid increase in the diversity of metabolite profiling applications. Critically, GC-MS and LC-MS based profiling will play a continued role in functional genomics and will increasingly be integrated into systems biology. However, perhaps the most rapid growth will be seen in applications related to medical, toxicological or pharmaceutical research. The possibility of nutrigenomics to foster an integrated understanding of the role of small molecules in plant and animal systems is certainly exciting. Plant biotechnology, through breeding or engineering of healthier crop plants, may follow strategies built upon such knowledge. Thus metabolite profiling is a rare example of a technology with broad reaches across multiple, diverse, scientific fields.

REFERENCES

Aharoni A, Ric de Vos CH, Verhoeven HA, Maliepaard CA, Kruppa G, Bino R and Goodenowe DB. Nontargeted metabolome analysis by use of Fourier Transform Ion Cyclotron Mass Spectrometry. *OMICS*, 6: 217–234 (2002).

Annesley TM. Ion suppression in mass spectrometry. *Clin. Chem.*, 47: 1041–1044 (2003).

Arlt K, Brandt S and Kehr J. Amino acid analysis in five pooled single plant cell samples using capillary electrophoresis coupled to laser-induced fluorescence detection. *J. Chromatogr. A*, 926: 319–325 (2001).

Dallüge J, Beens J and Brinkman UAT. Comprehensive two-dimensional gas chromatography: a powerful and versatile analytical tool. *J. Chromatogr. A*, 1000: 69–108 (2003).

Davidson W and Frego L. Micro-high-performance liquid chromatography/Fourier transform mass spectrometry with electron-capture dissociation for the analysis of protein enzymatic digests. *Rapid Commun. Mass Spectrom.*, 16: 993–993 (2002).

Farré EM, Tiessen A, Roessner U, Geigenberger P, Trethewey RN and Willmitzer L. Analysis of the compartmentation of glycolytic intermediates, nucleotides, sugars, organic acids, amino acids, and sugar alcohols in potato tubers using a nonaqueous fractionation method. *Plant Physiol.*, 127: 685–700 (2001).

Fernie AR, Willmitzer L and Trethewey RN. Sucrose to starch: a transition in molecular plant physiology. *Trends Plant Sci.*, 7 :35–41 (2002).

Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN and Willmitzer L. Metabolite profiling for plant functional genomics. *Nat. Biotechnol.*, 18: 1157–1161 (2000a).

Fiehn O, Kopka J, Trethewey RN and Willmitzer L. Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal. Chem.*, 72: 3573–3580 (2000b).

Fiehn O. Metabolomics. The link between genotype and phenotype. *Plant Mol. Biol.*, 48: 155–171 (2002).

Fiehn O. Metabolic networks of *Cucurbita maxima* phloem. *Phytochemistry*, 62: 875–886 (2003).

Galili G and Hofgen R. Metabolic engineering of amino acids and storage proteins in plants. *Metab. Eng.*, 4: 3–11 (2002).

German JB, Roberts MA and Watkins SM. Personal metabolomics as a next generation nutritional assessment. *J. Nutr.*, 133: 4260–4266 (2003).

Glassbrook A and Ryals J. A systematic approach to biochemical profiling. *Curr. Opin. Plant Biol.*, 4: 186–190 (2001).

Guo D, Chen F, Inoue K, Blount JW and Dixon RA. Downregulation of caffeic acid 3-O-methyltransferase and caffeoyl CoA 3-O-methyltransferase in transgenic alfalfa. impacts on lignin structure and implications for the biosynthesis of G and S lignin. *Plant Cell*, 13: 73–88 (2001).

Hayen H and Karst U. Strategies for the liquid chromatographic-mass spectrometric analysis of non-polar compounds. *J. Chromatogr. A*, 1000: 549–565 (2003).

Horning EC and Horning MG. Metabolic profiles: gas-phase methods for analysis of metabolites. *Clin. Chem.*, 17: 802–809 (1971).

Hsieh Y, Merkle K, Wang G, Brisson JM and Korfmacher WA. High-performance liquid chromatography-atmospheric pressure photoionization/tandem mass spectrometric analysis for small molecules in plasma. *Anal. Chem.*, 75: 3122–3127 (2003).

Huhman DV and Sumner LW. Metabolic profiling of saponins in *Medicago sativa* and *Medicago trunculata* using HPLC coupled to an electrospray ion-trap mass spectrometer. *Phytochemistry*, 59: 347–360 (2003).

Lange BM, Ketchum REB and Croteau RB. Isoprenoid biosynthesis. Metabolite profiling of peppermint oil gland secretory cells and application to herbicide target analysis. *Plant Physiol.*, 127: 305–314 (2001).

Lindon JC, Holmes E and Nicholson JK. Metabonomics and its role in drug development and disease diagnosis. *Expert Rev. Mol. Diagn.*, 4: 189–199 (2004).

Luque-Garcia JL and Luque de Castro MD. Ultrasound-assisted Soxhlet extraction: an expeditive approach for solid sample treatment - Application to the extraction of total fat from oleaginous seeds. *J. Chromatogr. A*, 1034: 237–242 (2004).

Martin SE, Shabanowitz J Hunt DF and Marto JA. Subfemtomole MS and MS/MS peptide sequence analysis using nano-HPLC micro-ESI Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.*, 72: 4266–4274 (2000).

Matuszewski BK, Constanzer ML and Chavez-Eng CM. Strategies for the assessment of matrix effect in quantitative bioanalytical methods based on HPLC-MS/MS. *Anal. Chem.*, 75: 3019–3030 (2003).

May GD. An integrated approach to Medicago functional genomics, in: *Recent Advances in Phytochemistry, Vol. 36*, J. T. Romeo, and R. A. Dixon, eds., Pergamon, Oxford, UK, pp. 179–195 (2002).

Mazur BJ. Plant metabolite profiling en route to destination. *Nat. Biotechnol.*, 21: 875–876 (2003).

Meyermans H *et al.* Modifications in lignin and accumulation of phenolic glucosides in poplar xylem upon down-regulation of caffeoyl-coenzyme A O-methyltransferase, an enzyme involved in lignin biosynthesis. *J. Biol. Chem.*, 275: 36899–36909 (2000).

Morris CR, Scott JT, Chang HM, Sederoff RR, O'Malley D and Kadla JF. Metabolic profiling: a new tool in the study of wood formation. *J. Agric. Food Chem.*, 52: 1427–1434 (2004).

Nicholson JK, Connelly J, Lindon JC and Holmes E. Metabonomics: a platform for studying drug toxicity and gene function. *Nat. Rev. Drug Discov.*, 1: 153–161 (2002).

Niessen WMA. Natural products and endogenous compounds, in: *Liquid Chromatography-Mass Spectrometry*, 2nd ed., Marcel Decker, New York, pp. 465–500 (1999).

Niessen WMA. Progress in liquid chromatography-mass spectrometry instrumentation and its impact on high-throughput screening. *J. Chromatogr. A*, 1000: 413–436 (2003).

Niwa T. Metabolic profiling with gas chromatography-mass spectrometry and its application to clinical medicine , *J. Chromatogr.*, 20: 313–345 (1986).

Oliver SG. Yeast as a navigational aid in genome analysis. *Microbiology*, 143: 1483–1487 (1997).

Plumb RS, Stumpf CL, Gorenstein MV, Castro-Perez JM, Dear GJ, Anthony M, Sweatman BC, Connor SC and Haselden JN. Metabonomics: the use of electrospray mass spectrometry coupled to reverse-phase liquid chromatography shows potential for the screening of rat urine in drug development. *Rapid Commun. Mass Spectrom.*, 16: 1991–1996 (2002).

Plumb RS, Granger J, Stumpf C, Wilson ID, Evans JA and Lenz EM. Metabonomic analysis of mouse urine by liquid-chromatography-time of flight mass spectrometry (LC-TOFMS): detection of strain, diurnal and gender differences. *Analyst*, 128: 819–823 (2003a).

Plumb RS, Stumpf CL, Granger J, Castro-Perez JM, Haselden JN and Dear GJ. Use of liquid chromatography/time-of-flight mass spectrometry and multivariate statistical analysis shows promise for the detection of drug metabolites in biological fluids. *Rapid Commun. Mass Spectrom.*, 17: 2632–2638 (2003b).

Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh MC, Beerden JA, Brindle KM, Kell DB, Rowland JJ, Westerhoff HV, van Dam K and Oliver SG. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.*, 19: 45–50 (2001).

Richter BE, Ezzell JL, Felix D, Roberts KA and Later DW. An accelerated solvent extraction system for the rapid preparation of environmental organic compounds in soil. *Am. Lab.*, 27: 24–28 (1995).

Roepenack-Lahaye Ev, Degenkolb T, Zerjeski M, Franz M, Roth U, Wessjohann L, Schmidt J, Scheel D and Clemens S. Profiling of arabidopsis secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. *Plant Physiol.*, 134: 548–559 (2004).

Roessner U, Wagner C, Kopka J, Trethewey RN and Willmitzer L. Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant J.*, 23: 131–142 (2000).

Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, Willmitzer L and Fernie AR. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell.*, 3: 11–29 (2001a).

Roessner U, Willmitzer L and Fernie AR. High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. *Plant Physiol.*, 127: 749–764 (2001b).

Roessner U, Willmitzer L and Fernie AR. Metabolic profiling and biochemical phenotyping of plant systems. *Plant Cell Rep.*, 21: 189–196 (2002).

Rostagno MA, Palma M and Barroso CG. Ultrasound-assisted extraction of soy isoflavones. *J. Chromatogr. A*, 1012: 119–128 (2003).

Santos FJ and Galceran MT. Modern developments in gas chromatography-mass spectrometry-based environmental analysis. *J. Chromatogr. A*, 1000: 125–151 (2003).

Sauter H, Lauer M and Fritsch H. Metabolic profiling of plants - A new diagnostic technique, in: *Synthesis and Chemistry of Agrochemicals II*, D. R. Baker, J. G. Fenyes, and Moberg WK, eds., American Chemical Society, Washington, DC, pp. 288–299 (1991).

Schwartz JC, Senko MW and Syka JE. A two-dimensional quadrupole ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.*, 13: 659–669 (2002).

Stolker AL, Niesing W, Fuchs R, Vreeken RJ, Niessen WM and Brinkman UAT. Liquid chromatography with triple-quadrupole and quadrupole-time-of-flight mass spectrometry for the determination of micro-constituents - a comparison. *Anal. Bioanal. Chem.*, 378: 1754–1761 (2004).

Sumner LW, Mendes P and Dixon RA. Plant metabolomics: large-scale phytochemistry in the functional genomics area. *Phytochemistry*, 62: 817–836 (2003).

Sweetlove LJ, Last RL and Fernie AR. Predictive metabolic engineering: a goal for systems biology. *Plant Physiol.*, 132: 420–425 (2003).

Taylor J, King RD, Altmann T and Fiehn O. Application of metabolomics to plant genotype discrimination using statistics and machine learning. *Plant J.*, 18: 241–248 (2002).

Tolstikov VV and Fiehn O. Analysis of highly polar compounds of plant origin: Combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Anal. Biochem.*, 301: 298–307 (2002).

Tolstikov VV, Lommen A, Nakanishi K, Tanaka N and Fiehn O. Monolithic silica-based capillary reversed-phase liquid chromatography/electrospray mass spectrometry for plant metabolomics. *Anal. Chem.*, 75: 6737–6740 (2003).

Trethewey RN. Metabolite profiling as an aid to metabolic engineering in plants. *Curr. Opin. Plant Biol.*, 7: 196–201 (2004).

Trethewey RN, Geigenberger P, Riedel K, Hajirezaei MR, Sonnewald U, Stitt M, Reismeier JW and Willmitzer L. Combined expression of glucokinase and invertase in potato tubers leads to a dramatic reduction in starch accumulation and a stimulation of glycolysis. *Plant J.*, 15: 109–118 (1998).

Trethewey RN, Krotzky AJ and Willmitzer L. Metabolic profiling: a Rosetta stone for genomics? *Curr. Opin. Biotechnol.*, 2: 83–85 (1999a).

Trethewey, R. N., Reismeier, J. W., Willmitzer, L., Stitt, M., and Geigenberger, P. Tuber specific expression of a yeast invertase and a bacterial glucokinase in potato leads to an activation of sucrose phosphate synthase and the creation of a futile cycle. *Planta*, 208: 227–238 (1999b).

van Ommen B and Stierum R. Nutrigenomics: Exploiting systems biology in the nutrition and health arena. *Curr. Opin. Biotechnol.*, 13: 517–521 (2002).

Wagner C, Sefkow M and Kopka J. Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry*, 62: 887–900 (2003).

Watkins SM and German JB. Metabolomics and biochemical profiling in drug discovery and development. *Curr. Opin. Mol. Ther.*, 4: 224–228 (2002).

Watkins SM, Hammock BD, Newman JW and German JB. Individual metabolism should guide agriculture toward foods for improved health and nutrition. *Am. J. Clin. Nutr.*, 74: 283–286 (2001).

Wink M. Plant breeding: importance of plant secondary metabolites for protection against pathogens and herbivores. *Theor. Appl. Genet.*, 75: 225–233 (1988).

Yamazaki M, Nakajima J, Yamanashi M, Sugiyama M, Makita Y, Springob K, Awazuhara M and Saito, K. Metabolomics and differential gene expression in anthocyanin chemovarietal forms of *Perilla frutescens*. *Phytochemistry*, 62: 987–995 (2003a).

Yamazaki Y, Urano A, Sudo H, Kitajima M, Takayama H, Yamazaki M, Aimi N and Saito K. Metabolite profiling of alkaloids and strictosidine synthase activity in camptothecin producing plants. *Phytochemistry*, 62: 461–470 (2003b).

Zubritsky E. The best of both worlds with LC/FTMS. *Anal. Chem.*, 72: 633A (2000).

Chapter 8

THE APPLICATION OF ELECTROCHEMISTRY TO METABOLIC PROFILING

David F. Meyer, Paul H. Gamache and Ian N. Acworth.
ESA Inc. 22 Alpha Road, Chelmsford, MA 01824

1. INTRODUCTION

Electrochemistry is one of the most sensitive and versatile techniques available for the study of biomolecules (Kissinger and Heineman, 1996; Brajter-Toth and Chambers, 2002). This discussion will focus on the use of 3-electrode electrochemical (EC) flow cells for controlled-potential applications. This encompasses some of the most common uses of electrochemistry, including amperometric and coulometric detection for HPLC (LCEC) (Acworth *et al.*, 1997; Sabbioni *et al.*, 2004; Gonzalez de la Huebra *et al.*, 2003; Riis, 2002) and hydrodynamic voltammetry (Nagels *et al.*, 1989). These techniques all involve the use of EC cells to produce oxidative and / or reductive (redox) reactions of analyte in flowing solutions. LCEC is well known for its selectivity and sensitivity. A logical extension of these reaction-based techniques is the coupling of EC with analytical devices such as mass spectrometry (MS) and nuclear magnetic resonance spectroscopy (NMR). When in parallel, EC may be used to complement the analytical capabilities of MS and NMR (Gamache *et al.*, 2004). When in series, EC may be used to analyze species that can be studied with these and other structurally informative devices (Zhou and Van Berkel, 1995; Jurva *et al.*, 2003; Hayen and Karst, 2003). This chapter will focus on quantitative and qualitative aspects of multivariate LCEC approaches to metabolic profiling and will briefly discuss synthetic applications of EC as a potential tool for metabolite characterization.

There are many excellent sources of information on theory and applications of controlled-potential techniques (Kissinger and Heineman, 1996; Chen *et al.*, 1996; Lacourse, 2001; Lund and Baizer, 1991). Briefly, a 3-electrode EC cell includes reference (RE), working (WE) and auxiliary (AE) electrodes that comprise a circuit that is controlled by a potentiostat. The cell functions by establishing a specific applied potential (E_{appl}), or voltage difference, between RE and WE. The AE, through electrical feedback via the potentiostat, provides the energy necessary to establish and maintain E_{appl} and relies on solution conductivity to complete the circuit. Supporting electrolyte (e.g., >20mM buffer) is typically incorporated in the mobile phase to provide sufficient conductivity for optimal functioning of the circuit. E_{appl} thus drives redox reactions of solution-phase species at the WE.

Whether or not a metabolite is 'redox active' depends critically upon its structure and also on the conditions (e.g., pH, solvent properties). For a given condition, characteristics of the WE and E_{appl} are two of the primary determinants. EC cells with carbon-based WE are the primary focus of this discussion. Other WE materials such as noble metals (e.g., Au, Ag and Pt) have surface properties that are advantageous for specialized applications (e.g., Au WE for carbohydrate detection (Rocklin, 1984; Bowers, 1991)). These WE often take part in the redox reaction (e.g., through complexation) and, as a result, the WE itself may be gradually consumed (Rocklin, 1984). Also, when used for electro-oxidation, noble metal WE often form oxide layers that gradually renders their surface less active (Neuburger and Johnson, 1987). Carbon-based WE, by contrast, typically serve as relatively inert electron donors, which is dependent on E_{appl}. These WE are relatively resistant to surface oxide effects, are typically not consumed as part of analyte electrolysis, and offer a relatively wide useable potential window for many solvent and pH conditions (Rocklin, 1984). For these reasons, carbon-based WE are the most widely used for LCEC.

There are several possible EC flow cell designs, but only 3 basic geometries are in general use. Thin-layer and wall-jet amperometric cells have small surface area WEs and, when using normal bore HPLC flow rates (i.e., 0.2 to 2.0mL/min.), only a small percentage (typically <5%) of analyte comes into close enough proximity of the WE to be oxidized or reduced. Response is therefore affected by changes in flow rate and active WE surface area. A decrease in response due to WE passification is commonly observed with these small WE (Catarino *et al.*, 2003) and maintenance is frequently required. While this design is primarily applicable to relatively pure samples, an advantage is that cell volume can be relatively easily scaled to accommodate microbore and capillary LC methods. These cells are thus in widespread use for targeted analysis of volume-limited sample types, such

as microdialysis perfusates (Leis *et al.*, 2004). However, these cells are rarely used for complex multivariate analysis due to the above considerations, their limited resolving capacity and difficulties encountered when using gradient elution techniques.

A third class of EC flow cell utilizes high surface area micro-porous graphitic carbon WE. With this design, close to 100% of analyte mass comes into close enough contact with the WE surface to be electrolyzed when using normal bore flow rates. This design is termed "coulometric" since the integrated response (peak area) represents the total charge transferred (coulombs) from the reaction as in Faraday's law. Advantages to this design include stable and reproducible response and low susceptibility to loss of response due to WE contamination (Sabbioni *et al.*, 2004; Gonzalez de la Huebra *et al.*, 2003; Riis, 2002; Masuda *et al.*, 1997). A disadvantage is that coulometric EC cells are not as easily scaled for low flow applications.

The most significant advantage of coulometric EC is obtained when multiple coulometric EC cells are used in series (*viz.*, EC-Array (Matson *et al.*, 1984)). A further advantage relates to the high synthetic yield of reaction products. Given these advantages, the discussion below will focus on two general approaches with coulometric EC technology. First, the use of EC-Array alone or in parallel with MS for metabolic profiling will be discussed. Quantitative and qualitative characteristics and practical aspects will be exemplified by recent studies of xenobiotic toxin exposure in animals. The relevance of this technique to multivariate (e.g., 'omic') studies of gene and protein levels of biomolecular organization will also be discussed. Second, the use of decoupled EC flow cells in series with MS ion sources will be described as a potential tool in metabolite structural elucidation studies.

2. METABOLIC PROFILING WITH LC-EC-ARRAY

The general concepts of EC-Array are described in detail elsewhere (Gamache *et al.*, 1999; Acworth and Gamache, 1996). Briefly this technique employs up to 16 series EC cells with porous graphitic carbon-based WE. Each cell is typically poised at a different fixed potential, thereby spanning a wide potential window to allow detection of a broad scope of redox active metabolites. Efficient electrolysis obtained with high surface area WE allows selective detection and resolution of co-eluting analytes, based on differences in their relative ease of oxidation and / or reduction. Hydrodynamic voltammograms (HDV) for each redox active metabolite are obtained from the response across adjacent EC-Array sensors. These data are a reflection of the kinetic and thermodynamic components of electron transfer reactions with demonstrated use for peak identification and peak

purity assessment. LC-EC-Array has been widely used for analysis of redox active substances with primary *in-vivo* application to clinical (Cheng *et al.*, 1991; Gamache *et al.*, 1993; Gamache *et al.*, 1999), neuroscience (LeWitt *et al.*, 1992; Beal *et al.*, 1992; Volicer *et al.*, 1985), and redox biochemistry (Yanagawa *et al.*, 2001; Christen *et al.*, 2002; Hensley *et al.*, 1997; Hensley *et al.*, 2000; Acworth *et al.*, 1999; Collins *et al.*, 1998; Sofic *et al.*, 1992; Kristal *et al.*, 2002). Multi-component analyses are based on potential-dependent selectivity, fmole sensitivity and data-dependent acquisition (auto-ranging) which facilitates use with gradient elution and provides a 10^5 dynamic response range (Gamache *et al.*, 1999; Ferruzzi *et al.*, 1998).

2.1 Xenobiotic toxicity studies with parallel EC-array-MS

The following data, although from different experiments, are described within the context of an overall scheme that progresses from exploratory to targeted analysis. This includes: a) generating HPLC EC-Array-MS multivariate fingerprints with rapid (6 to 15min.) LC analyses; b) pattern recognition analysis to investigate the presence of natural sample groupings; and c) interrogating stored EC and MS data to characterize metabolite peaks, or peak patterns that significantly contribute to interesting sample groupings.

2.1.1 Analytical conditions

Data presented are based on reversed-phase HPLC with water/ acetonitrile binary gradient and various conditions (e.g., flow rate, column, gradient profiles). Either an Agilent 1100 or ESA LC system was used for solvent and sample delivery. Post-column flow was split between EC-Array (CoulArray®, ESA Inc) and MS detectors (Agilent 1100 MSD and Sciex QTRAP) with split ratios optimized for MS performance (see Figure 1). Specific conditions are described with results and practical considerations of coupling EC-Array with MS (e.g., requirements for supporting electrolyte, grounding of fluidic lines, flow rates) are described in more detail elsewhere (Zhou and Van Berkel, 1995; Gamache *et al.*, 2004).

2.1.2 Biological samples

Examples shown include analysis of urine samples from adult rats administered with single doses of the xenobiotic compounds: acetaminophen (APAP), acetylsalicylic acid, maleic acid and chloroethanamine. Samples from animal studies were kindly provided by Dr. Timothy Maher (Mass College of Pharmacy, Boston MA) and by Dr. Elaine Holmes (Imperial

College, London, UK), and details of the animal experimental protocols are described elsewhere (Holmes *et al.*, 2000; Nicholls *et al.*, 2001).

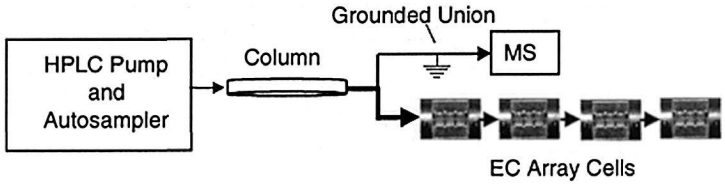


Figure 1. Diagram of LC system with parallel EC-Array and MS detection.

2.1.3 Results and discussion

A typical chromatogram obtained from urine analyzed by parallel EC-Array - MS is shown in Figure 2. From the analysis of 22 rat urine samples, inclusive of several treatment groups, an average of 168 ± 34 distinct redox active species (peak clusters) were detected within a 12-minute chromatographic run time (signal to noise ratio > 10). Differences in voltammetric behavior (i.e., relative ease of oxidation) among metabolites served as the basis for resolution of co-eluting species. EC-Array response corresponded to mass quantities of approximately 50 pg to 1 μg on column, based on calibration data from representative standards. Using the described conditions (Figure 2 legend), this corresponds to a urinary concentration range of approximately 25 ng/mL to 500 μg/mL, which provides some evidence of the dynamic range and sensitivity (e.g., 10^2 to 10^3 times better limits of detection than UV/vis absorbance (Ferruzzi *et al.*, 1998; Gamache *et al.*, 1999)) of EC-Array detection.

Since chemical structure is a critical determinant of a metabolite's redox behavior, the intrinsic generation of voltammetric data with EC-Array (i.e., response across channels) provides qualitative information for each redox active species. In general, EC reactions were observed according to the following general rank order (by relative ease of oxidation): *o*, *p*-quinol and *o*, *p*-aminophenol > tertiary amine > *m*-quinol ≈ phenol ≈ arylamine > secondary amine ≈ thiol > thioether ≠ primary amines, aliphatic alcohols. These HDV data were thus useful to track and normalize these complex profiles and to provide some indication of possible functional groups for a given unknown metabolite.

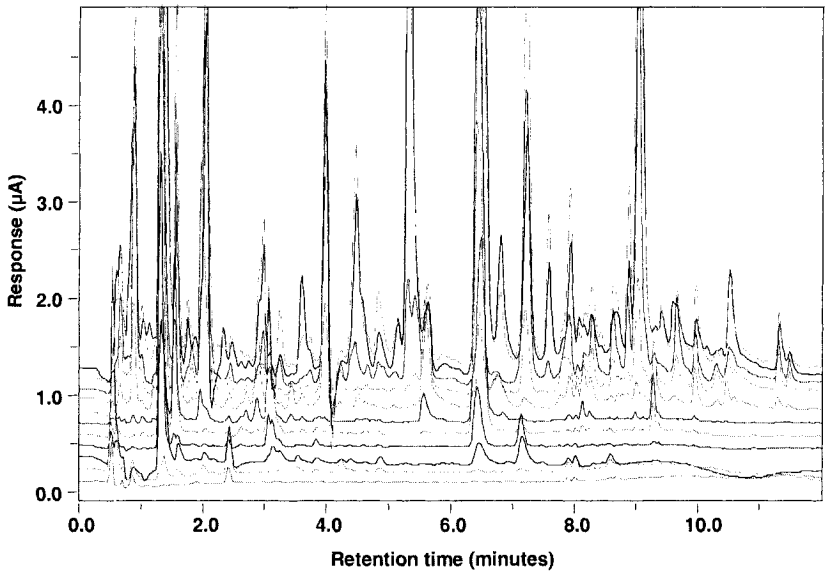


Figure 2. Representative EC-Array chromatogram (12 of 16 channels shown) from 20 μL injection of 10-fold diluted rat urine. Gradient elution 1% to 100% aqueous acetonitrile with 10 mM ammonium formate and 50 mM formic acid; flow rate 1.5 mL/min; Shiseido C18, 3μm, 75 mm x 4.6 mm i.d. column; 4:1 passive post-column flow split to EC-Array: MS, respectively. EC-Array potentials were 0 to 1050 mV in increments of 70 mV and data from ESI-MS, positive mode, scan range m/z 50-850 was acquired in parallel.

The combined use of MS and EC-array resulted in highly complementary detection. For example, the observation of a particular redox active metabolite peak allowed a more informed and targeted interrogation of corresponding MS data. Furthermore, our results suggest that many redox active urinary metabolites exist as solution phase neutral species under a variety of reversed-phase chromatographic conditions. For example, some prominent redox active urinary metabolites detected by EC-Array (e.g., ascorbic, uric, 5-hydroxyindoleacetic and homovanillic acids) were not detected by MS using various combinations of ESI, APCI, positive and negative ionization, neutral or acidic mobile phase conditions and even with targeted selected ion monitoring. The combined use of MS and EC-Array therefore has the potential to enhance the capabilities of MS and to provide broader coverage of the metabolome.

2.1.4 Pattern recognition analysis

We have focused on the use of EC-Array data for pattern recognition analyses. MS data were initially used to help distinguish xenobiotic metabolites and subsequently to characterize specific variables, revealed from chemometric analyses. A CoulArray® (ESA Inc., Chelmsford, MA) software utility was used to adjust for chromatographic variability followed by conversion of otherwise raw EC-Array data into a generic format for pattern recognition analysis. This allowed rapid data processing - typically < 5 minutes for 100 samples. Subsequent exploratory pattern recognition analysis was performed using Pirouette® (Infometrix, Inc., Seattle, WA).

In a model study of APAP-induced hepatotoxicity, results from principal components analysis (PCA) showed consistent differentiation (Figure 3A and B) of high dose APAP (200 and 300 mg/kg, 0-8 hr collection) from control, low dose (20 mg/kg) APAP, and high dose (200 mg/kg) acetylsalicylic acid. Differences were observed after exclusion of xenobiotic metabolite variables and PCA results were qualitatively similar (Figure 3A *vs* 3B), even when using different analytical conditions (i.e., different mobile phase pH and gradient). This is evidence of the robust nature of these small molecule redox profiles in differentiating the effects of this hepatotoxin.

High dose APAP is believed to result in toxicity via oxidative metabolic activation to form reactive N-acetyl-p-benzoquinoneimine (NAPQI), which can bind to macromolecules and also lead to production of reactive oxygen species. In this study, changes in endogenous metabolite profiles associated with a single high dose of APAP were clearly evident. Redox active metabolite peaks with significant contribution to the sample groupings, shown in Figure 3D, were inferred from the corresponding PCA loadings plots. HDV data for an endogenous metabolite, which was lower in high-dose APAP samples *vs* controls is shown in Figure 3E. These data suggest that this peak may possess a hydroxyindole, hydroxypurine or methoxycatechol structure, but additional EC-Array-MS studies are required.

While endogenous metabolites were of primary interest, variables associated with APAP metabolism provide a good example of the complementary nature of EC and MS. Both MS and EC-Array data provided evidence that peak M3 (Figure 4) consisted of two major components (*m/z* 232, oxidation potential (E_{ox}) 840mV and *m/z* 313, E_{ox} 600mV). The higher E_{ox} observed with *m/z* 232 suggests phenol substitution while the lower E_{ox} with *m/z* 313 implies an intact amidophenol structure.

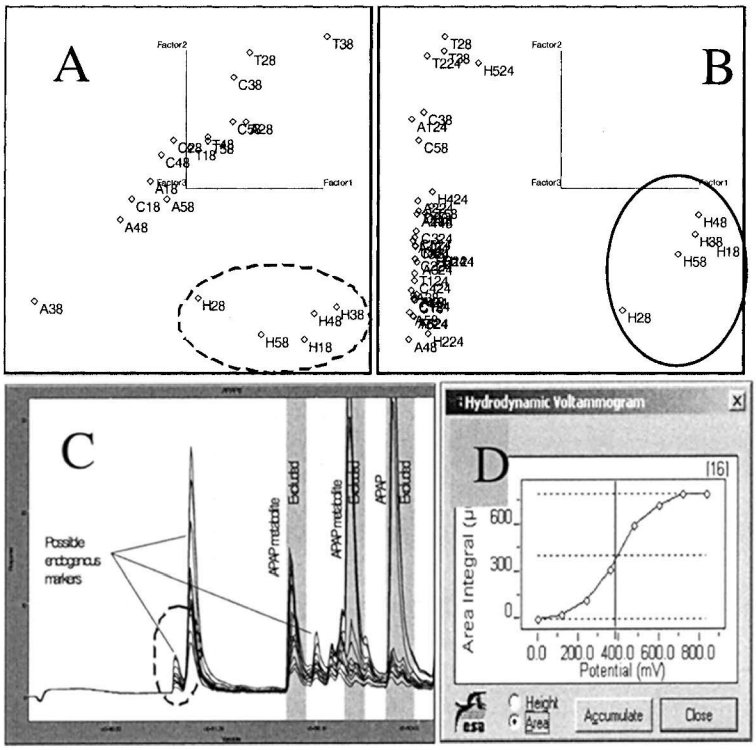


Figure 3. A) PCA scores plot of Factor 1 (65.8% of total variance) vs. Factor 2 (10.1% of total variance) showing separate grouping of high-dose APAP (H18 – H58) samples based on EC-Array profiles using a gradient of 1 to 80% acetonitrile in 6 min., pH 7. B) PCA scores plot of Factor 1 (78.9% of total variance) vs. Factor 2 (13.1% of total variance) from EC-Array profiles using a gradient of 1 to 10% acetonitrile in 6 min. then to 100% in 2min, pH3.9. C) Overlay of EC-Array data from 3 treatment categories (20mg/kg APAP, 200mg/kg APAP and control) with some differentiating features indicated. D) Hydrodynamic voltammogram of one possible endogenous marker peak.

These voltammetric and MS data are consistent with APAP-sulfate and APAP-mercapturate (APAP-M) structures, respectively. The presence of APAP-M, a known urinary marker of reactive quinoneimine species formation, provides evidence of the oxidative pathway associated with toxicity of this xenobiotic. Importantly, in our investigations this metabolite was only noticed when *both* EC-Array and MS data were interrogated.

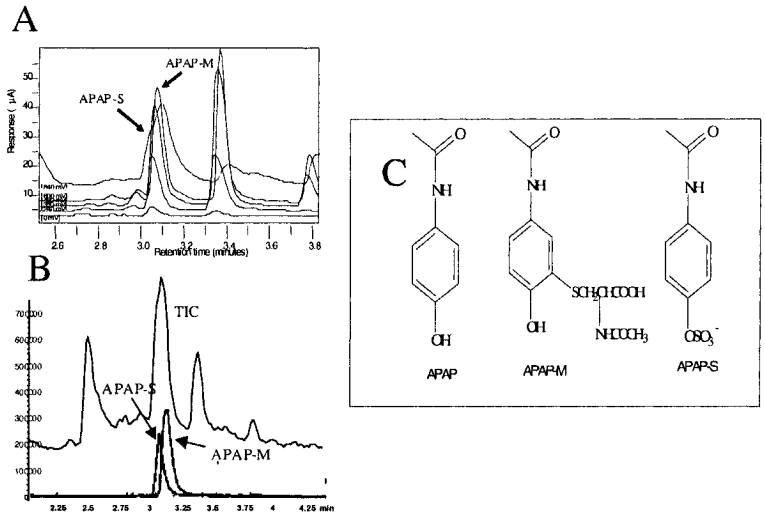


Figure 4. A) EC-Array response from 0-840 mV showing evidence of two constituents with differing voltammetric response. B) Overlay of total ion and extracted ion MS chromatograms showing evidence of sulfate (APAP-S, *m/z* 232) and mercapturate (APAP-M, *m/z* 313) metabolites of APAP. C) Structures of APAP, APAP-M, APAP-S.

In a similar example, urine obtained from rats exposed to renal toxins, maleic acid (MA) and chloroethanamine (CE) were analyzed by LC-EC-Array-MS. Histopathological data showed maximal toxic response in the second day after receiving MA (300 mg/kg) and CE (750 mg/kg), respectively, with complete recovery by the fifth day. A PCA scores plot (see Figure 5) showed clear differentiation between profiles obtained from MA and CE treated (MA and CE, Day 2) as compared to controls, recovered, and subtoxic-dosed animals. Repeated LC/EC/MS analyses using several chromatographic conditions resulted in very similar PCA results. These results were in good qualitative agreement with NMR-based metabonomic analysis of these samples (personal communication, Dr. Elaine Holmes, Imperial College, London, UK).

2.2 What is being measured electrochemically?

Our data demonstrate high informational content with EC-Array detection of > 100 urinary metabolites and some reports of several hundred in a single chromatographic run also exist (Shi *et al.*, 2002). Our ongoing

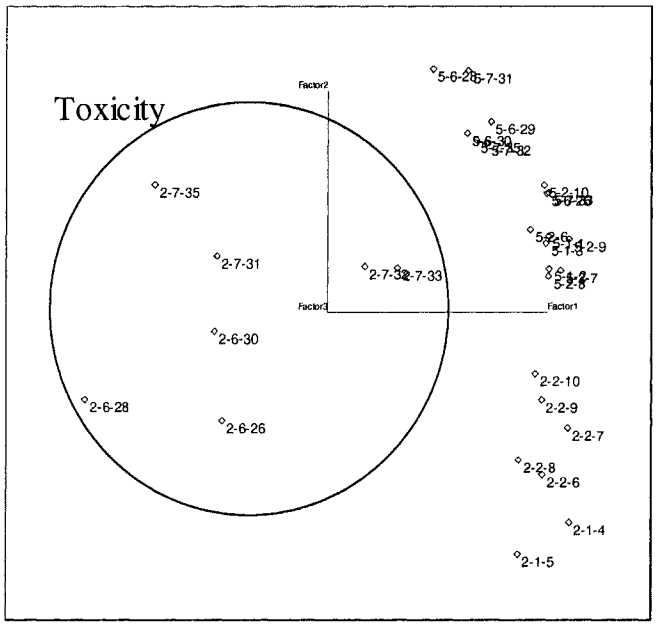


Figure 5. PCA scores plot of Factor 1 (73.6% of total variance) *vs* Factor 2 (8.2% of total variance) generated from EC-Array data showing differentiation of sample groups having evidence of toxin-induced renal histopathology (encircled) from control and recovered animals.

studies involve the use of voltammetric and MS data along with metabolic knowledge bases (e.g. KEGG) and compound libraries to provide peak annotation. Within the working potential range of carbon-based cells, electron transfer reactions can occur for a limited range of chemical structures, each compound reacting at a relatively specific applied potential. In general, oxidation reactions occur if the gain or loss of charge can be stabilized, for example, through π-electron delocalization. Common redox-active structures include aryl alcohols, aryl amines, secondary and tertiary aliphatic amines, sulfides and conjugated polyenes, aryl nitro and quinoid species. The scope of carbon-based EC therefore spans a significant range of endogenous metabolites including many antioxidants, co-factors, hormones, neurotransmitters, peptides and vitamins. Table 1 is a brief list of redox active chemical or metabolite classes for which carbon-based LCEC detection has been used for targeted quantitative bioanalysis.

Metabolites not detected with carbon-based LCEC methods include many carbohydrates, lipids and amino acids that lack these redox active substituents. However, there are hundreds of known endogenous metabolites

Table 1. Representative Redox Active Metabolites.

Chemical Structure or Metabolite Class	Examples	Comments
Amino acids	Cysteine, methionine, tryptophan, tyrosine	Peptides, numerous metabolites including biogenic amines, aminothiols, etc.
Aryl amine	Kynurenine, 3-hydroxykynurenine	Indole pathway implicated in many neurological disorders
Biogenic amine	Dopamine, epinephrine, norepinephrine, serotonin	Transmitters and Hormones
Chroman	α-, β-, δ-, γ-tocopherol, 5-nitro-γ-tocopherol	Anti-oxidants and markers of oxidative stress
Hormone and metabolite	Estrogens, thyroxines	Hormones, implicated in many processes including hormonal carcinogenesis
Conjugated polyenes	Carotenoids, retinoids, vitamin D	Vitamins and antioxidants, numerous functions
Pterin and Pteridine	Folates, biopterins	Co-factors, 1-carbon metabolism
Purines	Guanine, 8-hydroxy, 2'-deoxyguanosine, hypoxanthine, uric acid	Nucleic acid bases, markers of DNA damage
Pyridine	Pyridoxal, pyridoxine	Vitamins, many functions
Quinone/hydroquinone	Coenzyme Q10, plastoquinone, vitamin K1	Numerous functions
Sulfide	Glutathione, homocysteine	Numerous functions, marker of folate, B12, B6 deficiencies

that retain the redox active structure of primary metabolites and building blocks such as the electroactive amino acids (cysteine, methionine, tyrosine, tryptophan), purines, estrogens, folates, retinoids, pyridoxals and kynurenines.

2.3 Relevance to genomics and proteomics

There is an enormous range of multivariate approaches in biological studies for which the word "profiling" or the suffix "omics" is used as a descriptor. Genomic and proteomic approaches, for example, can range from purely exploratory and hypothesis-generating to highly directed studies, towards the testing of a well-developed hypothesis. The relevance of EC-Array to genomics and proteomics is therefore very dependent on the investigational and practical aspects (e.g., sample type, throughput requirements) of a given approach. As previously described, LC-EC-Array has been widely applied as a medium-to-low throughput technique for a

wide range of metabolic profiling studies. This includes multivariate analysis in plants and cell systems and in tissue or bio-fluids from higher organisms.

Of particular relevance is the selectivity of EC-Array toward redox active substances. Redox processes are very important biochemical processes. Many enzymes have redox centers that catalyze electron transfer reactions involving both endogenous and xenobiotic molecules through a variety of mechanisms. Radical and non-radical reactive species are also generated as part of normal redox metabolism. Many studies suggest that redox metabolism of a wide range of chemical structures leads to formation of reactive electrophiles, which act in a diverse array of toxic processes that typically involve covalent binding or other modifications to small and large molecules (e.g. DNA, proteins, peptides, lipids), redox cycling, antioxidant / scavenger depletion, and other elements of oxidative stress. Biological systems are always being challenged by pro-oxidants and other reactive species and there is a complex and redundant protective system involving prevention, detoxification, clearance and repair. The production of reactive species becomes particularly relevant during pathological processes, including drug toxicity and disease. A metabolic imbalance that favors the production of pro-oxidant species over antioxidant protection is often referred to as oxidative stress. This condition is heavily implicated as a causal, adaptive or ancillary factor in most disease and toxic processes, for example, being a key aspect in immune and inflammatory response.

The concept of redox regulation, broadly defined as biological response to maintain homeostasis against oxidative stress, is highly relevant to the use of EC-Array in the context of genomics and proteomics. There is much interest in the relationships between cellular redox status and networks involving signal transduction, alterations in gene expression, etc., particularly with respect to immune and inflammatory responses. The applicability of LC-EC-Array to redox active species (e.g., hormones, neurotransmitters, antioxidants, markers of oxidative stress) has therefore led to its widespread use to study oxidative metabolism and redox biochemical processes including those related to aging (Cadenas *et al.*, 1997; Yanagawa *et al.*, 2001), immune response (Bugianesi *et al.*, 2000), inflammation (Christen *et al.*, 2002, Hensley *et al.*, 1997), and many pathological processes (Collins *et al.*, 1998; Hensley *et al.*, 1997; Sofic *et al.*, 1992; Russell *et al.*, 1992; Beal *et al.*, 1992; Acworth, 2003)). The sensitivity, resolving capacity and qualitative information obtainable with combined EC-Array and MS detection may provide a particularly powerful tool for a wide variety of metabolic profiling studies.

3. SERIAL EC-MS FOR SYNTHESIS AND CHARACTERIZATION

The use of upstream decoupled EC flow cells in series with MS has been previously described as a means of synthesizing and characterizing potential drug metabolites (Deng and Van Berkel, 1999; Jurva et *al.*, 2000). Our studies have also included the synthesis of endogenous compounds using endogenous metabolite precursors (Gamache *et al.*, 2003).

Extracted ion voltammograms (Figure 6) show MS ion abundance as a function of potential associated with EC oxidation of estrogen metabolites in the presence of glutathione. Our data indicate that the most abundant ions

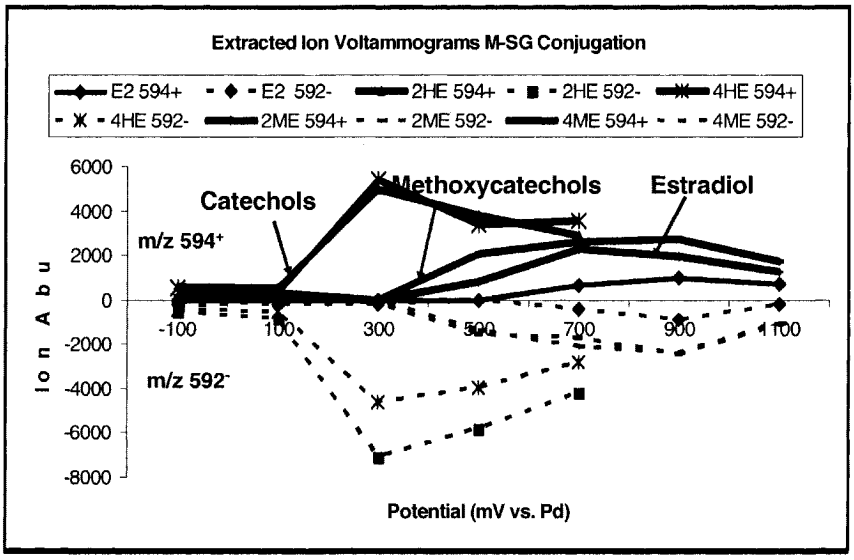


Figure 6. Extracted ion voltammograms, *m/z* 594 from positive ion electrospray (ESI) MS and *m/z* 592 from negative ion ESI-MS corresponding to protonated and deprotonated catecholestrogen glutathionyl adducts obtained by flow injection analysis of various estrogen metabolites and upstream EC oxidation in the presence of glutathione.

produced from estradiol (E), 2- and 4-hydroxyestradiol (2HE and 4HE) and 2- and 4-methoxyestradiol (2ME and 4ME) all correspond to reactive electrophilic quinone and catecholestrogen glutathionyl conjugates (CE-SG). EC reactions proceeded at specific potentials including aromatic hydroxylation of E at 1000mV, O-dealkylation of 2ME and 4ME at 600mV, and dehydrogenation of 2HE and 4HE at 300mV. These results demonstrate that this technique is capable of very closely simulating the proposed

biotransformation reactions related to estrogen-dependent carcinogenesis (Devanesan *et al.*, 2001). Furthermore, EC reactions may be carried out before an LC column to enable on-line separation, purification and analysis of reaction products. Recent studies have demonstrated the feasibility of scaling up to produce sufficient quantities for structural confirmation by NMR (Gamache *et al.*, 2004). The simplicity and speed of on-line EC-LC-MS may thus provide an effective means of characterizing some of the many unknown metabolites encountered in multivariate profiling of endogenous metabolites.

4. CONCLUSION

The concurrent acquisition of EC-Array and MS data showed several advantages in exploratory multivariate profiling including broader coverage of the chemical diversity and concentration range of endogenous metabolites. The qualitative information from both techniques was useful for data normalization, peak purity and structural elucidation studies. Chemometric analysis of raw EC-Array profiles demonstrated the ability to reproducibly differentiate sample groups consistent with xenobiotic-induced histopathological changes. As many organic chemicals are thought to exert toxicity via redox processes, the acquired redox profiles may be particularly useful for tissue and regio-specific modeling, diagnostic marker identification and mechanistic insight to xenobiotic-induced toxicity.

The feasibility of electrochemically synthesizing endogenous metabolites has been demonstrated for estrogen metabolites including those proposed as potential biomarkers of hormonal carcinogenesis. The simplicity of forming reaction products on-line with LC-MS using the same conditions as biological sample analysis provides a potentially efficient means of characterizing some of the many unknown metabolites encountered in metabolic profiling studies.

ACKNOWLEDGEMENTS

The authors thank Dr. Elaine Holmes, Imperial College, London, UK and Dr. Timothy Maher, Massachusetts College of Pharmacy, Boston, MA for their helpful correspondence and for providing biological samples used in the described studies.

REFERENCES

Acworth IN. *Handbook of Redox Biochemistry,* (2003).

Acworth IN *et al.* Estimation of hydroxyl free radical levels *in vivo* based on liquid chromatography with electrochemical detection. *Methods Enzymol*, 300: 297-313 (1999).

Acworth IN and Gamache PH. The coulometric electrode array for use in HPLC analysis: Part 1:Theory. *American Laboratory*, (1996).

Acworth IN *et al. Progress in HPLC-HPCE: Coulometric electrode array detectors for HPLC,* VSP, Utrecht, The Netherlands. (1997).

Beal MF *et al.* Kynurenic acid concentrations are reduced in Huntington's disease cerebral cortex. *J Neurol Sci.*, 108: 80-87 (1992).

Bowers ML. A new analytical cell for carbohydrate analysis with a maintenance-free reference electrode. *J. Pharma. Biomed. Anal.*, 9: 1133-1137 (1991).

Brajter-Toth A and Chambers JQ *Electroanalytical Methods for Biological Materials,* Marcel Dekker, Inc., New York. (2002).

Bugianesi R *et al.* High-performance liquid chromatography with coulometric electrode array detector for the determination of quercetin levels in cells of the immune system. *Anal. Biochem.*, 284: 296-300 (2000).

Cadenas S *et al.* Oxidative DNA damage estimated by oxo8dG in the liver of guinea-pigs supplemented with graded dietary doses of ascorbic acid and alpha-tocopherol. *Carcinogenesis*, 18: 2373-2377 (1997).

Catarino RI *et al.* Flow amperometric determination of pharmaceuticals with on-line electrode surface renewal. *J. Pharm. Biomed. Anal.*, 33: 571-580 (2003).

Chen JG, Woltman SJ and Weber SG. Electrochemical detection of biomolecules in liquid chromatography and capillary electrophoresis. *Adv. Chromatogr.*, 36: 273-313 (1996).

Cheng MH *et al.* Automated analysis of urinary VMA, HVA, and 5-HIAA by gradient HPLC using an array of eight coulometric electrochemical detectors. *Lab. Robot. Automat.*, 4: 297-303 (1991).

Christen S *et al.* Analysis of plasma tocopherols alpha, gamma, and 5-nitro-gamma in rats with inflammation by HPLC coulometric detection. *J. Lipid Res.*, 43: 1978-1985 (2002).

Collins AR *et al.* Oxidative DNA damage measured in human lymphocytes: large differences between sexes and between countries, and correlations with heart disease mortality rates. *Faseb J.*, 12: 1397-1400 (1998).

Deng H and van Berkel GJ. A thin-layer electrochemical flow cell coupled online with electrospray-mass spectrometry for the study of biological redox reactions. *Electroanalysis*, 11: 857-865 (1999).

Devanesan P *et al.* Catechol estrogen metabolites and conjugates in mammary tumors and hyperplastic tissue from estrogen receptor-alpha knock-out (ERKO)/Wnt-1 mice: implications for initiation of mammary tumors. *Carcinogenesis*, 22: 1573-1576 (2001).

Ferruzzi MG *et al.* Carotenoid determination in biological microsamples using liquid chromatography with a coulometric electrochemical array detector. *Anal. Biochem.*, 256: 74-81 (1998).

Gamache P, Freeto SM and Acworth IN. Coulometric array HPLC analysis of lipid soluble vitamins and antioxidants. *Amer. Clin. Lab.* (1999).

Gamache P *et al.* Metabolomic Applications of Electrochemistry / Mass Spectrometry. *J. Amer. Assoc. Mass Spectrom. (Submitted).*

Gamache P *et al.* ADME/Tox Profiling Using Coulometric Electrochemistry and Electrospray Ionization Mass Spectrometry. *Spectroscopy*, 18: 14-21 (2003).

Gamache P *et al.* Rapid on-line electrochemical synthesis of pharmaceutical degradants and metabolites for profiling, identification and quantitation. Poster presentation, *Pittcon* Chicago, IL (2004).

Gamache PH, Kingery ML and Acworth IN. Urinary metanephrine and normetanephrine determined without extraction by using liquid chromatography and coulometric array detection. *Clin. Chem.*, 39: 1825-1830 (1993).

Gonzalez de la Huebra MJ, Bordin G and Rodriguez AR. Comparative study of coulometric and amperometric detection for the determination of macrolides in human urine using high-performance liquid chromatography. *Anal. Bioanal. Chem.*, 375: 1031-1037 (2003).

Hayen H and Karst U. Analysis of phenothiazine and its derivatives using LC/electrochemistry/MS and LC/electrochemistry/fluorescence. *Anal. Chem.*, 75: 4833-4840 (2003).

Hensley K *et al.* Quantitation of protein-bound 3-nitrotyrosine and 3,4-dihydroxyphenylalanine by high-performance liquid chromatography with electrochemical array detection. *Anal. Biochem.*, 251: 187-195 (1997).

Hensley K, Williamson KS and Floyd RA. Measurement of 3-nitrotyrosine and 5-nitro-gamma-tocopherol by high-performance liquid chromatography with electrochemical detection. *Free Radic. Biol. Med.*, 28: 520-528 (2000).

Holmes E *et al.* Chemometric models for toxicity classification based on NMR spectra of biofluids. *Chem. Res. Toxicol.*, 13: 471-478 (2000).

Jurva U, Wikstrom HV and Bruins AP. *In vitro* mimicry of metabolic oxidation reactions by electrochemistry/mass spectrometry. *Rapid. Commun. Mass Spectrom.*, 14: 529-533 (2000).

Jurva U *et al.* Comparison between electrochemistry/mass spectrometry and cytochrome P450 catalyzed oxidation reactions. *Rapid. Commun. Mass Spectrom.*, 17: 800-810 (2003).

Kissinger PT and Heineman WR. *Laboratory Techniques in Electroanalytical Chemistry,* Marcel Dekker, New York. (1996).

Kristal BS, Vigneau-Callahan K and Matson WR. Simultaneous analysis of multiple redox-active metabolites from biological matrices. *Methods Mol. Biol.*, 186: 185-194 (2002).

Lacourse WR. Electrochemical detectors: functional group analysis. *Enantiomer*, 6: 141-152 (2001).

Leis S *et al.* Catecholamine release in human skin-a microdialysis study. *Exp. Neurol.*, 188: 86-93 (2004).

LeWitt PA *et al.* Markers of dopamine metabolism in Parkinson's disease. The Parkinson Study Group. *Neurology*, 42: 2111-2117 (1992).

Lund H and Baizer MM *Organic electrochemistry, an introduction and guide,* Marcel Dekker, New York. (1991).

Masuda, S, *et al.* A novel high-performance liquid chromatographic assay for vitamin D metabolites using a coulometric electrochemical detector. *J. Pharm. Biomed. Anal.*, 15: 1497-1502 (1997).

Matson WR *et al.* n-Electrode three-dimensional liquid chromatography with electrochemical detection for determination of neurotransmitters. *Clin. Chem.*, 30: 1477-1488 (1984).

Nagels LJ, Mush G and Massart DL. Rapid-scan hydrodynamic voltammetry and cyclic voltammetry of pharmaceuticals in flow injection analysis conditions. *J. Pharm. Biomed. Anal.*, 7: 1479-1483 (1989).

Neuburger GG and Johnson DC. Comparison of the pulsed amperometric detection of carbohydrates at gold and platinum electrodes for flow injection and liquid chromatographic systems. *Anal. Chem.*, 59: 203-204 (1987).

Nicholls AW *et al.* Metabonomic investigations into hydrazine toxicity in the rat. *Chem. Res. Toxicol.*, 14: 975-987 (2001).

Riis B. Comparison of results from different laboratories in measuring 8-oxo-2'-deoxyguanosine in synthetic oligonucleotides. *Free Radic. Res.*, 36: 649-659 (2002).

Rocklin RD. Working-electrode materials. *LC*, 2: 588-594 (1984).

Russell IJ *et al.* Cerebrospinal fluid biogenic amine metabolites in fibromyalgia/fibrositis syndrome and rheumatoid arthritis. *Arthritis Rheum.*, 35: 550-556 (1992).

Sabbioni C *et al.* Simultaneous liquid chromatographic analysis of catecholamines and 4-hydroxy-3-methoxyphenylethylene glycol in human plasma. Comparison of amperometric and coulometric detection. *J. Chromatogr. A*, 1032: 65-71 (2004).

Shi H *et al.* Characterization of diet-dependent metabolic serotypes: proof of principle in female and male rats. *J. Nutr.*, 132: 1031-1038 (2002).

Sofic E *et al.* Reduced and oxidized glutathione in the substantia nigra of patients with Parkinson's disease. *Neurosci. Lett.*, 142: 128-130 (1992).

Volicer L *et al.* Serotoninergic system in dementia of the Alzheimer type. Abnormal forms of 5-hydroxytryptophan and serotonin in cerebrospinal fluid. *Arch. Neurol.*, 42: 1158-1161 (1985).

Yanagawa K *et al.* Changes in antioxidative mechanisms in elderly patients with non-insulin-dependent diabetes mellitus. Investigation of the redox dynamics of alpha-tocopherol in erythrocyte membranes. *Gerontology*, 47: 150-157 (2001).

Zhou F and van Berkel GJ. Electrochemistry combined online with electrospray mass spectrometry. *Anal. Chem.*, 67: 3643-3649 (1995).

Chapter 9

DIFFERENTIAL METABOLIC PROFILING FOR BIOMARKER DISCOVERY

A mass spectrometric approach

Haihong Zhou, Aaron B. Kantor and Christopher H. Becker
SurroMed, Inc.,1430 O'Brien Drive, Menlo Park, CA 94025, USA

1. INTRODUCTION

Effective biomarkers are urgently needed in a range of diseases for early and accurate diagnosis, and for monitoring disease progression and the effects of therapeutic intervention. A biomarker is defined as an attribute that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention (GROUP, 2001). Biomarkers (Figure 1) range from simple low molecular weight molecules such as sugars, fatty acids, steroids and free-floating peptides, to soluble proteins and cell surface proteins, to complex integrated properties. Established metabolite biomarkers include cholesterol and glucose for monitoring the risk of heart disease and diabetes, respectively. Established protein markers include HDL/LDL and glycated hemoglobin (HbA1c) for the same indications. Cellular markers include Her2 for breast cancer and CD4 T cells for immune function in diseases like HIV infection. Biomarker discovery can benefit from a range of technology platforms that are most appropriate for evaluating different classes of analytes.

Advances in genomics allow quantification of changes in the expression of every gene within a genome. Subsequent developments in proteomics permit relatively detailed investigation of proteins within a proteome (McDonald and Yates, 2002; Wu and MacCoss, 2002). However, the emerging field of metabolomics needs substantial effort. Metabolomics refers to a global analysis of the entire complement of endogenous

metabolites (the metabolome) in cells, tissues or fluids. Quantifying the changes in metabolite concentrations due to disease events or therapeutic agents should prove valuable in identifying new biomarkers.

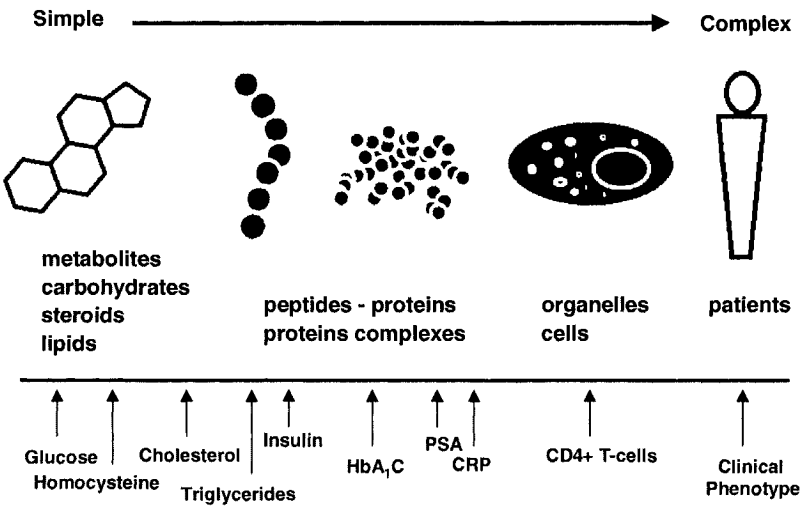


Figure 1. Examples of biomarkers used today. Biomarkers can range from simple molecules to complex integrated properties.

Metabolites are end products of biological processes, and changes in their expression level may provide insight into disease mechanisms. Many investigators have employed the first two strategies, genomics and proteomics, in the discovery of biomarkers (Adam *et al.*, 2001; He and Chiu, 2003; Krieg *et al.*, 2002; Pang *et al.*, 2002; Tugwood *et al.*, 2003; Vernon *et al.*, 2002; Wang *et al.*, 2003). This chapter focuses on the third strategy, namely, metabolomics and metabolic profiling. Pros and cons of various metabolic profiling approaches are discussed with the emphasis on hyphenated mass spectrometric methods where mass spectrometry is directly coupled to a separation method. General strategies for biomarker discovery and validation are discussed and apply regardless of which technologies are being used. Specific examples are included to illustrate the enormous potential and the challenges of this field in biomarker discovery.

2. APPROACHES TO METABOLIC PROFILING

Traditional metabolite research focuses only on a few metabolites or a class of metabolites at one time, where analytical techniques can be optimized for best detection and quantification. In contrast, metabolic profiling in the context of biomarker discovery aims to simultaneously detect and quantify all metabolites in a given biological system. Challenges in achieving this goal include the thousands of compounds in the metabolome (Beecher, 2003), their structural diversity and complexity, and their wide range of concentrations. A variety of techniques have been used in metabolic profiling including nuclear magnetic resonance (NMR) spectroscopy, mass spectrometry (MS), electrochemical detection (ECD), and optical spectroscopy such as infrared (IR), Raman spectroscopy and ultraviolet/visible (UV/VIS) spectroscopy. These techniques have been reviewed in the previous volume (Harrigan and Goodacre, 2003).

NMR spectroscopy and MS are the two dominating approaches because of the information-rich data sets they produce in terms of number of components (100's to 1000's), relative quantification and information on chemical identity. For example, high resolution NMR spectroscopy has been applied to the search for toxicity markers, an effort pioneered by Nicholson's group in Imperial College (Beckonert *et al.*, 2003; Griffin *et al.*, 2001; Lenz *et al.*, 2000; Nicholson *et al.*, 2001; Scarfe *et al.*, 2000; Warne *et al.*, 2000; Waters *et al.*, 2002). Applications include the study of the effect of environmental stressors on organism health (Viant *et al.*, 2003) and diagnosis of the presence and severity of coronary heart disease (Brindle *et al.*, 2002). There are several excellent reviews (Lindon *et al.*, 2004a; 2004b). The advantage of NMR over MS is that NMR requires little or no sample preparation (Nicholson and Wilson, 2003) and is a non-destructive method. Additionally, the recent development of high-resolution magic-angle spinning NMR spectroscopy permits metabolic profiling of intact tissue samples (Garrod *et al.*, 2001). Unfortunately, sensitivity still is a major drawback of this technique (low micrograms of total material at the best for routine analysis), a restriction limiting investigations to abundant metabolites. New developments with the cryoflow probe and extremely high field magnets (up to 900 MHz) improve the sensitivity 15-20 fold (Spraul *et al.*, 2003; van der Greef *et al.*, 2004), which is still several orders of magnitude less than that of other techniques, such as mass spectrometry.

There are more sensitive detection systems such as the electrochemical detection (ECD), which can detect picogram leves of redox active molecules (Kaddurah-Daouk *et al.*, 2004), and laser induced fluorescence (Bonato, 2003; Huck *et al.*, 2003; Stobiecki and Makkar, 2004). However, these techniques produce significantly less information per sample in terms of

number of compounds and chemical identity, compared to the data from NMR and MS.

Mass spectrometry is often chosen as a primary metabolic profiling method because a wide range of metabolites are readily ionized in MS with a detection limit of nanograms or less. Furthermore, hyphenated mass spectrometric methods, such as gas chromatography-mass spectrometry (GC-MS), liquid chromatography-mass spectrometry (LC-MS) and capillary electrophoresis-mass spectrometry (CE-MS) provide separation of biological complex mixtures, leading to a higher overall coverage of the metabolome. GC-MS was first used to study human metabolites in 1970 (Horning and Horning, 1970; Zlatkis and Liebich, 1971) and since then it has been used rountinely in metabolite research.

GC-MS is suitable for analyzing small volatile and thermally stable metabolites. Non-volatile polar metabolites can be converted to less-polar, volatile and thermally-stable compounds by derivatization. Due to high energy electron impact ionization, most of ions observed in GC-MS are fragmented ions, which helps determine the structure of a metabolite. However, parent ions rarely seen in the GC-MS spectra can be a problem for structure elucidation of unknown metabolites. Although chemical ionization, as an alternative, can produce pseudo-molecular ions, its limitation is low sensitivity.

The introduction of electrospray ionization (ESI) in the 1980s (Whitehouse *et al.*, 1985; Yamashita and Fenn, 1984) enabled many metabolites to be analyzed at one time using LC-MS without the need for derivatization. Moreover, most of the ions produced with this soft ionization technique are molecular ions rather than fragments. Elemental composition of observed ions is readily available owing to recent advancements in mass accuracy and mass resolution.

Metabolites are usually quantified in mass spectrometry by comparison to spiked standards. However, when measurements require tracking hundreds to thousands of biomolecules, this strategy becomes impractical. We have developed a method which works without the need for spiked standards or isotopic labeling (Wang *et al.*, 2003). There have been concerns that non-linearities and ion suppression effects of electrospray ionization would lead to poor quantification, especially for complex mixtures like serum. For these reasons, we conducted a study to demonstrate that direct quantification of relative differences among samples by LC-MS can be achieved (Wang *et al.*, 2003). Linear or near-linear quantification is achieved for both the proteome and metabolome for thousands of serum components. Repeat processing of serum samples demonstrated good analytical reproducibility with median CVs of 26 and 24 per cent for the proteome and metabolome respectively. Substantial on-line chromatographic

separation is often used in this approach to reduce the complexity of the mixture. Moreover, similar samples are compared (e.g. serum *vs* serum), so that the chemical matrix and any ion suppression effects are similar between samples.

Despite the fact that no single technology is able to investigate the entire metabolome simutaneously, there are practical constraints on what and how many technologies can be employed, such as available sample volume, throughput, robustness and operating costs.

3. IMPORTANT ASPECTS IN BIOMARKER DISCOVERY BY METABOLIC PROFILING USING MASS SPECTROMETRY

A general strategy for biomarker discovery involves comparison of metabolic profiles from different groups (e.g. disease *vs* healthy control) and quantification of relative changes in metabolite concentration between groups. Comparisons are made without any *a priori* assumptions about the metabolites that are different between groups. Molecular identification is made subsequently and is initially achieved by comparing with established metabolite libraries. Some components are not identified initially. Large differences and low p-values between groups drive the identification of these components by tandem MS and other methods.

Generally, biomarker discovery proceeds with the sequence starting with a study design in which the type and number of samples, technology platform, and statistical methods are defined. The samples are often easily accessible body fluids (blood or urine) from either humans or animals. In some cases, other fluids (e.g. cerebrospinal and synovial fluid) or tissue extracts are appropriate because it is expected that the concentration of disease-relevant metabolites is higher in these samples than in plasma or urine.

After samples are collected and processed, LC-MS and GC-MS are performed. Usually, raw data from analytical instruments requires software processing to derive a list of components and their intensities for each sample. Here components represent the common monoisotopic peaks found in multiple samples.

Statistical methods are then applied to derive a list of components that show significant differences. The next step is to identify the structure of these components and confirm the relevance of these components to the disease by checking known pathways or databases. The final stage in biomarker discovery is the validation of candidate biomarkers.

The following section discusses several important aspects involved in biomarker discovery by metabolic profiling using mass spectrometry.

3.1 Sample collection and handling

A sample of reliable quality is of paramount importance in biomarker discovery. Any systematic bias among cohorts related to sample collection and down-stream handling can lead to wrong conclusions or simply add too much noise to the study results. In addition, clinical samples may have been acquired without a metabolomics perspective in mind. For example, in the case of blood collection, fasting blood is preferred to reduce short-term variations due to metabolism. Standardization for physical activity and collection time are also potentially useful. Moreover, the type of blood tubes used directly affects impurities that are introduced into the metabolome mixture. If possible, the number of freeze-thaw cycles should be minimized. Therefore, to ensure sample quality, conditions of sampling, transportation, pre-treatment (e.g. coagulation) and storage should be documented and well controlled (Bischoff and Luider, 2004).

3.2 Sample preparation

Metabolic profiling of body fluids such as blood plasma usually involves an initial simplification step to remove high molecular weight proteins, which interfere in the detection of metabolites. Their removal can be achieved by organic precipitation, ultrafiltration, or size exclusion chromatography. For GC-MS, the low molecular weight portion is chemically derivatized to improve thermal stability and ionization efficiency. In general, there is no need for chemical modification for ESI-MS analysis; however, removal of biological salts from the sample can greatly improve the robustness and sensitivity of the mass spectrometric detection. In the case of LC-MS, reverse-phase columns are often preferred to couple with ESI-MS since the mobile phase is directly compatible with the requirement for electrospray ionization. Polar metabolites that are not retained on reverse-phase columns can be separated using normal-phase chromatography or hydrophilic-interaction chromatography (Schlichtherle-Cerny *et al.*, 2003; Tolstikov and Fiehn, 2002). CE-MS provides extremely high resolution for charged species and can be an alternative when sample size is limited (Soga *et al.*, 2003). The main drawback of CE is its low loading capacity. Generally, by using a combination of multiple sample preparation methods, overall coverage of the metabolome can be extensive. However, there is a balance among the amount of information that is

retrieved per sample, personnel and equipment required, time consumed, cost and sample losses over processing steps prior to MS analysis.

3.3 Mass spectrometry instrumentation

A variety of mass spectrometers have been used in metabolic profiling research including the triple-quadrupole mass spectrometer, ion-trap mass spectrometer, time-of-flight mass spectrometer (TOF-MS), and Fourier-transform mass spectrometer (FT-MS). Profiling complex mixtures requires mass spectrometric system to have high sensitivity, mass accuracy, mass resolution and a wide dynamic range. High-mass accuracy and resolution are particularly useful to resolve peaks that overlap in mass and retention time, which is often the case for profiling body fluids.

Figure 2 presents how high-mass accuracy with a time-of-flight mass spectrometer can facilitate peak determination. One hundred femtomole of angiotensin III peptide was spiked into 100 μL of a human serum metabolome mixture and analyzed by LC-ESI-TOF MS. The retention time of angiotensin III peptide was 27.4 min with a dominant triply charged protonated ion (theoretical *m/z* 306.5051). A peak from the serum metabolome sample with similar mass was eluted at 34.6 min. By narrowing the mass window from 1 Da (top) to 0.01 Da (bottom), the mass chromatogram of angiotensin III peptide was more and more distinguishable. The neighboring peak that previously dominated the whole chromatogram dropped significantly. In this case, the two peaks were well separated in retention time. For those that cannot be well separated in retention time, high mass accuracy is clearly advantageous.

Fourier-transform mass spectrometry can be an alternative for metabolic profiling due to its exceptional resolution and mass accuracy. In the past, the cost, robustness and the difficulty in coupling with separation techniques have limited its use in large-scale metabolic profiling research. It is foreseeable that FT-MS will find a wider application as commercial instruments become more accessible to the research community (Aharoni *et al.*, 2002).

3.4 Data processing and quantification

The lack of commercially available algorithms and software tools for mass spectrometric data handling and processing has hampered the application of mass spectrometry, especially LC-MS, in metabolic profiling. Although in general MS data processing steps are technique dependent, they can be summarized into following steps: (1) detection of peaks; (2) correction of shifts (e.g. mass and retention time); (3) normalization of

intensities; and (4) construction of component lists. Here a MassView™ software developed in SurroMed, Inc. (Hastings *et al.*, 2002; Wang *et al.*, 2003) is discussed as an example to illustrate various steps involved in LC-MS data processing.

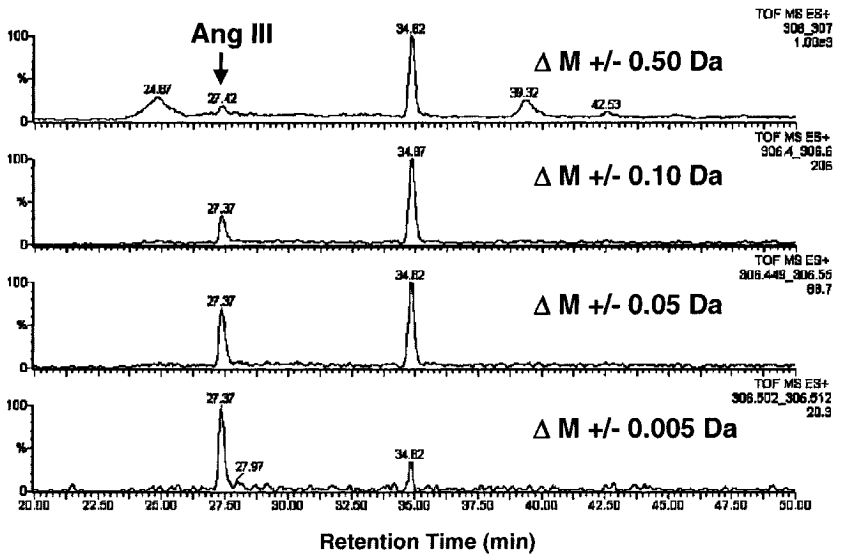


Figure 2. Mass-selected chromatograms of angiotensin III peptide (theoretical m/z 306.5051) with mass window of 1 Da (top), 0.2 Da, 0.1 Da and 0.01 Da (bottom), respectively. One hundred femtomole of angiotensin III peptide was spiked into 100 μL of pooled human serum metabolome and analyzed by LC-ESI TOF mass spectrometer. The peak at 34.8 min was from human serum.

For peak detection, the software first performs baseline subtraction and data smoothing for each mass spectrum at a given elution time. A vectorized peak detection algorithm (Hastings *et al.*, 2002) which considers both mass and retention time dimensions is used to identify valid spectral peaks in the presence of noise sources, notably chemical noise. Next, the isotopic pattern of the peaks is assigned and an intensity value recorded for the monoisotopic peak. Any mass shift due to environmental conditions or instrumental factors is then adjusted using an internal calibrant. Using an arbitrarily chosen reference sample, the LC-MS retention time shifts between samples are adjusted by dynamic time warping (Wang *et al.*, 2003). This method corrects both linear and non-linear shifts in retention time. Peak intensity normalization is then performed. One file is chosen as a reference and all other files are normalized relative to it. The median value for the intensity

ratio of a set of spectral peaks between the files is used as the normalization constant. Monoisotopic peaks found to correspond to each other in multiple samples become assigned as common components of the study. Note that no spiked standards are required in this methodology.

The MassView™ software is also used to visualize the LC-MS data. Figure 3 shows the richness of the human serum metabolome. The sample was prepared from 50 µL of human serum using organic precipitation and 10 µL of which were analyzed by reverse-phase LC-MS in a two-hour run. Roughly 2000 monoisotopic peaks (components) are detected with intensity threshold of 35 ion counts. Although there are components below 35 ions counts, a conservative threshold results in a more reliable and robust analysis.

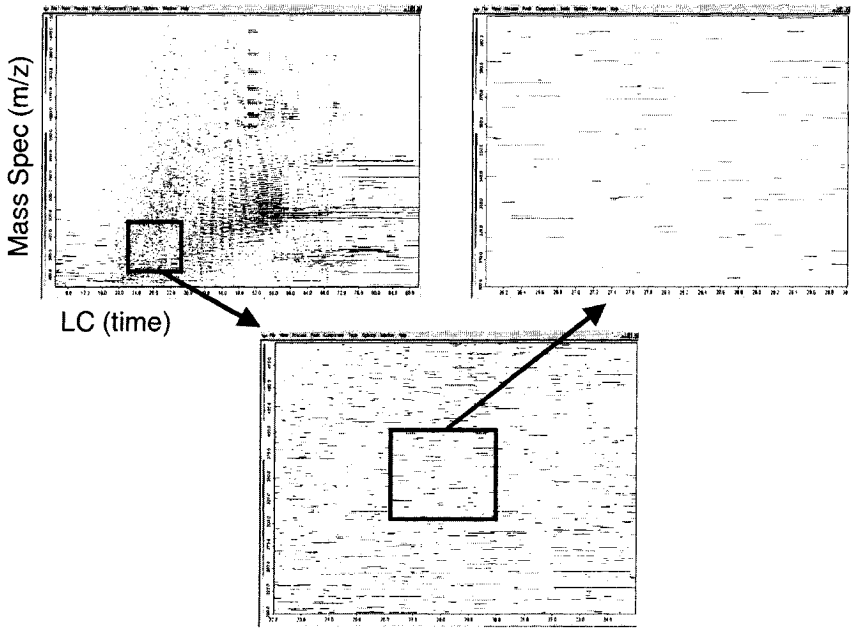


Figure 3. LC-MS analysis of 10 µL of human serum metabolome. Chromatography elution time is plotted vs. mass-to-charge (*m/z*) ratio. About 2000 components (monoisotopic peaks) were tracked with a two-hour chromatographic run.

Recently, several commercial sources introduced their version of data processing software. For example, Waters Corporation introduced Markerlynx™ software for processing metabolic profiling data from LC-MS analysis. This software currently can only process data generated by Waters' mass spectrometers. The ACD/MS Manager from Advanced Chemistry

Development, Inc. can be used to process profiling data from various instruments. Details such as how the profiling data are processed are not yet available to users. We can expect to see further commercial software developments for analysis of large data sets.

3.5 Statistics and data mining

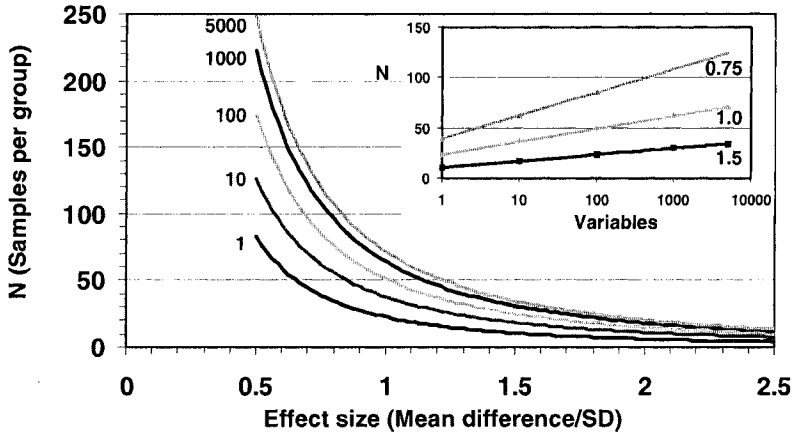


Figure 4. Samples needed to reach statistical significance. The number of samples needed as a function of effect size when an increasing number of variables (mass spectrometry components in this case) are compared. The effect size is the difference between the means of the measure for the two groups, relative to the weighted standard deviation in the groups (MD/SD). The number of variables compared is constant for each curve and is listed to the left of each curve (1 to 5000). Sample number as a function of the number of variables (on a logarithmic scale) is plotted in the inset. The calculation is based on unpaired two group comparisons, a power of 90% and an overall study-wise p-value of 0.05 based on the Bonferroni correction. It assumes that the variables are independent.

Since clinical samples have non-trivial costs associated with collection and processing, it is important to know how many samples are required to derive statistically meaningful results in an initial biomarker discovery study. A calculation for evaluating the required numbers of samples (Kantor, 2002) demonstrates that the number needed can be quite manageable. Figure 4 shows the number of samples needed as the number of independent variables increases from 1 to 5000. Variables can be components in mass spectrometry data. The calculation here assumes an unpaired comparison of two groups (e.g. disease *vs* healthy group) and a power of 90%. To maintain the *overall* positive rate at 0.05, the p-values from the univariate test statistics have been adjusted using the Bonferroni correction (Blair, 1996;

Holm, 1979). The utility of any given variable is determined by the effect size, which is the difference between the means of the measure for the two groups, relative to the weighted standard deviation of the measure in the groups (MD/SD). For initial studies in which there is no *a priori* estimate of the effect size for most variables, it is reasonable to power a first study to detect effect sizes of around one. In this case, for profiling 5000 variables, 70 samples are required per group (Figure 4). Of course, larger effect sizes can be detected with fewer numbers of samples. For example, for 1000 variables the number of samples needed per group drops from 62 to 30 and then 18 when the effect size increases from 1 to 1.5 to 2.0, respectively. These levels might be appropriate for pilot studies. It is also fair to say that the Bonferroni correction is extremely conservative, and does not take account of meaningful biological information acquired during the experiments. It is often useful to evaluate profiling results at multiple levels of univariate p-values.

Metabolic and proteomic profiling data are broad, with many more variables than samples. Both univariate and multivariate approaches have been applied to analyze such data sets. Univariate methods such as the *t* test and its nonparametric equivalent, the Wilcoxon test, can be used to determine the significance of observed changes. Other univariate methods have also been developed specifically for analyzing expression profile data (Zhu *et al.*, 2003). Univariate analysis is often straighforward and thus useful for initial evaluation of the data.

The limitation of univariate analysis is that it does not account for the interdependence of the variables (White *et al.*, 2004). In contrast, multivariate approaches can address the correlations among variables.

Examples of multivariate analysis methods include principal component analysis (PCA) (Hilsenbeck *et al.*, 1999), clustering (Eisen *et al.*, 1998), linear discriminant analysis (LDA) (Dumas *et al.*, 2002), artificial neural networks (ANN) (Ott *et al.*, 2003), and self-organizing maps (SOM) (Dow *et al.*, 2004). Regardless of the statistical methods employed, it is generally useful to apply more than one approach to gain a greater confidence and insight into the results.

After statistically significant differences are determined, the next step is chemical structure identification of these differing components. Identification is necessary to rule out spurious findings due to experimental bias (Diamandis, 2004). The structural complexity and diversity of metabolites often make identification more difficult than for genes and proteins. To date, there are no publicly available databases containing a repertoire of all metabolites in a metabolome such as GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) and SwissProt (http://us.expasy.org/sprot/) for genes and proteins. Metabolomic databases

such as KEGG (http://www.genome.jp/kegg/) and Dictionary of Natural Products (http://www.chemnetbase.com/scripts/dnpweb.exe?welcome-main) are far from comprehensive.

Identification of metabolites by GC-MS usually involves matching of the observed electron-impact fragmentation pattern to a library containing fragmentation patterns from a large number of organic molecules such as in the National Institute of Standards (NIST) database (http://www.nist.gov/srd/nist1.htm).

Since there is no comparable database currently available for ESI-MS data, the identification of metabolites from ESI-MS generally includes: (1) obtaining elemental composition based on accurate masses; (2) querying known chemical databases for candidate chemical structures; (3) acquiring tandem mass spectrometry data (MS/MS) or even MS^n data to confirm the proposed structure or eliminate candidate structures; and (4) if possible, testing the pure compound for retention time and MS/MS fragmentation pattern. If the above steps still cannot determine the structure of a metabolite (e.g. isomers), NMR can be used for further structure elucidation. The challenge for NMR analysis is to purify enough material from complex biological fluids, especially when the metabolite of interest is present in low abundance. Positive identification of the metabolites will then lead the way for further data mining, e.g. querying reference databases containing biological pathways, literature curation for known disease mechanisms and correlating with any available data from proteomic profiling and gene expression experiments.

3.6 Validation

The discovery of meaningful changes in a complex chemical system (e.g. body fluids) requires validation of both the analytical techniques and the biological results (candidate biomarkers). For profiling techniques, data variability from sources such as sample collection, processing and instruments should be evaluated (Glassbrook and Ryals, 2001).

Recently, error models as a function of local peak intensity were established for LC-MS profiling data (Anderle, 2004). In that work, pooled human serum samples were used to estimate the variation contributed from (a) instruments (LC-MS) or (b) sample processing before LC-MS. Such exercises help establish baseline performance for the evaluation of biological variation in the metabolite concentration among individuals. Validation of analytical techniques can be performed by checking the reproducibility of statistically significant components. For example, in a recent biomarker discovery project, we noted that human serum samples from rheumatoid arthritis (RA) and control subjects could be re-processed and analyzed 10

months after the first study and that the statistically meaningful changes agreed between the two studies (unpublished data).

The accuracy of quantification can be tested with spiked standards to build standard curves. For example, we demonstrated the quantification capability of LC-MS technology without using spiked standards (Wang *et al.*, 2003). Different amounts of vitamin B_{12} were spiked into pooled human serum samples, which were then analyzed by LC-MS. Note that intensity normalization was performed to correct overall intensity drifts. Figure 5 shows that normalized intensities of vitamin B_{12} were linearly proportional to its spiked concentration from both an ion-trap and a time-of-flight mass spectrometer.

Complete validation of biomarkers generally requires follow-up studies on a large population. However, several initial validation steps can be taken. For example, observed candidate markers should include known biomarkers reported in the literature, provided that these biomarkers are above the detection limit of current methods. Confirming the relavance of candidate biomarkers to disease mechanisms also increases the confidence of the results.

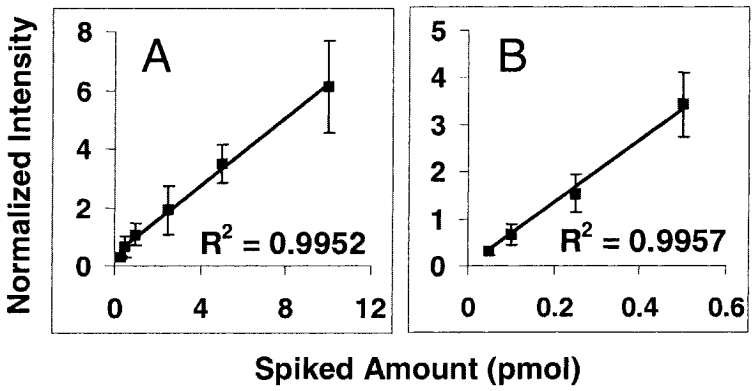


Figure 5. Standard curves for signals from normalized peak intensities of spiked vitamin B_{12} versus the spiked amount in 100 μL serum. These data were collected from on-line reverse-phase liquid chromatography directly coupled to either (A) an ion-trap mass spectrometer, or (B) an ESI-TOF mass spectrometer. Each data point represents the mean of normalized intensity of doubly charged vitamin B_{12} for 25 LC-MS runs. The error bars refer to the standard deviation (n=25). (Reprinted with permission from (Wang *et al.*, 2003). Copyright 2003 American Chemical Society).

4. CLINICAL APPLICATIONS

Metabolomics is becoming increasingly important and holds great potential in biomarker discovery research. Examples based on LC-MS and GC-MS will be illustrated below.

4.1 Disease biomarkers

Rheumatoid arthritis (RA) involves inflammation in the lining of the joints and/or internal organs and is a leading cause of long-term disability. Identification of biomarkers that predict subtypes of disease, clinical outcome or response to therapy can be valuable to the clinician. We are undertaking a longitudinal, noninterventional study in RA and have presented initial cytometry and proteomic profiling results (Kantor *et al.*, 2004). Here we present some initial metabolic profiling results. Samples from 18 RA subjects with active disease and 18 controls were compared. 50 μL of serum from each subject were processed to remove molecular weight components greater than 5000 Da and 10 μL was analyzed by reverse-phase LC-MS for metabolites and free floating peptides. MassView™ software was used to process raw data files and construct a list of components.

As mentioned earlier, components are common monoisotopic peaks found in multiple samples. Each component is a variable in the comparison. An unpaired *t* test or nonparametric test, as appropriate for each variable, was used for data analysis. Among 2200 variables (components) observed, 23 variables showed p-value smaller than 0.001, 136 variables p-value smaller than 0.01. There were total of 355 variables with p-value less than 0.05. Assuming all variables were independent, 110 variables would be expected to have p-value less than 0.05 by chance. This indicates that there are real metabolic differences between the active RA and control groups. The biggest fold change observed was 3.6 fold. Components (variables) with p-value less than 0.05 were plotted in Figure 6. Separations between RA subjects and controls were clearly observed. Identification and correlation of these changing metabolic components with biological pathways are currently ongoing.

Many researchers have demonstrated the utility of mass spectrometry in targeted metabolic profiling in a clinical setting. For example, tracking changes in metabolite concentrations using GC-MS was demonstrated (Shoemaker and Elliott, 1991). More than 90 urine samples were screened and 103 metabolites were quantified to confirm genetic metabolic defects. Another example is steroid profiling by LC-tandem mass spectrometry to improve the positive predictive value of newborn screening for congenital adrenal hyperplasia (Minutti *et al.*, 2004).

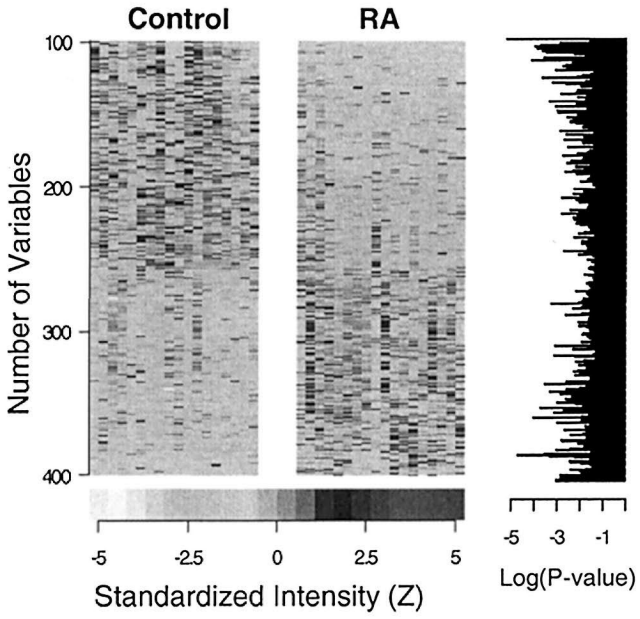


Figure 6. Heat map of 355 metabolic components (variables) that show statistically significant changes (p-value < 0.05) between control (n=18) and subjects with active rheumatoid arthritis (n=18). The x-axis represents the number of subjects. The y-axis represents the number of variables sorted by fold change. The intensity of all variables is standardized (Z = (individual measure – mean of all measures)/ standard deviation).

4.2 Drug discovery and development

Metabolomics can play an important role in drug discovery and development. Biomarkers that indicate early-stage efficacy or toxicity of a compound can help reduce the time and costs of developing a drug.

Screening toxicity markers by LC-MS and multivariate analysis was recently demonstrated by a proof-of-concept study on drug-induced phospholipidosis (Idborg *et al.*, 2004; Idborg-Bjorkman *et al.*, 2003). In that study, urine samples were collected from 12 male Wistar rats, half were dosed with citalopram, an antidepressant drug, and the other half given regular drinking water. Solid-phase extraction was used to prepare samples for LC-MS. The raw data were first reduced to peaks by automatic curve resolution (Manne and Grande, 2000; Shen *et al.*, 2001) and the peaks were aligned among samples. Multivariate statistical methods such as PCA were

performed for pattern recognition. PCA loadings were studied as a means of discovering potential biomarkers. Figure 7 shows a score plot based on PCA of all samples. There is an obvious separation between treated (rats #7-12) and untreated rats (#1-6). It is worth pointing out that several drug metabolites were falsely assigned as potential biomarkers (Helena Idborg, personal communication). Therefore, all possible drug metabolites should be filtered out from the data before PCA analysis. In this case, a publicly accessible reference database containing drug metabolites would be helpful. Nevertheless, results showed that differences in the metabolic pattern and time dependence could be captured using this methodology.

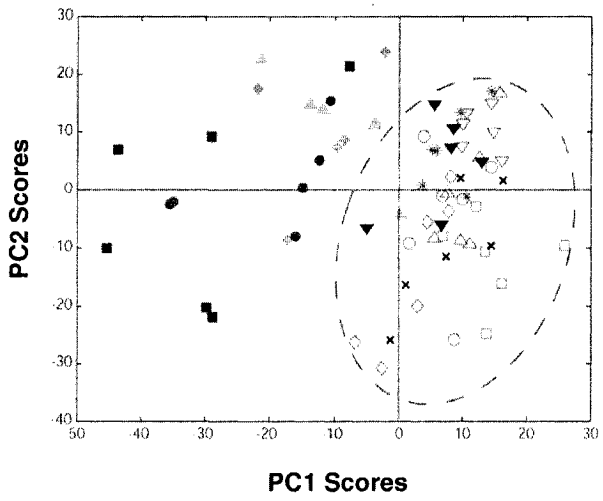


Figure 7. Scores of PC1 and PC2 obtained from PCA of all samples are plotted. Control samples (rats #1-6) are separated from the dosed animal samples (#7-12) collected after day 3. The control samples are shown as open markers on the right. Open markers of different kinds were used to separate each control rat, e.g., data for rat #4 on different collection days are shown as open diamonds (◊). Predosed samples (rats #7-12) are labeled **x**. Post-dose samples (filled markers) collected after 1 (▼), 3 (▲), 7 (●), 10 (♦) and 14 (■) days are distributed to the left of the plot. (Reprinted with kind permission from (Idborg-Bjorkman *et al.*, 2003). Copyright 2003 American Chemical Society).

5. CONCLUSION

Metabolomics is emerging as another important "omics" science. Considerable progress has already been made in terms of analytical technologies for tracking and quantifying thousands of variables. There are a number of challenges to overcome. Publicly available databases are needed

that link retention time and mass spectra to metabolites for LC-MS data, allowing facile identification of metabolites. Ultimately, only when mass spectral profiles (peak patterns) are translated to names of metabolites on a large scale can biologists readily determine the relevance of these metabolites to the disease or drug effect. In addition, an informatics infrastructure is desirable for easy integration of metabolic information with genomic and proteomic data. Finally, technologies for metabolic profiling, including sample preparation methods and instrumentation, need further improvement for more comprehensive coverage of the metabolome. With rapid advances in all aspects of metabolomics, we anticipate a greater contribution of this field to early disease diagnosis and improved treatments.

ACKNOWLEDGEMENTS

We thank the many colleagues at SurroMed for their contributions: Michael Natan for Figure 1, Hua Lin for Figure 2, Markus Anderle for Figure 6, Gary Frenzel, Jennifer Thompson, Thomas Shaler, Sushmita Roy, Weixun Wang, Praveen Kumar, and Jeff Satkofsky for general technology development. We thank Andrea Perrone and Grace Davis for clinical support. We thank David J. Fisher and his colleagues at the Palo Alto Medical Foundation, Arthur M. Bobrove, Melvin C. Britton, Barbara Anderson, Dianna Hill, Paula Whited and Kathleen Boice for subject recruitment and clinical samples. Special thanks go to Ed Rubenstein for thoughtful reading of the manuscript.

REFERENCES

Adam B-L, Vlahou A, Semmes OJ and Wright GL Jr. Proteomic approaches to biomarker discovery in prostate and bladder cancers. *Proteomics*, 1: 1264-1270 (2001).

Aharoni A, de Vos CHR, Verhoeven HA, Maliepaard CA, Kruppa G, Bino R and Goodenowe DB. Nontargeted metabolome analysis by use of Fourier transform ion cyclotron mass spectrometry. *Omics*, 6: 217-234 (2002).

Allen J, Davey HM, Broadhurst D, Heald JK, Rowland JJ, Oliver SG and Kell DB. High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nat. Biotechnol.*, 21: 692-696 (2003).

Anderle M, Roy S, Lin H, Becker C. Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum. *Bioinformatics*, 20: 3575 - 3582 (2004).

Beckonert O, Bollard ME, Ebbels TMD, Keun HC, Antti H, Holmes E, Lindon JC and Nicholson JK. NMR-based metabonomic toxicity classification: hierarchical cluster analysis and k-nearest-neighbour approaches. *Anal. Chim. Acta*, 490: 3-15 (2003).

Beecher CWW. The human metabolome, In *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*, Harrigan GG, Goodacre R. (Eds), Kluwer Publishers, New York, 311-319 (2003).

Bischoff R and Luider TM. Methodological advances in the discovery of protein and peptide disease markers. *J. Chromat. B*, 803: 27-40 (2004).

Blair RC, Troendle JF, Beck RW. Control of familywise errors in multiple endpoint assessments via stepwise permutation tests. *Stat. Med.*, 15: 1107-1121 (1996).

Bonato PS. Recent advances in the determination of enantiomeric drugs and their metabolites in biological fluids by capillary electrophoresis-mediated microanalysis. *Electrophoresis*, 24: 4078-4094 (2003).

Brindle JT, Antti H, Holmes E, Tranter G, Nicholson JK, Bethell HWL, Clarke S, Schofield PM, McKilligin E, Mosedale DE and Grainger DJ. Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using ^{1}H-NMR-based metabonomics. *Nat. Medicine*, 8: 1439-1445 (2002).

Diamandis EP. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool. *Mol. Cell. Proteomics*, 3: 367-378 (2004).

Dow LK, Kalelkar S and Dow ER. Self-organizing maps for the analysis of NMR spectra. *Drug Discov. Today: BIOSILICO*, 2: 157-163 (2004).

Dumas M-E, Canlet C, Andre F, Vercauteren J and Paris A. Metabonomic assessment of physiological disruptions using ^{1}H-^{13}C HMBC-NMR spectroscopy combined with pattern recognition procedures performed on filtered variables. *Anal. Chem.*, 74: 2261-2273 (2002).

Eisen MB, Spellman PT, Brown PO and Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95: 14863-8 (1998).

Fiehn O. Metabolomics - the link between genotypes and phenotypes. *Plant Mol Biol.*, 48: 155-71 (2002).

Garrod S, Humpher E, Connor SC, Connelly JC, Spraul M, Nicholson JK and Holmes E. High-resolution ^{1}H NMR and magic angle spinning NMR spectroscopic investigation of the biochemical effects of 2-bromoethanamine in intact renal and hepatic tissue. *Mag. Reson. Medicine*, 45: 781-790 (2001).

Glassbrook N and Ryals J. A systematic approach to biochemical profiling. *Curr. Opin. Plant Biol.*, 4: 186-90 (2001).

Griffin JL, Walker LA, Shore RF and Nicholson JK. Metabolic profiling of chronic Cadmium exposure in the rat. *Chem. Res. Toxicol.*, 14: 1428-1434 (2001).

Group BDW. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin. Pharmacol. Ther.*, 69: 89-95 (2001).

Harrigan GG, Goodacre R. (Eds) *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*, pp. 335 pp. Kluwer Publishers, New York (2003).

Hastings CA, Norton SM and Roy S. New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data, *Rapid Commun. Mass Spectrom.*, 16: 462-7 (2002).

He Q-Y and Chiu J-F. Proteomics in biomarker discovery and drug development. *J. Cell. Biochem.*, 89: 868-886 (2003).

Hilsenbeck SG, Friedrichs WE, Schiff R, O'Connell P, Hansen RK, Osborne CK and Fuqua SA. Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *J. Natl. Cancer Inst.*, 91: 453-9 (1999).

Holm S. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6: 65-70 (1979).

Horning EC and Horning MG. Human metabolic profiles obtained by gas chromatography and gas chromatography-mass spectrometry. *Advan. Chromatogr. Proc. Int. Symp., 6th*:226-43 (1970).

Huck CW, Stecher G, Bakry R and Bonn GK. Recent progress in high-performance capillary bioseparations. *Electrophoresis*, 24: 3977-3997 (2003).

Idborg H, Edlund PO and Jacobsson SP. Multivariate approaches for efficient detection of potential metabolites from liquid chromatography/mass spectrometry data. *Rapid Commun. Mass Spectrom.*, 18: 944-54 (2004).

Idborg-Bjorkman H, Edlund PO, Kvalheim OM, Schuppe-Koistinen I and Jacobsson SP. Screening of biomarkers in rat urine using LC/electrospray ionization-MS and two-way data analysis. *Anal Chem.*, 75: 4784-92 (2003).

Kaddurah-Daouk R, Beecher C, Kristal BS, Matson WR, Bogdanov M and Asa DJ. Bioanalytical advances for metabolomics and metabolic profiling. *PharmaGenomics* 4: 46-52 (2004).

Kantor AB. Comprehensive phenotyping and biological marker discovery. *Dis. Markers* 18: 91-7 (2002).

Kantor AB, Wang W, Lin H, Govindarajan H, Anderle M, Perrone A and Becker C. Biomarker discovery by comprehensive phenotyping for autoimmune diseases. *Clin. Immunol.*, 111: 186-195 (2004).

Krieg RC, Paweletz CP, Liotta LA and Petricoin EF. III. Clinical proteomics for cancer biomarker discovery and therapeutic targeting. *Technol. Cancer Res. Treat.*, 1: 263-272 (2002).

Lenz EM, Wilson ID, Timbrell JA and Nicholson JK. A ^{1}H NMR spectroscopic study of the biochemical effects of ifosfamide in the rat: evaluation of potential biomarkers. *Biomarkers*, 5: 424-435 (2000).

Lindon JC, Holmes E, Bollard ME, Stanley EG and Nicholson JK. Metabonomics technologies and their applications in physiological monitoring, drug safety assessment and disease diagnosis. *Biomarkers*, 9: 1-31 (2004a).

Lindon JC, Holmes E and Nicholson JK. Metabonomics and its role in drug development and disease diagnosis. *Expert Rev. Mol. Diagn.*, 4: 189-199 (2004b).

Manne R and Grande BV. Resolution of two-way data from hyphenated chromatography by means of elementary matrix transformations. *Chemomet. Intell. Lab. Syst.*, 50: 35-46 (2000).

McDonald WH and Yates JR 3rd. Shotgun proteomics and biomarker discovery. *Dis. Markers*, 18: 99-105 (2002).

Minutti CZ, Lacey JM, Magera MJ, Hahn SH, McCann M, Schulze A, Cheillan D, Dorche C, Chace DH, Lymp JF, Zimmerman D, Rinaldo P and Matern D. Steroid profiling by tandem mass spectrometry improves the positive predictive value of newborn screening for congenital adrenal hyperplasia. *J. Clin. Endocrin. Metabol.*, 89: 3687-3693 (2004).

Nicholson J, Lindon J, Scarfe G, Wilson I, Abou-Shakra F, Sage A, Harland G and Castro-Perez J. Quantification and identification of 2-bromo-4-trifluoromethylaniline metabolites in rat urine using HPLC-ICP-MS/TOF/MS. *Adv. Mass Spectrom.*, 15: 659-661 (2001).

Nicholson JK and Wilson ID. Opinion: Understanding 'global' systems biology: Metabonomics and the continuum of metabolism. *Nat. Rev. Drug Discov.*, 2: 668-676 (2003).

Ott K-H, Aranibar N, Singh B and Stockton GW. Metabonomics classifies pathways affected by bioactive compounds. Artificial neural network classification of NMR spectra of plant extracts. *Phytochemistry*, 62: 971-985 (2003).

Pang JX, Ginanni N, Dongre AR, Hefta SA and Opiteck GJ. Biomarker discovery in urine by proteomics. *J. Proteome Res.*, 1: 161-169 (2002).

Scarfe GB, Clayton E, Wilson ID and Nicholson JK. Identification and quantification of metabolites of 2,3,5,6-tetrafluoro-4-trifluoromethylaniline in rat urine using ^{19}F nuclear magnetic resonance spectroscopy, high-performance liquid chromatography-nuclear magnetic resonance spectroscopy and high-performance liquid chromatography-mass spectrometry. *J. Chromatog, B*, 748: 311-319 (2000).

Schlichtherle-Cerny H, Affolter M and Cerny C. Hydrophilic interaction liquid chromatography coupled to electrospray mass spectrometry of small polar compounds in food analysis. *Anal. Chem.*, 75: 2349-54 (2003).

Shen H, Grung B, Kvalheim OM and Eide I. Automated curve resolution applied to data from multi-detection instruments. *Anal. Chimic. Acta*, 446: 313-328 (2001).

Shoemaker JD and Elliott WH. Automated screening of urine samples for carbohydrates, organic and amino acids after treatment with urease. *J. Chromatography*, 562: 125-38 (1991).

Soga T, Ohashi Y, Ueno Y, Naraoka H, Tomita M and Nishioka T. Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. *J. Proteome Res.*, 2: 488-94 (2003).

Spraul M, Freund AS, Nast RE, Withers RS, Maas WE and Corcoran O. Advancing NMR sensitivity for LC-NMR-MS using a cryoflow probe: Application to the analysis of acetaminophen metabolites in urine. *Anal. Chem.*, 75: 1536-1541 (2003).

Stobiecki M and Makkar HPS. Recent advances in analytical methods for identification and quantification of phenolic compounds. *EAAP Publication* 110 (*Recent Advances of Research in Antinutritional Factors in Legume Seeds and Oilseeds*): 11-28 (2004).

Tolstikov VV and Fiehn O. Analysis of highly polar compounds of plant origin: Combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Anal. Biochem.*, 301: 298-307 (2002).

Tugwood JD, Hollins LE and Cockerill MJ. Genomics and the search for novel biomarkers in toxicology. *Biomarkers*, 8: 79-92 (2003).

van der Greef J, Stroobant P and van der Heijden R. The role of analytical sciences in medical systems biology. *Curr. Opin. Chem. Biol.*, 8: 559-565 (2004).

Vernon SD, Unger ER, Dimulescu IM, Rajeevan M and Reeves WC. Utility of the blood for gene expression profiling and biomarker discovery in chronic fatigue syndrome. *Dis. Markers*, 18: 193-199 (2002).

Viant MR, Rosenblum ES and Tieerdema RS. NMR-based metabolomics: A powerful approach for characterizing the effects of environmental stressors on organism health. *Environ. Sci. Technol.*, 37: 4982-9 (2003).

Wang W, Zhou H, Lin H, Roy S, Shaler TA, Hill LR, Norton S, Kumar P, Anderle M and Becker CH. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.* 75: 4818-4826 (2003).

Warne MA, Lenz EM, Osborn D, Weeks JM and Nicholson JK. An NMR-based metabonomic investigation of the toxic effects of 3-trifluoromethyl-aniline on the earthworm *Eisenia veneta.Biomarkers*, 5: 56-72 (2000).

Waters NJ, Holmes E, Waterfield CJ, Farrant RD and Nicholson JK. NMR and pattern recognition studies on liver extracts and intact livers from rats treated with a-naphthylisothiocyanate.*Biochem.Pharmacol.*,4: 67-77 (2002).

White CN, Chan DW and Zhang Z. Bioinformatics strategies for proteomic profiling.*Clin. Biochem.*, 37: 636-41 (2004).

Whitehouse CM, Dreyer RN, Yamashita M and Fenn JB. Electrospray interface for liquid chromatographs and mass spectrometers. *Anal. Chem.*, 57: 675-9 (1985).

Wu CC and MacCoss MJ. Shotgun proteomics: tools for the analysis of complex biological systems. *Curr. Opin. Mol. Ther.*, 4:242-50 (2002).

Yamashita M and Fenn JB. Electrospray ion source. Another variation on the free-jet theme. *J. Phys. Chem.*, 88: 4451-9 (1984).

Zhu W, Wang X, Ma Y, Rao M, Glimm J and Kovach JS. Detection of cancer-specific markers amid massive mass spectral data. *Proc. Natl. Acad. Sci. USA*, 100: 14666-71 (2003).

Zlatkis A and Liebich HM. Profile of volatile metabolites in human urine. *Clin. Chem.*, 17: 592-4 (1971).

Chapter 10

NMR-BASED METABONOMICS IN TOXICOLOGY RESEARCH

Laura K. Schnackenberg, Richard D. Beger, and Yvonne P. Dragan
Division of Systems Toxicology, 2, National Center for Toxicological Research, Food and Drug Administration, Jefferson, AR 72079-9502

1. INTRODUCTION

The search for biomarkers of toxicity and disease is one of the major research initiatives at the National Center for Toxicological Research (NCTR) of the FDA. A systems biology approach that integrates biomarkers from functional genomics, proteomics and metabonomics research into a comprehensive understanding of toxicological events and disease states has been initiated at the NCTR. Since genomics, proteomics and metabonomics have become accepted research platforms within the pharmaceutical industry, understanding 'omics' research has become a priority for the FDA. Thus, 'omics' research falls within the NCTR mission statement on the NCTR public website that is defined as "to conduct peer-reviewed scientific research that supports and anticipates the FDA's current and future regulatory needs. This involves fundamental and applied research specifically designed to define biological mechanisms of action underlying the toxicity of products regulated by the FDA." The goals of our metabonomics research are to investigate and validate metabonomics drug toxicity results (Robertson *et al.*, 2000) and to use systems biology approaches to understand toxic mechanisms and disease states (Aardema and MacGregor, 2002; Nicholson and Wilson, 2003; Coen *et al.*, 2004). Transcriptomics describes the analysis of gene expression while proteomics is the comprehensive analysis of the proteome. Metabonomics evaluates the temporal changes in cell metabolism, which is downstream from genomic

and proteomic events. Monitoring the changes in the cellular concentrations of metabolites represents an opportunity to identify phenotypical responses to drug toxicity and disease state (Fiehn, 2002).

Metabonomics is the study of metabolite levels in biofluids and tissues. Biofluids contain hundreds of compounds in dynamic equilibrium with various tissues and reflect ongoing cellular processes. While these processes have been studied extensively over the last century, there are a limited number of endogenous metabolites that have been linked to cellular metabolism. The direct link between metabonomics and cellular metabolism represents an opportunity to determine metabolic changes associated with changes in cell function. Since serum and urine represent biofluids that are readily accessible through minimally invasive techniques, they are a useful source of potential biomarkers of drug toxicity, drug efficacy and disease state. Systematic studies of tissues allow the metabonomic, protein and gene expression changes measured to be traced back to mechanisms responsible for drug toxicity and disease state.

The nature of nuclear magnetic resonance (NMR) spectroscopy permits a quantitative investigation of changes in biofluid metabolite composition. NMR spectroscopy is uniquely suited to the analysis of the complex composition of biofluids following a toxic insult because it is quantitative, reproducible and has a linear response. A toxic insult will induce a characteristic variation in the metabolite concentrations in the biofluid that will be reflected in the NMR spectrum. Changes in the NMR spectra of biofluids can thus be used to understand toxicity in terms of the perturbation of a biochemical pathway. Metabonomics provides an approach to the study of metabolic changes in biofluids, cells or tissues following "pathophysiological insult or genetic modification" (Nicholson *et al.*, 1999). Metabonomics uses a training set of samples exhibiting a specific toxicity or pathology and compares them with controls in order to generate a classification scheme for a particular response. Once a metabolic model is generated and spectral biomarkers determined, standard analytical techniques and spectral libraries can be used to identify the specific compounds responsible for the spectral biomarkers. In this manner, the spectral biomarkers and their associated metabolites that define differences between groups exhibiting a toxic response or disease, compared to a control, can be identified. This approach permits the metabolic response to dietary changes or drug administration to be queried with high resolution proton (^{1}H) NMR spectra of biofluids or tissues. In combination with multivariate statistics, NMR-based metabonomics can provide an integrated analysis of cellular function. The overall pattern of the spectrum and the hundreds of metabolites detected provide new biomarker development for a particular toxic response or disease state. This approach has been pioneered

by Jeremy Nicholson of Imperial College. A COnsortium for MEtabonomic Toxicology, called COMET, was formed between Imperial College and six pharmaceutical companies to develop a database of 150 kidney and liver toxins as assessed by urinary ^{1}H NMR analysis (Lindon *et al.*, 2003). NMR-based metabonomics analysis is not limited to urine and serum studies. Several investigations have demonstrated the utility of an NMR-based approach for other biofluids including plasma, bile and seminal fluid among others (Nicholson and Wilson, 1989; Nicholson and Foxall, 1995; Lynch *et al.*, 1994). In addition, tissue samples including liver have been examined using magic angle spinning (MAS) NMR technology (Waters *et al.*, 2000; Waters *et al.*, 2001). Furthermore, analysis of cell culture media and tissue extracts has been performed using NMR-based methods in order to give a complete metabolic picture of the toxicity caused by tamoxifen (Griffin *et al.*, 2003) and acetaminophen (Coen *et al.*, 2003).

At NCTR, we have presently decided to focus on solution state NMR for metabonomics research because it is the most quantitatively accurate spectroscopic technique (Keun *et al.*, 2002). However, mass spectrometric (MS)-based metabonomics techniques offer higher sensitivity and molecular selectivity over NMR (see Chapter 19 by Lenz *et al.*). For these reasons, MS-based methods will be combined with NMR-based methods in future metabonomics studies at NCTR.

NMR-based metabonomics has also been shown to be useful for clinical investigations. Urine from children has been analyzed to detect diseases that result from inborn errors in metabolism including maple syrup disease, mevalonic aciduria and orotic aciduria (Holmes *et al.*, 1997; Bamforth *et al.*, 1999). Currently, one company in North Carolina is analyzing thousands of serum samples a day for lipoprotein and cholesterol levels by NMR. This data is then being used to generate models of cardiovascular disease (Brindle *et al.*, 2002). Finally, metabonomics methods in the clinical research environment have been demonstrated to be useful to evaluate the efficacy of immunosuppressants relative to their cytotoxicity in renal transplant patients through analysis of urine (Foxall *et al.*, 1993).

2. ANALYSIS OF NMR DATA

Unsupervised and supervised models can be built for each toxin or disease using the NMR spectra from urine, serum or tissue extracts. Most statistical analyses of metabonomics spectra start with principal component analysis (PCA), a data reduction technique that allows visualization of the multidimensional spectral data in two or three dimensions. Prior to integration, normalization and PCA, the spectral regions for water, urea,

drug and drug metabolites are removed. PCA is then applied for the preliminary visual analysis of the metabonomics data and to identify outliers that should be removed before further analysis.

High resolution 1D proton NMR spectral data are normally reduced into 256 spectral bins that are 0.04 ppm wide (Holmes *et al.*, 1994). An alternative to the standard binning technique in which each bin has an equal width is known as "intelligent" bucketing. This bucketing technique uses algorithms to determine the optimal bin size for the spectra as a group, typically somewhere between 0.02 to 0.06 ppm. The use of the intelligent bucketing method helps reduce errors that may arise due to small changes in pH that can shift a peak. Further, the algorithms used are designed to look for local minima in the spectra as a group, which can prevent the same peak from being split between two bins. Once the spectra have been bucketed, the areas within each bucket are integrated and these values are used for PCA. This unsupervised, multivariate PCA statistical method allows the reduction of the multidimensional data represented by NMR spectra to a couple of dimensions that can be more readily visually interpreted. Principal components (PCs) will have contributions or "loadings" from all the variables, or in this case, the normalized intensities from each bin. The variation that is explained by the first PC is removed from the original data and the axis of maximum variation from the remaining data is used to form the second PC, which is orthogonal to the first PC. This process continues until all the variation in the data has been exhausted. The loadings plot can be used to distinguish the variables that are major contributors to a PC and that lead to the clustering of groups within the dataset. Within each bin is information about the concentrations of the metabolites within a 0.02 to 0.06 ppm wide region of the NMR spectrum. This allows the identity of affected metabolites to be narrowed down and potentially determined by the application of 2D NMR and other analytical techniques.

When the variation between groups is greater than the variation within groups, the PCA plot will show isolated clustering of the different groups. If the PCA sample clusters correlate directly with pathology, then supervised statistical analysis becomes unnecessary. If the clusters overlap or do not directly correlate to pathology, then supervised statistical methods like soft independent modeling of class analogy (SIMCA), partial least squares discriminant analysis function (PLS-DA), and artificial neural networks (ANN) models can be derived (Holmes *et al.*, 2000).

Since obtaining urine or serum is relatively non-invasive, it is possible to follow individual animals over a period of time. In this case, PCA of the serum or urine spectra can be used to depict the metabolic trajectory of an animal throughout the length of a study. Under toxic doses, an animal will initially move away from the control cluster in time followed by a return to

the control cluster if the toxic insult is reversible. The magnitude of the toxic and recovery responses is proportional to the distance from the control region. The PCA loadings responsible for the initial metabolic changes from control can be interpreted as early biomarkers of that specific toxicity. Our preliminary data indicate that metabonomics profiles of drug-induced toxicity depend on the dose and the amount of time following drug administration. The cross correlation of the time- and dose-dependent profiles will provide useful information about the toxin under investigation.

A metabonomics method that uses the temporal trajectory path of metabolic changes with PCA statistics called scaled-to-maximum, aligned, and reduced trajectories (SMART) can be used to assess a particular drug response (Keun *et al.*, 2004). SMART homothetic trajectory analysis is a process that aligns each animal's principal component trajectories by pre-dose subtraction and then scales the trajectories by the maximum differences between trajectories. The spectra are then averaged as a final reduction. SMART analysis is possible because of the linear response of NMR and is used in an attempt to map metabolic trajectories onto one another. If a metabolic principal component trajectory overlaps another trajectory using SMART analysis, this implies that the same temporal biological events are responsible for the measured responses. SMART analysis can be used to remove inter-laboratory variation and normal physiological and phenotypical variation, to correlate different dose-response trajectories and to find correlations between inter-species responses.

Currently, the identification of many of the metabolites in NMR and MS spectra is not possible without additional analytical experiments. Several companies are building NMR and MS spectral databases and associated software that deconvolute NMR or MS data into the relative concentrations of their endogenous metabolic components. One of these companies is currently developing a metabonomics database of 200 endogenous metabolites to aid in evaluating 1D high resolution proton NMR spectra of clinical urine and serum. The current state of metabonomics research relies on statistical analysis of the NMR and MS spectra and the hope that the analysis will identify spectral features that can be associated with specific metabolites that are biomarkers of drug toxicity or disease state. The future analysis of metabonomics data will involve identifying not only which set of metabolites is responsible in the detection of drug toxicity and disease state, but to link this set of metabolites to altered metabolic pathways and the associated enzymes and genes.

3. ADVANTAGES OF METABONOMICS IN TOXICOLOGICAL RESEARCH

The major advantages of NMR-based metabonomics research are the minimally invasive nature of sample collection and the ease of sample preparation. There is little sample preparation necessary, which makes implementation of the metabonomics method both fast and economical. Further, since collection of urine and serum for metabonomics analyses is non-destructive to the animal, temporal analyses using the same animal can be performed. This provides the opportunity to use fewer animals per study, which also makes metabolic analyses more economical. The minimally invasive nature of sample collection means that metabonomics can be easily applied to both nonclinical and clinical research. This is a major advantage over methods that require tissue for analysis when moving from nonclinical to clinical studies. Further, endogenous metabolites such as sugars, amino acids, lipids, steroids and triglycerides are species independent, unlike gene transcripts and proteins that may show large interspecies variation. Because metabonomics research is very fast and has lower costs per sample analyzed, metabonomics can be used to direct and validate proteomic, genomic and clinical studies.

Another benefit of using a temporal study with the same animal is that it is not necessary to have a complete understanding of the pharmacodynamic and pharmacokinetic properties of a drug prior to the investigation. This removes the need for preparatory pharmacokinetic research prior to initial *in vivo* toxicity studies saving time and money. This is especially important in the early ADME (absorption, distribution, metabolism and excretion) stages of drug discovery. Temporal metabonomics trajectory studies allow a quick determination as to whether the endogenous metabolite concentrations return to normal after a period of time following the toxic insult or remain in a perturbed toxic state. The metabolic trajectory can be used to evaluate whether and when the animal recovers from the toxic stress (Nicholson *et al.*, 2002). If the animal recovers from the toxic stress, the cause of the toxic response may be understood by relating the temporal metabolic response to other metabonomics trajectories where the mechanism is known. The magnitude of the toxic response and the speed of animal recovery can be used by sponsors and regulatory agencies to determine if the threats of toxicity outweigh the health benefits of the drug.

4. EXAMPLES OF METABONOMICS RESEARCH ON URINE

4.1 Ethanol toxicity in rats

The changes in metabolite concentrations found in biofluids following a toxic stress or in response to a disease are dependent on specific mechanisms. To examine this, urine from adult male Sprague-Dawley (SD) rats that were fed a diet containing 13 g/kg/day ethanol by continuous intragastric infusion was analyzed. This paradigm has previously been reported to result in urinary ethanol concentrations (UECs) that cycle between 0 and 500 mg/dL over a 6-7 day period (Badger *et al.*, 2000).

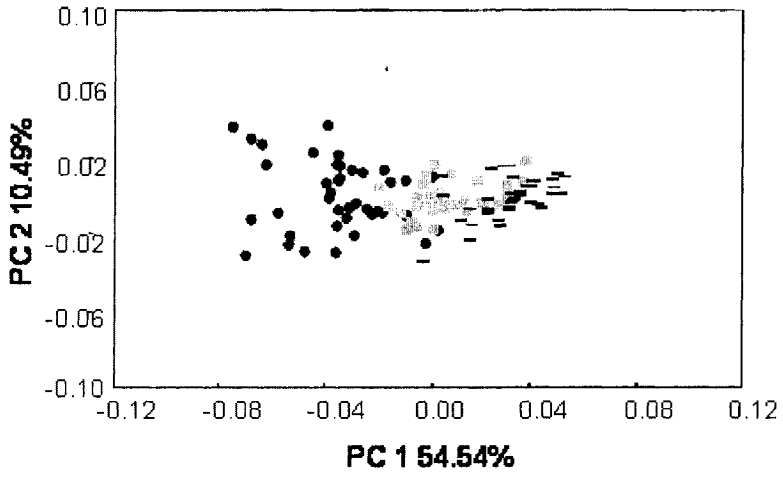


Figure 1. Principal component analysis of NMR spectra of urine (refer text for details). Samples with the highest UEC are shown in dark dashes, samples with intermediate levels of UEC in gray squares, and the samples with lowest levels of UEC are shown in dark diamonds

Further, class I alcohol dehydrogenase (ADH) expression was shown to be elevated at high UECs, which results in greater ethanol metabolism that drives the UEC down. After 3 days, the mean UEC levels are reduced to near zero and the ADH expression decreased. In the current study, 1D proton NMR was used to monitor the UECs. Bins corresponding to the ethanol, acetate and acetaldehyde peaks were removed prior to statistical analysis and

principal components were calculated from the binned 1D ^{1}H NMR spectra. Figure 1 is 2D principal component (PC1 versus PC2) analysis plot of the NMR spectra of urine collected at the 24 hour mark from four SD rats over a 28 day period. PC1 is responsible for 54.5% of the spectral variation and PC2 is responsible for 10.5% of the spectral variation seen in the urine from ethanol diet study. Figure 1 shows that high UEC samples (UEC > 250 mg/dL), medium UEC samples (250 mg/dL < UEC < 100 mg/dL) and low UEC samples (UEC < 100 mg/dL) are separated along the first PC axis. Further inspection of the NMR spectra of urine showed an increase in creatine that occurred during the first week of the ethanol study in animals with low UEC output. Taurine and creatine have been previously reported as markers of hepatic dysfunction (Timbrell *et al.*, 1995) but both metabolites are not very specific to liver. Based on information from the loadings plots, 2D NMR and MS techniques were used to identify ethyl glucuronide as the primary xenobiotic metabolite responsible for the variation between high, medium and low UEC noted in the first principal component of the PCA plot. The metabonomic analysis of urine from an ethanol feed study in rats identified liver toxicity biomarkers and an exogenous metabolite of ethanol.

4.2 Drug toxicity in mice

Cisplatin is a commonly used chemotherapeutic drug for several types of cancer (Rosenberg, 1978). However, the doses must be limited since repeated dosing with cisplatin can result in acute renal failure. Four 129/SV mice were dosed intraperitoneally with the renal toxin cisplatin at a dose of 20 mg/kg body weight. Urine was collected over a 24 hour period prior to dosing and for three days after dosing. Regions containing the water and urea resonances were removed prior to bucket integration and PCA. No signals from cisplatin or its metabolites were seen in any of the spectra. A 3D PCA plot of the cisplatin metabonomics study shows that the treatment group one day after cisplatin administration clusters slightly away from the control, while points for treatment groups on days 2 and 3 move further from the control, indicating that the mice has not begun to recover from the initial toxic insult. Hippurate, glucose, fucose, fumarate, succinate, dimethylamine, trimethylamine N-oxide and numerous amino acids showed dramatic concentration changes in urine with respect to control levels during the three day metabonomic investigation of cisplatin toxicity. Metabonomics analysis of urine from cisplatin-treated mice identified several potential biomarkers of toxicity before histopathological changes could be detected and suggested a rather prolonged toxic response.

5. FUTURE RESEARCH OBJECTIVES AT NCTR

5.1 Integration of NMR and MS techniques

The use of NMR-based metabonomics methods has been well documented in the literature. Mass spectrometry (MS) has been used in metabonomics because of the high sensitivity (Plumb *et al.*, 2002). Both platforms clearly have advantages and disadvantages as metabonomics platforms. With MS, each molecule is detected differently due to ionization potential differences between molecules, whereas each molecule is detected the same in an NMR spectrum. Further, the detection of specific metabolites by NMR may be confounded by overlap with the chemical shifts of other endogenous metabolites or by removal of solvent or drug metabolite resonances. On the other hand, NMR is much faster since there is no need to separate the metabolites by GC or LC prior to analysis. NMR methods are also highly quantitative with 1-2% accuracy (Keun *et al.*, 2002). While NMR methods are inherently more accurate, NMR is not as sensitive as MS techniques.

In order for MS to be applicable, the metabolites must be in an ionized form. This means that some of the metabolites detected by NMR may not be detected using MS. Conversely, MS results may pinpoint metabolites that were not significant in NMR analysis due to spectral overlap or were not detected by NMR because of low concentration. A combination of NMR-based and MS-based methods should result in a more complete analysis of the system. The use of MS in conjunction with NMR also affords the opportunity to identify and quantify specific metabolites that are seen in tissues and biofluids.

5.2 NMR-Based metabonomics of serum

Serum samples can be evaluated in several ways. The straight serum sample can be analyzed or various extractions can be used to isolate particular groups of metabolites. Acetonitrile can be used to precipitate the proteins and to extract the aqueous metabolites. Subsequently, the pellet from the acetonitrile precipitation can be extracted with a 2:1 volume ratio of chloroform and methanol to obtain the lipophilic metabolites. Results obtained from studies of serum can be directly correlated with clinical chemistry measurements of serum enzyme levels including ALT and AST values. Metabonomics investigations of serum have been applied to assess the severity of coronary heart disease, as well as to correlate the serum metabolic profiles with hypertension (Brindle *et al.*, 2002, 2003). In both cases, metabonomics studies of the serum presented a relatively non-

invasive means of screening that could be readily applied in a clinical environment.

5.3 Metabonomics investigations of whole tissue and tissue extracts

Metabonomics studies are not limited to biofluids or serum extracts. The ability to investigate whole tissue and tissue extracts also exists. Advances in high resolution magic angle spinning (MAS) NMR spectroscopy make it possible to investigate whole tissue or biopsy samples. However, tissue extracts can be prepared and analyzed using solution state NMR spectroscopic methods. A range of metabolites can be investigated depending on the type of extract that is obtained. Particularly useful is an acetonitrile extraction followed by chloroform/methanol extraction that allows analysis of the aqueous and lipophilic metabolites respectively. The incorporation of tissue studies provides important information about changes in endogenous metabolites that occur due to a tissue-specific toxicity. Analysis of specific target tissues gives specific information about a toxic response within a particular organ whereas; in urine changes are related to the entire animal system. Further, correlation of metabolic changes in the tissue with the urine could provide more direct information about the pathways involved in the toxic response or disease state condition.

MAS solid state NMR has been used to investigate the effects of 2-bromoethanamine (BEA) on renal and hepatic tissue (Garrod *et al.*, 2000). Previous studies had looked at the effects of BEA on urinary metabolites. This study allowed the metabolites found in the urine study to be related to specific tissues. Further, the mode of action of BEA could be elucidated by comparison of the effects on different tissues.

Waters and coworkers (2001) also used an integrated approach to study the effect of α-naphthylisothiocyanate (ANIT) on male Han-Wistar rats. MAS-NMR was used to analyze the intact livers. Following MAS-NMR measurements, high resolution NMR of tissue extracts, plasma and urine were obtained. The time points of metabolic perturbation could be assessed in each biological matrix showing how the metabolite patterns were associated with the response to the toxin over the time course of the experiment.

6. CONCLUSION

NMR-based metabonomics research can be used to produce preclinical models that aim to identify hepatic and renal toxicities. Metabonomics research on urine and serum is non-invasive, can be automated, has a low cost per sample but has a large initial investment in NMR and MS equipment. The metabonomics data can be evaluated quickly to produce predictive toxicological models. Although the temporal biomarkers found in spectra of biofluids have been shown to be excellent for discriminating drug toxicity from control cases, the markers are not tissue specific. For metabonomics of biofluids to become an acceptable technique, it will be necessary to relate the metabolic changes seen in biofluids to the associated metabolic, proteomic and genomic changes observed in target tissues. The focus of some of the FDA/NCTR metabonomics research will be to understand the biochemical mechanisms ongoing in the tissue that result in altered metabolite changes in the serum and urine. Our research will apply metabolic, proteomic and genomic techniques to tissue studies to determine the mechanisms that are directly responsible for the metabolic changes seen in biofluids. Increases or decreases in metabolite concentration due to a toxic response or disease process are not necessarily the same in the tissue and biofluid. This may require the measurement of some metabolites at levels below NMR detection limits, requiring the use of other metabonomics platforms such as MS. In the future, we anticipate that the integration of NMR with MS technologies will be required to address the multiple challenges facing metabolic profiling. NMR and MS spectral databases of metabolites will not only aid the interpretation of metabolic profiles, but will identify NMR- or MS-detected metabolites that are not in current databases. Reducing the NMR and MS spectra to relative endogenous metabolic levels may be very important when dealing with clinical trials where the dietary intake is not controlled. Finally, metabonomics is a new technology that holds a lot of promise for toxicological research, evaluation of drug efficacy, clinical diagnostics and regulatory drug evaluation.

ACKNOWLEDGMENTS

We would like to acknowledge Dr. Thomas Badger and Dr. Martin Ronis of the Arkansas Children's Hospital for providing urine samples used in the ethanol study, and Dr. Robert Safirstein of the Central Arkansas Veterans Administration Healthcare System for providing the urine samples in the cisplatin study

REFERENCES

Aardema MJ, MacGregor JT. Toxicology and genetic toxicology in the new era of "toxicogenomics": impact of "omics" technologies. *Mutat. Res.*, 499: 13-25 (2002).

Badger TM *et al.* Cyclic expression of class I alcohol dehydrogenase in male rats treated with ethanol. *Biochem. Biophys. Res. Commun.*, 274: 684-688 (2000).

Bamforth FJ *et al.* Diagnosis of inborn errors of metabolism using ^{1}H NMR spectroscopic analysis of urine. *J. Inherit. Metab. Dis.*, 22: 297-301 (1999).

Brindle JT *et al.* Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using ^{1}H-NMR-based metabonomics. *Nat. Med.*, 8: 1439-1445 (2002).

Brindle JT *et al.* Application of chemometrics to ^{1}H NMR spectroscopic data to investigate a relationship between human serum metabolic profiles and hypertension. *Analyst*, 128: 32-36 (2003).

Coen M *et al.* Integrated metabonomic investigation of acetaminophen toxicity in the mouse using NMR spectroscopy. *Chem. Res. Tox.*, 16: 295-303 (2003).

Coen M *et al.* Integrated application of transcriptomics and metabonomics yields new insight into the due to paracetamol in the mouse. *J. Pharm. Biomed. Anal.*, 35: 93-105 (2004).

Fiehn O. Metabolomics—the link between genotypes and phenotypes. *Plant Mol. Biol.*, 48: 155-171 (2002).

Foxall PJD *et al.* NMR spectroscopy as a novel approach to the monitoring of renal transplant function. *Kidney Int.*, 43: 234-245 (1993).

Garrod S *et al.* High resolution ^{1}H NMR and magic angle spinning NMR spectroscopic investigation of the biochemical effects of 2-bromoethanamine in intact renal and hepatic tissue. *Magn. Reson. Med.*, 45: 781-790 (2000).

Griffin JL *et al.* Cellular environment of metabolites and a metabonomic study of tamoxifen in endometrial cells using gradient high resolution magic angle spinning ^{1}H NMR spectroscopy. *Biochim. Biophys. Acta*, 1619: 151-158 (2003).

Holmes E *et al.* Automatic data reduction and pattern recognition methods for analysis of ^{1}H nuclear magnetic resonance spectra of human urine from normal and pathological states. *Anal. Biochem.*, 220: 284-296 (1994).

Holmes E *et al.* 750 MHz ^{1}H NMR spectroscopy characterization of the complex metabolic pattern of urine from patients with inborn errors of metabolism: 2-hydroxyglutatric aciduria and maple syrup urine disease. *J. Pharm. Biomed. Anal.*, 15: 1647-1657 (1997).

Holmes E *et al.* Development of a model for classification of toxin-induced lesions using ^{1}H NMR spectroscopy of urine combined with pattern recognition. *NMR Biomed.*, 11: 235-44 (1998).

Holmes E *et al.* Chemometric models for toxicity classification based on NMR spectra of biofluids. *Chem. Res. Toxicol.*, 13: 471-478 (2000).

Keun HC *et al.* Analytical reproducibility in ^{1}H NMR-based metabonomic urinalysis. *Chem. Res. Toxicol.*, 15: 1380-1386 (2002).

Keun HC *et al.* Geometric trajectory analysis of metabolic responses to toxicity can define treatment specific profiles. *Chem. Res. Toxicol.*, 17: 579-587 (2004).

Lindon JC *et al.* Contemporary issues in toxicology: The role of metabonomics in toxicology and its evaluation by the COMET project. *Toxicol. Appl. Pharm.*, 187: 137-146 (2003)

Lynch MJ *et al.* Ultra high field NMR spectroscopic studies on human seminal fluid, seminal vesicle and prostatic secretions. *J. Pharm. Biomed. Anal.*, 12: 5-19 (1994).

Nicholson JK, Wilson ID. High-resolution proton NMR spectroscopy of biological fluids. *Prog Nucl. Mag. Res. Sp.*, 21: 449-501 (1989).

Nicholson JK, Foxall PJD. 750 MHz ^{1}H and ^{1}H-^{13}C NMR spectroscopy of human blood plasma. *Anal. Chem.*, 67: 793-811 (1995).

Nicholson JK, Lindon JC, Holmes H. "Metabonomic": understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, 29: 1181-1189 (1999).

Nicholson JK *et al.* Metabonomics: a platform for studying drug toxicity and gene function. *Nat. Rev. Drug Discov.*, 1: 153-161 (2002).

Nicholson JK, Wilson ID. Understanding "Global" Systems Biology: Metabonomics and the continuum of metabolism. *Nat. Rev. Drug Discov.*, 2: 668-676 (2003).

Plumb RS *et al.* Metabonomic analysis of mouse urine by liquid-chromatography-time of flight mass spectrometry (LC-TOFMS): detection of strain, diurnal and gender differences. *Analyst*, 128: 819-823 (2002).

Robertson DG *et al.* Metabonomics: evaluation of nuclear magnetic resonance (NMR) and pattern recognition technology for rapid *in vivo* screening of liver and kidney toxicants. *Toxicol. Sci.*, 57: 326-337 (2000).

Rosenberg B. Platinum complexes for the treatment of cancer. *Interdiscipl. Sci. Rev.*, 3: 134–147 (1978).

Timbrell JA, Waterfield CJ, Draper RP. Use of urinary taurine and creatine as biomarkers of organ dysfunction and metabolic perturbations. *Comp. Haematol. Int.*, 5: 112-119 (1995).

Waters NJ *et al.* High-resolution magic angle spinning ^{1}H NMR spectroscopy of intact liver and kidney: optimization of sample preparation procedures and biochemical stability of tissue during spectral acquisition. *Anal. Biochem.*, 282: 16-23 (2000).

Waters NJ *et al.* NMR and pattern recognition studies on the time-related metabolic effects of α-Naphthylisothiocyanate on liver, urine, and plasma in the rat: an integrative metabonomic approach. *Chem. Res. Toxicol.*, 14: 1401-1412 (2001).

Chapter 11

METHODOLOGICAL ISSUES AND EXPERIMENTAL DESIGN CONSIDERATIONS IN METABOLIC PROFILE-BASED CLASSIFICATIONS

Bruce S. Kristal,[1,2] Yevgeniya Shurubor,[1] Ugo Paolucci,[1] Wayne R. Matson[3]
[1]Dementia Research Service, Burke Medical Research Institute, 785 Mamaroneck Ave., White Plains, NY 10605; [2]Departments of Biochemistry and Neuroscience, Cornell University Medical College, 1300 York Ave, NY, NY 10021, [3]ESA, Inc., 22 Alpha Road, Chelmsford, MA 01824;

1. INTRODUCTION

The onset of the era of –omics and systems biology arguably brings with it fundamental shifts in every aspect of biological research. At the conceptual level, there is a shift in focus from mechanistic details to more generalistic hypotheses that is followed by data-driven (discovery-based) set of analyses. A similar shift from univariate statistical analyses to multivariate approaches to data analysis such as those which have been more commonly used in physical sciences and engineering is also apparent. In reality, however, successful implementation of these new approaches relies as much on experimental design and mechanistic understanding, as ever. Similarly, mathematical analysis eventually returns to an understanding of whether or not observed phenomena are primarily and inherently a result of statistical sampling or biological reality.

Not surprisingly, careful considerations of analytical, biological, and mathematical issues remain as important as ever, albeit with the complication that post-genomic technologies such as transcriptomics, proteomics and metabolomics bring (i.e., that they generate data that surpass

ready human interpretation). The days of the lone investigator glancing at raw data and recognizing the answer – or the problem with the experiment - are increasingly rare. Coupled with a growing understanding of –omics-specific, technology-specific, experimental and data analysis requirements and limitations, empirical observation suggests that each technology and experimental area will require the generation, and possibly standardization, of a series of experimental and mathematical approaches geared toward the specific applications of interest.

Whether there is an optimal approach will depend on the level at which optimality is sought. Clearly, there will be no optimal approach to all metabolomics experiments, but equally clearly, well-established experimental design standards will need to be present, as discussed below.

With respect to analytical issues, the comparative importance of precision, stability, sensitivity, dynamic range, resolution, and throughput will be highly dependent on the specific application. The requirement for precision is the most critical when differences between classes (or observations/test individuals/populations) approach the limitations of instrumental precision. For example, studies such as our work on dietary restriction (DR) (Shi *et al.,* 2002a; Shi *et al.,* 2002b; Shi *et al.,* 2002c; Shi *et al.,* 2004; Paolucci *et al.,* 2004a; Paolucci *et al.,* 2004b) often rely on measurements of metabolites that differ by 18% (median) to 23% (mean) between groups, requiring a strict attention to high precision measurements. In contrast, studies with state markers that differ 10-fold between classes, such as those analyzed in some toxicology studies, could have coefficients of variation at >50% without compromising the study. Studies conducted within short time frames, such as single drug-dose treatments, will be far more tolerant of reduced stability in an analytical platform than studies performed longitudinally over years, such as ours.

Studies where the key metabolites of interest are present at high levels (e.g., toxin derivatives, glucose, amino acids, etc.) may not require great sensitivity (Lindon *et al.,* 2003), whereas studies of less abundant metabolites, such as neurotransmitters and oxidative damage products, will require optimizing sensitivity and signal-to-noise ratios. Similarly, studies that focus on the production of critical, high concentration single metabolites will have less stringent requirements for resolution and dynamic range than those, such as ours, that build up profiles from multiple peaks, and thus require the ability to detect, and accurately quantify, as many of these peaks as possible.

Finally, throughput appears to be a tradeoff that must be made relative to the other variables. Certain techniques used to analyze the identities and concentrations of metabolites/small molecules, such as NMR (Lindon *et al.,* 2003), Raman spectroscopy (Jarvis and Goodacre, 2004; Lopez-Diez and

Goodacre, 2004), Fourier transform infrared spectroscopy (Winder *et al.*, 2004) and MALDI and electrospray MS (Vaidyanathan and Goodacre, 2004) offer tremendous throughput, but have disadvantages in other respects. Chromatography based systems, such as HPLC coupled with coulometric detection (Matson *et al.*, 1984; Matson *et al.*, 1987; Matson *et al.*, 1990; LeWitt *et al.*, 1992; Ogawa *et al.*, 1992; Beal *et al.*, 1992) that we use, can offer greater resolution and sensitivity, but at the price of low throughput and the related higher cost/sample. The redox selectivity of the instrument can be both an advantage (through simplifying and focusing analysis, particularly in studies relating to free radical damage or metabolism), or a disadvantage (because of the potential limitation in what the system can score).

In this chapter, we will describe the specific empirical lessons that we have learned from our studies; we propose that these lessons potentially provide a series of general approaches. But first, we offer a word about the path not taken.

2. THE "ROAD NOT TAKEN" – METABOLOMICS WITH HIGH ABUNDANCE STATE MARKERS

Sometimes metabolomics can be simplified, at least initially, to the analysis of a relatively small number of key metabolites (Ellis *et al.*, 2002; Gavaghan *et al.*, 2000; Griffin *et al.*, 2000). For example, this approach can be effective in studies of drug metabolism and/or toxicology, where differences between multiple cohorts in a study is usually relatively small, and the individuality of responses is most commonly limited to either kinetics, to ratios between a few metabolites, or to the non-responder/non-metabolizer (Keun *et al.*, 2004). Likewise, the effect of gender in studies involving powerful single markers is often – although not always – either irrelevant or one of temporal relevance of quantitative degree, not qualitative difference (e.g., glucose is increased in both diabetic men and diabetic women). Because these issues are comparatively minor points in conditions where there are strong markers, the specifics of experimental design are often less critical than in the studies described below (of course, in some cases, such as looking at early time points, design issues become critical, but these situations don't fall under the class of studies described in this section – see below). Likewise, variations due to analytical parameters and experimental sample preparation are also less likely to be influential, although this is clearly a matter of degree – careless design or preparation will still be met with disastrous consequences. As a result, experimenters working in these areas can afford to trade difficulties in analytical

measurements for progressively higher throughputs (at least until they begin to delve further into the less overt aspects of their metabolic profiles).

The critical point here is that, with reservations, it appears likely that experimental limitations in these systems will not be imposed by issues in metabolomics, but rather by outside issues, such as limitations of clinical samples, availability of drug analogues, etc.

3. THE ROAD LESS TRAVELED –THE DARK SIDE OF METABOLOMICS

The problems that we study in our investigation of caloric restriction differ from those described above for at least one of six major reasons: (i) there are no state markers; (ii) many or all of our markers are present at very low levels/concentrations; (iii) cohort effects are significant; (iv) gender effects are significant; (v) biochemical differences between individuals are significant; (vi) sampling issues and sample handling may create unwanted signal of comparable magnitude to the signal being followed. Before addressing the metabolomics issues involved, we will first discuss the biological problem we are studying.

3.1 Defining a serotype for long term, low calorie intake

In humans, the association between caloric balance and disease is most clearly seen in the association between increased body mass index (BMI) and increased risk of neoplasia, Type II diabetes, and cardiovascular and cerebral-vascular disease (Willett, Dietz, and Colditz, 1999; Willett, 2001). As one example, an increase of ~50% in cancer risk was observed in individuals with morbid obesity (BMI>40) in an American Cancer Society study of 900,000 adults (Calle *et al.,* 2003). In general, this study associated obesity with an increased risk of colorectal, pancreatic, liver, esophageal, kidney, gallbladder, prostate, breast, cervical, ovarian, and uterine cancers, as well as multiple myeloma and non-Hodgkin's lymphoma -- in other words the risk of nearly all cancers is increased by obesity (Calle *et al.,* 2003). The study concluded by estimating that "the current patterns of overweight and obesity in the United States could account for 14% of all deaths from cancer in men and 20% of those in women."

Evidence of increased risk is also seen in studies of specialized populations. For example, data from the Nurses' Health Study show that even a BMI of 26 – representing borderline obesity –results in an increased risk of coronary heart disease and hypertension and an eight-fold increase in

the frequency of Type II diabetes as compared with a BMI of 21 (Willett *et al.*, 1999). Similarly, a weight gain of 15 kg in adulthood was associated with similar increases in disease frequency (Willett *et al.*, 1999).

Studies in laboratory rodents clearly demonstrate the complementary observation that low-calorie diets are associated with increased longevity and delayed morbidity (McCay, 1935; McCay *et al.*,1935; Weindruch and Walford, 1988; Yu, 1994; 1996; Weindruch and Walford, 2000). Dietary restriction (DR), also called caloric or food restriction, is an experimental paradigm in which the dietary or caloric intake of a group of animals is reduced relative to the intake of *ad libitum* fed (AL) controls. The history of the exploration of this phenomenon is extensive (Kristal and Yu, 1994) and has been described in detail in other publications (see Weindruch and Walford, 1988; Yu, 1994; 1996), It is important to recognize that DR is the most potent, the most robust, and the most reproducible known means of reducing morbidity and mortality in mammals. A problem with the study of this phenomenon, however, is that nearly all systems, ranging from hormonal to anatomical to biochemical and metabolic, are affected to one extent or another. In short, this is a large and very complex system to study.

One approach to the study of these complex system level phenomena is to analyze them using high throughput and/or high data density techniques. In humans, for example, questions of this class might be initially addressed by using genomics techniques, such as the analysis of single nucleotide polymorphisms. Most studies of DR, however, including our own, are carried out in inbred animal strains in which there is no genetic diversity, so the phenomenon plays out entirely as a post-genetic effect. People have therefore looked for distinctions between DR and AL fed populations by examining global differences in mRNA and protein expression, using mRNA expression arrays (Lee *et al.*, 1999) and proteomics techniques (Heydari *et al.*, 1989; Butler *et al.*, 1989). By contrast, we have focused on metabolomics.

3.2 The basics: A metabolomic approach to study DR

Our studies have four goals: (i) to gain insights into the mechanisms by which DR exerts its effects; (ii) to recognize DR and AL feeding regimens in different species; (iii) to determine biochemically the effect of long-term caloric intake of an individual (an important issue in epidemiological studies); and (iv) to identify predictive markers of disease.

As we noted above, the basic components in a metabolomics study are a general hypothesis, an analytical platform, and an informatics tool. Our starting hypothesis was that long-term, low-calorie diets induce changes in metabolism that persist throughout the lifespan. This hypothesis makes two

predictions: (i) DR alters the sera metabolome and that, therefore; (ii) there exists a DR serotype – which would allow the equivalent of a blood test for DR – or more generally, for any level of caloric intake.

To test this hypothesis, we analyzed sera from Fischer x Brown Norway F1 rats maintained on AL feeding or experiencing a variety of different extents and durations of DR. The total experiment involved male and female rats of five different ages (6-30 months of age). Overall, the study included 36 diet groups divided into 82 cohorts. Notably, and we think generally, the scientific goals of the study and the constraints imposed by the subtle differences in the serotypes in question then became the dominant factor in the choice of both an analytical platform and an informatics platform. Specifically, in the analysis of metabolites present in the sera, we chose methods (HPLC-coulometry) that provided high sensitivity, resolution, and dynamic range advantages that were obtained at the expense of high throughput and information about the structure of the metabolite. Likewise, the type of data analysis we required drove our choice of an informatics platform. We chose to use relatively standard multivariate analysis approaches, including clustering, principal components (PCA), and their cognate pattern recognition tools, because they are well-suited to classification and to reducing large datasets into simpler, visual representations while still maintaining input from multiple metabolites.

4. EXPERIMENTAL DESIGN ISSUES: GENERAL/PRIMARY CONSIDERATIONS

In the interest of readers who are comparatively new to this field, it is worth briefly reviewing the basic issues involved in a metabolomics study of this sort. These first issues raised here, related to analysis, track back to elementary reasoning and rudimentary analytical chemistry. They are based on common sense, but are important and not to forget, as one moves into issues of experimental design. The first set of issues (concerned with accuracy, precision, coefficients of variation, etc., called "level one" issues, below) require an understanding of what multiple experimental runs look like and how quantitatively reproducible they are. "Level two" issues concern potential inter-sample, intra-individual differences. "Level three" issues address the individual differences between members of the same group. It is only after these issues have been considered that one can effectively begin to consider inter-group differences ("Level four").

Level one issues are essentially those of analytical chemistry: How should samples be acquired, handled, stored, and initially analyzed? Relevant particulars include, but are not limited to: (i) linearity of the assay

at the relevant ranges of concentrations; (ii) resolution of metabolites of interest from contaminants/other metabolites; (iii) reproducibility of results from samples that are split both before and after extractions to help determine the source of any errors); (iv) reproducibility of profiles based on split samples. Secondary issues include sensitivity of the assay to differences in sample acquisition and stability of the sample during storage. A third issue is stability over time, i.e., can one analyze samples over a period of weeks, months, or years under sufficiently stable conditions to collect useful data?

Level two issues concern potential inter-sample, intra-individual differences. For example, how different are sequential time resolved samples (i.e., sequential samples taken from an individual over a period of time) from a single individual? It is well known, for instance, that some measurements taken by dietary assessment instruments used in epidemiology display greater "within individual" variation (i.e., data taken from the same individual at different times) than "between individual" variation (i.e., data taken from two individuals). Even in this extreme case, however, the measurements are highly predictive at the population level. The goal of metabolomics, however, is to take such predictive and descriptive measurements to the level of the individual. It is therefore important, whenever possible, to determine the extent to which our measurements are robust for multiple measurements of the same individual.

Level three issues concern the level of variation of both single markers and overall profiles of individuals across all members of a class. Our own research directly suggests that the appropriate resolution of these issues is one of the most important elements in successful classification studies. One major issue that follows from our work is that non-spherical distributions in a class of interest seem incompatible with cluster algorithm-based separations, unless the interclass difference (distance) significantly exceeds the longest interclass differential. We have solved this problem through projection-based techniques such as PCA (Shi *et al.,* 2004; Paolucci *et al.,* 2004b), although other approaches, such as neural networks, genetic algorithms, and random forest methods might also be viable solutions. A second major issue is that those pursuing experiments in animals consider the use of multiple cohorts, as cohort-cohort differences can often dominate/obscure some group-dependent variations of interest. In our studies, these differences destroyed straight-forward descriptive based algorithms and necessitated moving to discriminant based algorithms (Paolucci *et al.,* 2004b), as described below. Notably, studies possessing strong state markers might be immune to level three issues.

Level four issues concern the differences between groups of interest. The logistics of these issues are addressed more directly below under informatics.

It is important to recognize that, in practice, one could simultaneously address several of these issues, or, arguably, skip levels in some cases. One might, for example, concurrently show that the instrument is analytically stable and that a specific metabolite is biologically stable within a model system of choice (e.g., Vigneau-Callahan *et al.,* 2001; Shi *et al.,* 2002c). At the "macroscopic" level, analysis of sera from AL fed female rats in one cohort by HPLC-coulometry gives results that look superficially identical to the results from AL fed female rats in a second cohort, but when one targets specific regions of the chromatograms, one can see near total conservation of a certain region but essentially complete differences in the other (Vigneau-Callahan *et al.,* 2001). There are several take-home lessons from this observation: (i) from the analytical side, we can make the observation that, in general, female rat serum looks like female rat sera. This is not surprising, but it is important in that it indicates that (ii) the platform is, broadly speaking, sufficiently analytically reproducible to consider a classification study; and (iii) there are metabolites that give highly conserved results and those that do not.

5. EXPERIMENTAL DESIGN ISSUES: SPECIFICS

Using the AL-DR study as a model, we will now discuss four issues that have arisen in our work, and our approaches to handling them: (i) "fuzzy" *vs.* tight analytical controls; (ii) analytical concerns; (iii) biological variability; (iv) gender and cohort effects. Analysis is dealt with in a subsequent section.

5.1 Fuzzy *vs* tight controls

One point to consider is whether one should tightly control the manner in which data is collected or whether one should deliberately, albeit selectively, "fuzzy" this process. For example, one could take blood samples at very defined times, using a much defined technique. The advantage of this approach is that one obtains data that is the most accurate about a very specific set of conditions. Alternatively, one may be a little bit fuzzy about this. In studies such as ours, for example, one may stretch out the time of sampling (i.e., "morning" *vs.* 9:00 am), the time samples sit at room temperature prior to moving to ice (~30 seconds to 2 minutes), the time samples sit at 4°C prior to centrifugation (~20 minutes to ~ 1 hour), and the

time the samples sit at 4°C before aliquoting and freezing (10 to 30 minutes). Samples from different groups must, of course, be handled identically. The advantage of this approach is that data about compounds that are either very labile or that display very sharp diurnal rhythms disappear into the noise. This effect helps to ensure that the compounds that do identify groups of interest, e.g., AL *vs* DR rats, are more robust.

5.2 Analytical concerns

The concentration of many metabolites in biological samples will approach signal/noise and detection limits in all analytical platforms, thus creating the problem of how to optimize the identification of potentially useful metabolites while simultaneously cutting out those which do not meet preset quality standards. In practice, we accomplished this by analyzing eight replicates of a single pool. Metabolites were incorporated into subsequent data analysis if they were found in 6/8 samples, with a mean ± 50% of the true mean concentration and a CV of <50%. These relatively low stringency conditions allowed us to confirm that we could measure ~ 300 redox active serum metabolites with sufficient reproducibility to make latter stages of the analysis possible. Two different sets of peak identification algorithms (Vigneau-Callahan *et al.,* 2001; Shi *et al.,* 2002b; 2002c; 2004) allowed us to validate a second set of analytes of approximately equal size, but with <50% overlap. By following each group separately, we increased coverage to close to 500 metabolites.

Given the focus here on experimental design, we will not further dwell on analytical issues such as maintaining high quality instrumentation.

5.3 Biological variability

In many practical ways, the problems that are associated with biological variation are closely tied to those related to analytical validation. In both cases, the critical concept for our approach was to eliminate those peaks that would not be useful at later stages for classification purposes. This serves to remove less informative metabolites, to simplify subsequent analysis and to reduce statistical noise. Thus, we can reduce the one component of biological variation, differences between concentrations of single metabolites, to the functional equivalent of the analytical test. These peaks were eliminated, in practice, by running an essentially equivalent analysis to that described above for the analytical control, but here using all of the individuals from a given class that were in a given cohort (Shi *et al.,* 2002c). Our primary interest, however, is in removing those peaks that are biologically uninformative – i.e., that have a high noise-to-signal ratio with

respect to class differences. We are willing, in contrast, to keep highly variable peaks that are informative. In practice, this can be readily done using multiple *t* tests at weak criteria, such as $p<0.20$ (See below and Shi *et al.*, 2002c).

5.4 Genders and cohorts

The next issue that must be addressed is whether or not there are gender and cohort effects on the markers of interest. As noted above, studies of severe toxicological injury and drug metabolism often proceed without major issues in these areas; for example, the presence of elevated blood glucose in male and female diabetics, and its reduction by insulin treatment. Even the shortest study reveals that this is not the case in our experiments. As one example, studies in two cohorts suggests that DR is associated with slight, but consistent, decreases in retinol in both male and female rats (Figure 1). This decrease, however, can only be observed when samples are matched by both gender and cohort. Examination of the data from female rats provides a perfect counter-example to the idea that it might be of general use across cohorts. Specifically, the decrease is such that the female AL rat in cohort 1 has identical levels of retinol as compared to the female DR rat in cohort 2. Analysis of retinol-palmitate reveals a different type of issue. Here, one sees that, in cohort 1, the male AL rats have nearly 5 times the level of this compound as compared to that observed in the DR rats, but the level in these DR rats was actually higher than that of the AL rats in the second cohort.

In practice, we have dealt with gender and cohort issues in two ways. The gender issue was addressed by carrying male and female models forward separately through several stages of analysis (clustering, principal components, and expert systems), before eventually determining that the two datasets could not be legitimately merged (Shi *et al.*, 2004; Paolucci *et al.*, 2004a, b). Cohort specificity was dealt with by testing each cohort separately. Within a given cohort, we found descriptive models valid, but our work eventually showed that we required discriminant models to address cohort differences (Shi *et al.*, 2004; Paolucci *et al.*, 2004a, b).

It is both possible – and important – to consider generalizing from these results. Changes in the metabolome that represent a biological signal in one system can often be biological noise in another. Two examples of this from our study, as noted above, are the effects of cohort (i.e., environment) and gender on the metabolome.

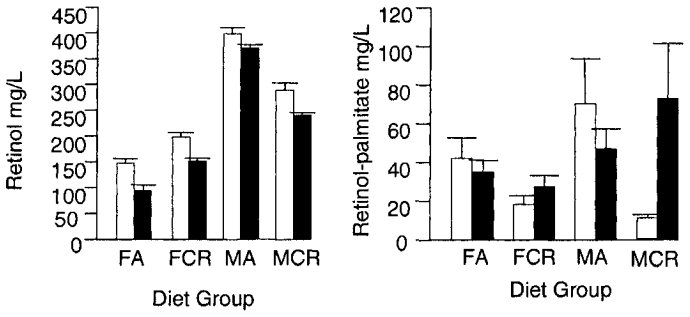


Figure 1. Cohort and Gender Specificity: Retinol (left panel) and retinol-palmitate (right panel) levels in the serum from 2 cohorts of male and female, AL and DR rats, each bar reflects N=8 rats. FA, female *ad libitum* fed; FCR, females maintained on caloric restriction (equivalent to DR); MA, males fed *ad libitum*; MCR, males maintained on caloric restriction. White bars, first cohort; black bars, second cohort. Assays were conducted using HPLC separations and coulometric array detection.

The impact of the environment on the metabolome takes several different forms with fundamentally distinct implications for those interested in metabolomics. One possibility occurs when the impact of some environments – or environmental factors – becomes lost in any given study because they vary on, essentially, an individual to individual basis. An example might be the effect of environment on a large epidemiological study, where the study participants are drawn broadly, and comparatively uniformly, from many different groups, locales, and populations. This type of confounding factor washes out many variables into the noise, but realistically, these would probably have been less than robust anyway. A second possibility occurs when the environment does influence the profile, but the impact should be equivalent for all individuals in the study. An example of this type of environmental impact is observed in studies of single cohorts. Such animals all eat exactly the same diet, are caged at the same time in the same room, and handled by the same people. Travel to and from the colony is often simultaneous, and thus occurs under identical circumstances. Under these conditions, the environmental impact is essentially invisible, but a concern arises as to the overall robustness of the profiles generated. This concern became immediately apparent to us when we examined our second cohort. Classification algorithms readily distinguished cohort 1 from 2, confounding attempts to distinguish AL and DR animals when the cohorts were pooled. In addition, we found that some of our best markers for distinguishing DR (i.e., $p<0.000001$) in cohort 1 were not statistically different between DR and AL fed animals in cohort 2.

Clearly, similar considerations hold for the impact of genetic variation on the metabolome – i.e., nature and nurture are both reflected in the metabolome.

These points lead to two critical observations. First, conducting studies in which the genome and environment are both very tightly controlled does yield the cleanest, tightest, strongest data, but the robustness of such data sets should be viewed with caution. Thus, secondly, metabolomic studies on experimental animals benefit from including multiple, separately handled cohorts, and possibly subsequent studies in alternative models. In our case, subtle changes in diet (i.e., batch number) and housing (unplanned physical plant work in the colony facility) most likely underlie the changes observed.

Gender issues must also be dealt with by design – not by chance. As noted earlier, metabolomic profiles generated in toxicological studies often appear to be substantially similar in both genders, as the major metabolites in question are often drug metabolites, or those few endogenous metabolites that are directly affected by the drugs in question (An example of such a study would be one that looked at the effects of allopurinol on purine catabolites). DR provides a very different, and potentially informative, lesson on the impact of gender on metabolomics. In nearly all strains of animals tested to date, DR extends longevity and delays morbidity equally well in both male and female animals. It is noted, however, that DR has significant effects on the sex hormone systems of both male and female animals. After several rounds of analysis, we determined that only about 25% of metabolites, identified as useful (i.e., for classification) in one sex, were also useful in the other. Perhaps more importantly, the profiles generated differ considerably. Whilst alternatives exist, our data suggests that building male and female profiles separately will greatly improve accuracy in many conditions.

6. INFORMATICS APPROACHES

6.1 Initial Cuts – where do we start?

If one wants to determine whether two groups can be distinguished based on metabolomic criteria, it is reasonable to begin by simply comparing a few members from each group. From experience, at least five possibilities exist: (i) clear, large and consistent differences exist between all the members in the two groups (consider fasting glucose in diabetics *vs* non-diabetics); (ii) clear and large differences exist between the groups, but these differences are not shared by all individuals in a group (consider diabetics maintained on insulin vs non-medicated diabetics); (iii) clear and large differences exist between individuals, but the differences are not related to group identity (as

an example, we have seen this in an unpublished study where metabolomics analysis readily identified two classes of observations, but the classes so identified were unrelated to any known variable – and the classes we were testing for were indistinguishable using metabolomics); (iv) differences exist at the levels of populations, but not individuals (i.e., there are overlapping distributions – all of our work on dietary restriction discussed here fits in this class), and; (v) there are no apparent differences. Case (i) examples are best evaluated to determine if the large differences observed are, in fact, consistent. If so, the metabolites that display these differences serve as state markers, and studies can be focused on a very limited subset of metabolites. Case (ii) examples suggest that there may be subgroups within one or more of the groups, and this possibility should be followed up initially by re-examining groups at both the level of metabolomics and at the level of other available information. Further studies can proceed most expediently when and if secondary groups are defined. Case (iii) studies suggest that binning- or tree-based informatics approaches should be used. Case (iv) studies, such as ours, suggest the need for more complex profiling, by techniques including clustering, projection methods, neural nets, and Bayesian analyses. Case (v) studies may also yield to such approaches, resulting in the identification of metabolites that differ between the groups, but one should recognize the possibility that the groups are metabolically equivalent. Over-fitting issues are particularly relevant with class (iv) and class (v) studies, which might result in the identification of apparent differences between groups even though no meaningful differences actually exist. One has to, for example, continually keep in mind that it is not surprising that some differences between groups can be found given that one has searched in excess of, perhaps, 1000 metabolites with small sample numbers. Case (i), (ii), (iii), and (v) studies will not be further addressed here; rather we will focus on class (iv) studies, i.e., those where differences exist at the population level but not at the level of individual

6.2 Generation and Validation of Robust Metabolic Profiles

We used a multi-component analysis platform beginning with t-tests, progressed to multivariate analyses, specifically hierarchical clustering analysis (HCA) and PCA and finally to pattern recognition and discriminant-based analysis.

As a general scheme, our approach took advantage of the fact that, given appropriate algorithm usage and follow-up validation studies, multivariate analyses are relatively efficient at identifying signal in the presence of random noise. We had two initial goals: (i) to cut the size of the data set, thus increasing our ability to focus on metabolites that could be detected

with high analytical accuracy and that had potentially high discriminatory power, and (ii) to minimize the loss of informative metabolites early on. In practice, these goals involve a tradeoff – we reduce the occurrence of false negatives (type II statistical errors) at the expense of increasing false positives (type I statistical errors). Initial cuts were thus made by cutting the data with a t-test at $p<0.20$. This is an extremely weak criterion, but it means that the type II errors are now primarily analytical errors – not primarily statistical noise in the biological signal. When we do this, we observed that about one-third of the 300 analytically valid peaks we were examining crossed the threshold (i.e., had $p<0.20$). We further estimated that as many as 60% of these represent type I statistical errors. We then took these markers forward into multivariate-based approaches, and asked whether HCA and PCA can recognize the groups of interest in the same set of samples that was used to generate the markers, i.e., to complete proof-of-principle (Shi *et al.,* 2002c). HCA is useful because it identifies the natural groups in the data, and PCA is useful for its ability to find linear combinations of original variables that account for maximal variation, thus shrinking the dataset while simultaneously beginning to associate certain mathematical relationships with groups.

6.3 Exploratory Analyses, "Proof-of-Principle," and Primary Validation

Most metabolomics-based classification studies can be broken into at least four phases. Phase I, described above, includes the identification of a series of metabolites that are potentially useful. Phases II and III concern the two initial phases of showing that these groups of metabolites are capable of distinguishing groups of interest. The first of these phases (Phase II), proof-of-principle, seeks the answers to four questions: (i) can my data distinguish groups of interest objectively, or at least subjectively (consider the case of selective rotations in a principal components space); (ii) How mathematically robust are my analyses? (iii) What do my data look like? and; (iv) what mathematical approaches are best suited to the next steps of analysis? The second of these phases (phase III validation) determines whether the metabolites selected, robustly distinguish between the groups of interest and begins to address the question of over-fitting. Successful validation of a profile enables one to move onward to "Phase IV" – model optimization and utilization.

Because of its objectivity, clustering is a useful approach to conduct proof-of-principle studies. Our HCA studies found that the group of metabolites having $p<0.20$ was able to distinguish groups with 100% accuracy in female rats (Shi *et al.,* 2002c). By changing the mathematical

details of the analysis, it was possible to show that the split between AL and DR groups in the proof of principle study was mathematically robust within the clustering algorithms tested.

The ability to simplify datasets and give an overview of the "structure" of a dataset makes PCA a useful tool to conduct proof of principle studies. Our PCA studies found that the group of metabolites having $p<0.20$ was able to distinguish groups with 100% accuracy in female rats. By changing the mathematical details of the analysis it was readily possible to show that the split between AL and DR groups in the proof-of-principle study was also mathematically robust.

Together, these "proof-of-principle" studies answer the four questions posed above. The data did distinguish groups of interest objectively; the analyses were mathematically robust; the samples had no overlap within a PCA space; multiple mathematical approaches were worth moving forward with, as all had worked. These data thus set the stage for determining if serotypes encode sufficient information to identify diet group, i.e., to move to primary validation studies (Shi *et al.*, 2002b). An example of this "proof of principle" PCA analysis is shown in Figure 2)

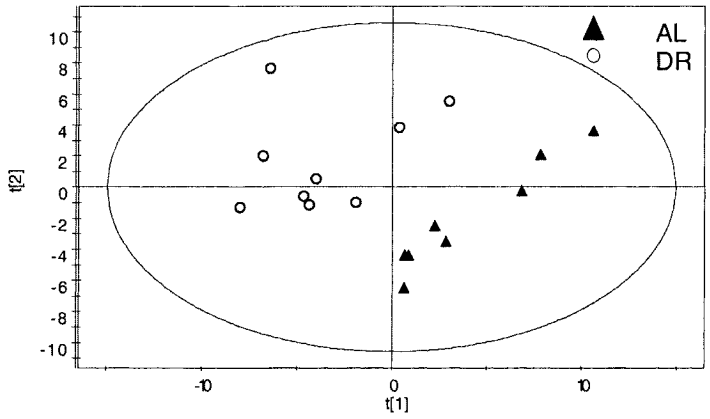


Figure 2. PCA scores plot based on serum metabolite analysis from 8 male rats fed *ad libitum* diets and 8 male rats maintained on DR regimens. Analysis was conducted using Umetrics SIMCA P-10.5. Data were unit variance scaled and the first two components were autofit. This plot is based on hand validated metabolite levels. See (Shi *et al.*, 2002b) for original analysis.

Validating nascent models represents the most critical stage in the use of metabolomic profiles for description, prediction or classification. In general,

metabolomics studies, like proteomics and transcriptomics studies, are done under conditions where the number of variables (metabolites) greatly exceeds the number of observations. Even the careful use of appropriate multivariate statistics, including permutation and leave-out validation testing, runs the risk of falling victim to over-fitting, both mathematically and biologically. The cleanest answer to these questions, and one that our work suggests is essential, is to validate with a completely independent dataset. In other words, to demonstrate that the model developed is predictive in a new cohort or group of interest.

In theory, at least, primary validation studies should be straight-forward, since we have already built reasonable models. Not surprisingly, validation studies often find over-fitting in the initial models and expose flaws in original reasoning. In practice, we suggest beginning validation studies by looking for statistical outliers. We have done this by looking at plots of Mahalanobis distance against the sums of sample residuals (i.e., the distance of the overall metabolomic profile of an observation from the overall metabolomic profiles of other observations in the study against the sum of the residuals of all variables in the profile), but other techniques are most likely equally efficacious. One can look for outliers using at least two different strategies. One strategy is to look at each class (e.g.., AL, DR) independently. This is a more exclusionary test for outliers, and is useful primarily if one wants to examine a core group. Alternatively, one may look for outliers in the context of all samples; this approach is more likely to include samples that serve as more stringent tests of the model. For example, a sample having intermediate characteristics is likely to be excluded by the first criteria, but not by the second. Regardless of the details of the tests chosen, testing for outliers helps ensure that valid models are not discarded early because of a small number of atypical samples. In practice, we found no outliers in our dataset, and were able to validate our models with high accuracy using either PCA or using cluster analysis (Shi *et al.*, 2002b). Surprisingly, as noted above, some of our strongest discriminating variables, in cohort one, were completely uninformative in cohort two – underlying the importance of validation in independent datasets.

6.4 Model Optimization – An Endless Pursuit of Perfection

Once a model has been validated, the goal shifts towards making the model "better." Here, "better" is likely to have different meanings for different groups. In our case, we were interested in four specific improvements: (i) model simplification; (ii) identifying optimal mathematical approaches; (iii) developing pattern recognition versions of the

model; (iv) increasing reliability of the models for more complex samples. Each offers potentially general lessons for this phase of classification study, and each is discussed below.

6.4.1 Model simplification

If all else is equal, simpler models are superior to complex models (Seasholtz and Kowalski, 1993). This is most clearly seen in the extreme case – a single, robust state marker is superior to an equally robust profile based on 500 metabolites. Although we have no state markers in our system, we expect, as noted above, that ~60% of our markers having $p<0.20$ represent statistical type I errors. We therefore reasoned that the same approach used previously, to cut the data set at $p<0.20$, could be used again. We conducted these analyses on the data from the second cohort, and then confirmed that, as expected, HCA and PCA separations were complete. Initial studies in a third cohort showed that the original dataset was able to distinguish diet group with 90% accuracy by HCA, but the shorter "improved" dataset lost accuracy to 63% – essentially equivalent to chance. This suggests that there are markers within the profile that contribute to separation without even reaching p values of 0.20. In other words, in the absence of state markers, larger profiles, built on multiple variables seem to aid considerably in classification.

6.4.2 Choice of Algorithm – Components *vs.* distance

We then used these both the short and long profiles to examine the utility of distance-based models (e.g., HCA) and projection or component-based models – models that include a vectorial or directional component. In contrast to the observations with HCA, PCA was capable of distinguishing groups based on either the longer or the shortened dataset. It must be considered that the separation plane plotted using PCA is subjective and therefore this is not definitive evidence, but this data supports a working technology model in which component or projection-based analyses are more powerful for metabolic profiling. Further work, in part noted below, supported this shift in modeling approaches in our work. Based on this and other metabolomics datasets we have looked at, component-based models generally appear equal or superior to distance based models.

6.4.3 Pattern Recognition

A major potential issue in the use of metabolic profiles is the desire for a profile whose use is as objective as possible and which incorporates past

knowledge. Expert systems/pattern recognition-based approaches simultaneously solve both problems. The use of a training set incorporates previous knowledge, and classification can then be scored without subjectivity, such as choosing a PCA rotation. We examined two different pattern recognition algorithms.

KNN, or k-nearest neighbor analysis (Cover and Hart, 1967), is a supervised version of HCA, with HCA itself being essentially KNN with k=1 (k is the number of neighbors polled for class assignment). KNN is a pure distance-based metric, and one of its great strengths lies in its utility with small training sets. The second approach that we tested is called SIMCA or soft independent modeling of class analogy (Wold, 1976). It is a supervised version of PCA, and thus a component or projection-based metric whose great strength is in modeling flexibility. Briefly, KNN was found to be unusable because it was impossible to set k in advance, whereas SIMCA was effective (Shi *et al.,* 2004). Indeed, SIMCA was more effective than KNN even when we set k after analysis, a statistically invalid way of optimizing analysis. This is strong evidence for preference of component compared to distance based algorithms in metabolomics studies. We note that this does appear somewhat metabolomics specific, as we have been involved in other non-metabolomics informatics studies in which clustering approaches out-performed component based studies.

6.4.4 Increasing reliability of the models for more complex samples

For many metabolomics studies, the dataset on which the models are built and validated are not the sets of final interest. It is therefore of critical importance to build models that are as robust as possible to the types of noise one might encounter in a "field test." As noted, we partially address this issue by a deliberately fuzzy design and sample collection approach. This approach, however, neither mitigated nor prepared us for one major surprise – significant metabolomic differences between cohorts one and two in our study. Although the profiles were robust in that they were able to distinguish diet in the new cohort, combined cohort studies were problematic.

Numerous straight-forward mathematical approaches (normalization, transformation, outlier removal, winsorization, [replacing outliers with defined values, e.g., all values >3 standard deviations from the mean with the value at 3 standard deviations from the mean] etc) were unsuccessful at resolving the cohort separations, telling us that the separations are not mathematical artifacts, but rather they are biological changes due to the differences between cohorts. We were able to solve these problems by switching from descriptive to discriminant techniques, specifically a

projection-based approach called partial least squares projection to latent structures discriminant analysis, or PLS-DA (Sjöström *et al.*, 1986, Ståhle and Wold, 1987, Paolucci *et al.*, 2004b). This projection-based technique is optimized to find the separation between groups rather than to best describe variation in the overall population. In other words, we are no longer looking for profiles of the groups in question, but rather we are looking for profiles of the space between the groups of interest. The models built using these techniques have been optimized empirically and have performed consistently with accuracy exceeding 95%. These optimization studies are beyond the scope of this chapter, but see Paolucci *et al.* (2004b).

7. THE AL-DR SYSTEM AS A MODEL FOR METABOLOMICS STUDY

Earlier in this chapter, we put forward an argument that the AL-DR system has been useful for working through relevant problems in metabolomics. Several characteristics of the system have potentially played a role in this utility. We can exert very tight experimental control over this system, allowing us to have well-defined biological groups to attempt to separate. The analytical chemistry, based on HPLC-coulometry is good, but not ideal, necessitating developing an approach to identify useful variables, which in turn greatly strengthened our approach and facilitated noise removal. The non-spherical profile distributions led to our identification of the problems with using clustering to solve metabolomic profiles. The existence of different cohorts highlighted the need for discriminant-based studies, and highlighted the tremendous sensitivity of these profiles. The subtleness of the AL/DR differences has led to a need to adapt finer level informatics. Although not discussed in detail here, differences that occur with ageing and gender have highlighted physiological influences on the mathematical vectors of interest.

8. SUMMARY

Metabolomics offers a new approach with the potential to complement the classification power offered by proteomics and transcriptomics. We have focused our efforts on identifying serum profiles that reflect nutritive intake, in the hopes that these profiles will have utility in several areas of investigation, relevant for human health. Conversely, the nutritive system we studied also appears to have value as a model system with which to test and

explore potential solutions for difficult problems in metabolomics. Our studies have utilized a highly reliable and reproducible animal model, and the lessons we have learned in building the models needed for our studies may be generally applicable to metabolomics-based classification studies. These lessons may be broadly divided into three classes: (i) conceptual issues (e.g., determining the exact question of interest and future use of classification models, focus on markers against observations, interest in answering descriptive against predictive questions); (ii) practical issues (analytical validation of markers, biological variability, addressing cohort-cohort differences, model validation), and; (iii) informatics issues (clustering compared to components, descriptive compared to discriminant, state markers compared to profiles).

REFERENCES

Beal MF *et al.* Kynurenic acid concentrations are reduced in Huntington's disease cerebral cortex. *J. Neurol. Sci.*, 108: 80-87 (1992).

Butler JA, Heydari AR and Richardson A. Analysis of effect of age on synthesis of specific proteins by hepatocytes. *J. Cell Physiol.*, 141: 400-409 (1989).

Calle EE *et al.* Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. *N. Engl. J. Med.*, 348: 1625-1638 (2003).

Cover T and Hart P. Nearest neighbor pattern classification. *IEEE Trans Information Theory*, 13: 21-27 (1967)

Ellis DI, Broadhurst D, Kell DB, Rowland JJ and Goodacre R. Rapid and quantitative detection of the microbial spoilage of meat using FT-IR spectroscopy and machine learning. *Appl. Environ. Microbiol.*, 68: 2822-2828 (2002)

Gavaghan CL, Holmes E, Lenz E, Wilson ID, Nicholson JK. An NMR-based metabonomic approach to investigate the biochemical consequences of genetic strain differences: application to the C57BL10J and Alpk:ApfCD mouse. *FEBS Lett.*, 484: 169-174 (2000)

Griffin JL *et al.* NMR spectroscopy based metabonomic studies on the comparative biochemistry of the kidney and urine of the bank vole (*Clethrionomys glareolus*), wood mouse (*podemus sylvaticus*), white-toothed shrew (*Crocidura suaveolus*) and the laboratory rat.*Comp. Biochem. Physiol. Pt B*, 127: 357-367 (2000).

Heydari AR *et al.* Age-related changes in protein phosphorylation by rat hepatocytes, *Mech. Ageing Dev.*, 50: 227-248 (1989).

Jarvis RM and Goodacre R. Ultra-violet resonance Raman spectroscopy for the rapid discrimination of urinary tract infection bacteria. *FEMS Microbiol. Lett.*, 232: 127-132 (2004).

Keun HC. Geometric trajectory analysis of metabolic responses to toxicity can define treatment specific profiles. *Chem. Res. Toxicol.*, 17: 579-587 (2004).

Kristal BS and Yu BP. Aging and its modulation by dietary restriction. In *Modulation of aging processes by dietary restriction.* 1st ed. Edited by B. P. Yu. Boca Raton: CRC press, Inc (1994).

Lee CK *et al.* Gene expression profile of aging and its retardation by caloric restriction. *Science,* 285: 1390-1393 (1999).

LeWitt PA *et al.* Markers of dopamine metabolism in Parkinson's disease: The Parkinson's Study Group. *Neurology*, 42: 2111-2117 (1992).

Lindon JC *.et al.* Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project. *Toxicol. Appl. Pharmacol.*, 187: 137-146 (2003).

Lopez-Diez EC and Goodacre R. Characterization of microorganisms using UV resonance Raman spectroscopy and chemometrics. *Anal.Chem.*, 76: 585-591 (2004).

Matson WR *et al.* Generating and controlling multiparameter databases for biochemical correlates of disorders. In *Basic, clinical and therapeutic aspects of Alzheimer's and Parkinson's diseases.* Vol. Volume II. New York: Plenum (1990).

Matson WR *et al.* EC array sensor concepts and data. *Life Sci.*, 41: 905-908 (1987).

Matson WR *et al.* n-electrode three dimensional liquid chromatography with electrochemical detection for determination of neurotransmitters. *Clin. Chem.*, 30: 1477-1488 (1984).

McCay CM. Cellulose in the diets of rats and mice. *J. Nutr.*, 10: 435-447 (1935).

McCay CM, Crowell MF and Maynard LA. The effect of retarded growth upon the length of lifespan and upon the ultimate body size. *J. Nutr.*, 10: 63-79 (1935).

Ogawa T *et al.* Kynurenine pathway abnormalities in Parkinson's disease. *Neurology*, 42: 1702-1706 (1992).

Paolucci U *et al.* Development of biomarkers based on diet-dependent metabolic serotypes: Concerns and approaches for cohort and gender issues in serum metabolome studies. *OMICS J. Integr. Biol.*, 8: 209-220, (2004a).

Paolucci U *et al.* Development of biomarkers based on diet-dependent metabolic serotypes: Characteristics of component-based models of metabolic serotype. *OMICS J. Integr. Biol.*, 8: 221-238 (2004b).

Seasholtz MB and Kowalski B. The parsimony principle applied to multivariate calibration. *Anal. Chim. Acta*, 277: 165-177(1993)

Shi H *et al.* Development of biomarkers based on diet-dependent metabolic serotypes: Practical issues in development of expert system-based classification models in metabolomic studies. *OMICS J. Integr. Biol.*, 8: 197-208, (2004).

Shi H *et al.* Attention to relative response across sequential electrodes improves quantitation of coulometric array. *Anal. Biochem.*, 302: 239-245 (2002a).

Shi H *et al.* Characterization of diet-dependent metabolic serotypes: Primary validation of male and female serotypes in independent cohorts of rats. *J. Nutr.*, 132: 1039-1046 (2002b).

Shi H *et al.* Characterization of diet-dependent metabolic serotypes: Proof of principle in female and male rats. *J. Nutr.*, 132: 1031-1038 (2002c).

Sjöström M, Wold S and Söderström B. PLS discriminant plots. In *Pattern recognition in practice II.* E.S.Gelsema and L.N.Kanal, Eds., Elsevier, Amsterdam, p.486 (1986).

Ståhle L and Wold S. Partial least squares analysis with cross-validation for the two-class problem A Monte Carlo study. *J. Chemometrics*, 1: 185-196 (1987).

Vaidyanathan S and Goodacre R. Metabolome and proteome profiling for microbial characterization. In *Metabolic profiling: Its role in biomarker discovery and gene function analysis.* G. G. Harrigan and R. Goodacre (Eds.) Kluwer Academic Publishers, Boston (2004)

Vigneau-Callahan KE *et al.* Characterization of diet-dependent metabolic serotypes: analytical and biological variability issues in rats. *J. Nutr.*, 131: 924S-932S (2001).

Weindruch R and Walford, R. *The retardation of aging and disease by dietary restriction.*: Charles C. Thomas (Ed.), Springfield, IL (1988).

Weindruch RH and Walford RL. Dietary restriction in mice beginning at one year of age: Effects on life-span and spontaneous cancer incidence. *Science*, 215: 1415-1418 (2000).

Willett WC. Diet and cancer: one view at the start of the millennium. *Cancer Epidemiol. Biomark. Prev.*, 10: 3-8 (2001).

Willett WC, Dietz WH and Colditz GA. Guidelines for healthy weight. *N. Engl. J. Med.*, 341: 427-434 (1999).

Winder CL *et al.* The rapid identification of *Acinetobacter* species using Fourier transform infrared spectroscopy. *J. Appl. Microbiol.*, 96: 328-339 (2004).

Wold S. Pattern Recognition by means of disjoint principal components models. *Patt. Recogn.*, 8: 127-139 (1976)

Yu BP. *Modulation of Aging Processes by Dietary Restriction.* Boca Raton: CRC Press (1994).

Yu BP. Aging, oxidative stress: Modulation by dietary restriction. *Free Radic.Biol.Med.*, 21: 651-668 (1996).

Chapter 12

MODELLING OF FUNGAL METABOLISM

Towards genome scale models

Helga David and Jens Nielsen
Center for Microbial Biotechnology, BioCentrum-DTU, Technical University of Denmark, DK-2800 Kgs Lyngby, Denmark

1. MODELLING CELLULAR METABOLISM

Mathematical modelling of cellular and metabolic systems dates back to the 1960s (Goodwin, 1963). From a physiological perspective, the aim of applying mathematical models to such systems is to provide further insight into the general principles governing cellular function. More recently, models have been developed in connection with targeted methods for the improvement of the metabolic capabilities of industrially important organisms – an approach referred to as metabolic engineering (Nielsen, 2001) - wherein the focus is on the rational engineering design (Wiechert, 2002).

1.1 Types of models

Metabolic models may be classified into two main categories, namely stoichiometric models and kinetic models. The former models rely exclusively on time invariant properties of metabolic networks, whereas the latter ones also incorporate kinetics (Gombert *et al.*, 2000; Patil *et al.*, 2004).

Since the stoichiometry of metabolic reactions is generally known, it is less complicated to build stoichiometric models, compared to kinetic models, as the development of the latter is hampered by a lack of knowledge of the kinetics of major processes, such as oxidative phosphorylation or transport. An issue that needs to be taken into account when building kinetic models is related to the validity of transferring reaction mechanisms of

enzymes observed *in vitro* to *in vivo* conditions, in addition to the estimation of usually a large number of kinetic parameters (Gombert *et al.*, 2000; Wiechert, 2002). In spite of these shortcomings, kinetic models can be applied to simulate the dynamics of metabolic systems, while stoichiometric models are restricted to describing (pseudo) steady state metabolic behaviour.

1.2 Model-based methods

A number of methods have been used for analysis of cellular function. Of special importance in the context of this chapter are constraint-based or flux-based modelling approaches. These focus on the quantification of metabolic fluxes, which are regarded as the final representation of a certain physiological state of the cell, resulting from different levels of cellular regulation (Stephanopoulos, 1999; Nielsen, 2003). These approaches make use of stoichiometric models and rely on the principle of mass conservation, and on the assumption of quasi-stationarity. Among the most frequently used methods are Metabolic Flux Analysis (MFA) (Stephanopoulos *et al.*, 1998), Metabolic Network Analysis (MNA) (Wiechert and Graaf, 1996; Szyperski, 1998; Christensen and Nielsen, 2000; Wiechert, 2001), and Flux Balance Analysis (FBA) (Varma and Palsson, 1994; Bonarius *et al.*, 1997; Schilling *et al.*, 1999a), which allow the determination of a particular flux distribution; as well as Extreme Pathway Analysis (Schilling *et al.*, 1999b, 2000) and Elementary Flux Mode Analysis (Schuster and Hilgetag, 1994; Schuster *et al.*, 2000), which provide a characterization of all possible flux distributions satisfying the mass balance constraints.

In addition, simulation methods relying on kinetic models have been developed, namely based on stationary and nonstationary mechanistic models, and models with gene regulation (Wiechert, 2002). The development of such models and modelling tools does, however, suffer from the limitations mentioned above and from the mathematical and computational complexities associated with dynamic models.

1.3 Focus of the present chapter

Whole-cell simulation is the ultimate challenge for the analysis of overall cellular function and several efforts are being undertaken at present to construct whole-cell kinetic models and to develop appropriate computer tools (Tomita, 2001). However, progress with these models has been hindered by the problem of lack of kinetic parameters, which is even more pronounced on a whole-genome scale. As a result, currently available *genome-scale models* or *metabolic whole cell models* are stoichiometric

models that include all metabolic reactions known to occur in an organism. When used in combination with flux-based approaches, genome-scale models provide a means of relating the organism's genotype to its phenotype, through the quantification of metabolic fluxes, and hence represent an important contribution to the field of functional genomics.

This chapter will focus on genome-scale models of organisms, in particular those developed for fungal systems, on their construction, properties and applications. Furthermore, attention will be drawn on the significance of these models for the process of integration of biological data from all levels of metabolism (genome, transcriptome, proteome, metabolome, fluxome, etc.), aiming at a systems level understanding of cellular function.

2. GENOME-SCALE MODELS

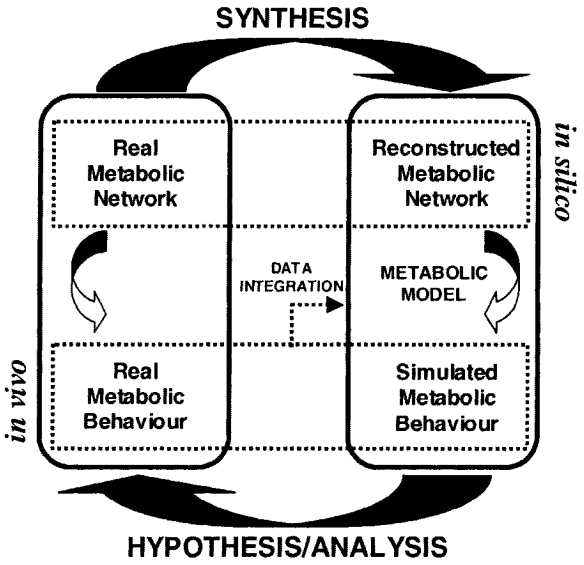


Figure 1. The iterative process of genome-scale model building. Once the metabolic network has been reconstructed and the genome-scale model has been set up, hypotheses about microbial behaviour can be generated, using constraint-based simulation methods. These hypotheses can be tested, by performing experiments, and subsequently the model can be refined such that its predictions are in closer agreement with experimental observations. Furthermore, data from different levels of cellular processes may be incorporated into the genome-scale model for improvement of model predictions.

Genome-scale models are based on complete metabolic reconstructions of the underlying networks (as far as they are known). The modelling process comprises several steps, starting with the reconstruction of the metabolic network, followed by the development of the corresponding stoichiometric model and subsequently the analysis of the metabolic network, using constraint or flux-based approaches. Thereby, quantitative hypotheses about microbial behaviour may be generated and evaluated experimentally to produce observations. These observations may extend the knowledge of the metabolic system, and hence may be fed into the model, which in this manner broadens in scope and predictive power. The process of building genome-scale models is therefore iterative. Figure 1 depicts the process of model development, and a detailed description of the steps involved is provided in the following sections.

2.1 Current models and their properties

Several genome-scale microbial models have been developed and published to date, namely, for the bacteria *Escherichia coli* (Edwards and Palsson, 2000), *Haemophilus influenzae* (Edwards and Palsson, 1999), and *Helicobacter pylori* (Schilling *et al.*, 2002), as well as the yeast *Saccharomyces cerevisiae* (Förster *et al.*, 2003). Table 1 presents the properties of these models along with some relevant characteristics of the microorganisms' genomes.

Table 1. Characteristics of current genome-scale metabolic networks in the context of the genomic data available.

Organism	Genome length (Mb)	Total no. ORFs	No. of metabolic ORFs (% of total)	No. of metabolic reactions	No. of metabolites
Helicobacter pylori[1]	1.66	1,590	291 (18%)	390	340
Haemophilus influenzae[1]	1.83	1,743	412 (24%)	461	367
Escherichia coli[1]	4.60	4,401	660 (15%)	720	436
Saccharomyces cerevisiae[2]	12.16	5,773	708 (12%)	1175	584

[1] (Price *et al.*, 2003).12
[2] (Nielsen, 2003).

As can be observed from Table 1, for all four organisms, the fraction of open reading frames (ORFs) in their genomes that code for proteins directly implicated in cellular metabolism is low, and there is a downward trend for

increasing sizes of the genome. Conversely, the fraction of ORFs coding for proteins involved in regulation is larger for higher organisms. Nevertheless, cellular metabolism is influenced by a considerable number of ORFs, directly or indirectly, through regulation of gene expression and enzyme activities (Nielsen, 2003).

Another important observation that can be made from the examination of reconstructed metabolic networks is the tight connection among the different parts of the metabolism, carried out by metabolites (e.g. cofactors) that are common to various metabolic reactions (Figure 2). As a result of this close interrelatedness, a perturbation in a given part of the metabolic network may be manifested as changes in many other parts of the metabolism. Thereby, information about the function of the complete metabolic network may be obtained from measurement of for instance, a few metabolic fluxes (Nielsen, 2003).

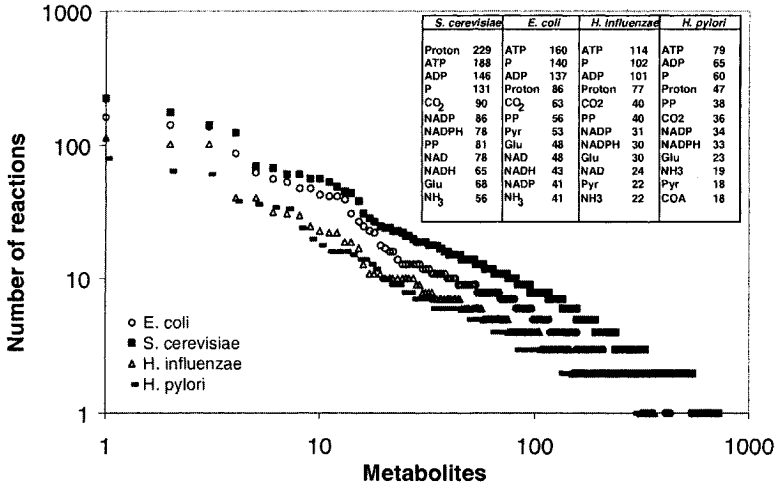


Figure 2. Frequency plot of the number of reactions that each metabolite appears in for the four reconstructed metabolic networks. For each metabolic network, the 10 metabolites that appear in the most reactions are listed (Nielsen, 2003).

2.2 Reconstruction of metabolic networks

The network of metabolic reactions of a given organism may be reconstructed from the annotated genome sequence and available

experimental data on the biochemistry and physiology of the organism under study.

Genomic data form the skeleton of the network. Therefore, the first step in the reconstruction process should be a detailed analysis of the organism's genome to identify metabolic genes that code for enzymatic reactions and transport processes. A number of genomic database, which provide information on the genome sequences, their annotations, and the corresponding metabolic maps for several organisms are accessible online (Table 2). Current genome-scale models rely on 60-70% of complete genome annotation (Price *et al.*, 2003).

Metabolic reactions for which there is no genomic information available may be identified through experimental and literature investigations of the biochemistry and physiology of the organism of interest. By this means, additional metabolic functions (reactions or pathways) are included in the reconstructed metabolic network, which can or cannot be subsequently assigned to orphan genes (ORFs) in the genome. Hence, model building may by itself play a significant role in the field of functional genomics. In addition to this, biochemical and physiological data may validate and complement information on already annotated metabolic genes (ORFs) (Covert *et al.*, 2001a).

Table 2. Databases available on-line with genomic information and links to metabolic maps.

Database	**Web address**
EcoCyc	http://ecocyc.org
MetaCyc	http://metacyc.org
Metabolic pathways database (MPW)	http://ergo.integratedgenomics.com/MPW
Kyoto Encyclopedia of Genes and Genomes (KEGG)	http://www.genome.ad.jp/kegg
What Is There (WIT)	http://wit.mcs.anl.gov
Biology Work Bench	http://workbench.sdsc.edu
Enzymes and Metabolic Pathways database (EMP)	http://www.empproject.com

Biochemical data have particular importance in the annotation of species-specific genes, namely those that code for enzymes with similar functions in different organisms, but whose sequences differ. In these cases, no conclusions can be drawn through sequence comparison. On the other hand, indication of the presence of certain metabolic reactions or pathways may be provided by the physiological characteristics of the organism, such as growth on a given substrate or secretion of a given product. These reactions should be incorporated in the metabolic reconstruction even if there are no genomic data supporting their inclusion (Covert *et al.*, 2001a).

Once the metabolic network has been reconstructed from the annotated genome sequence and experimentally determined biochemical and

physiological characteristics of the organism, a mathematical model may be formulated and used for simulation of cellular behaviour.

2.3 Model development and analysis of metabolic networks

In current stoichiometric genome-scale models, the metabolic reactions and metabolites included in the reconstructed network are interrelated through mass balances around intracellular metabolites, whose concentrations are assumed to be in (pseudo) steady state. Thereby, a system of equations - stoichiometric constraints - is defined, which is underestimated, as the number of reactions or fluxes is greater than the number of metabolites or mass balances. Hence, the number of degrees of freedom in the equation system is large and a multiplicity of solutions (flux distributions) exists. Further information about the metabolic system may be incorporated in the model, such as thermodynamics (reversibility of reactions) and enzyme capacities, resulting in additional constraints that reduce the number of possible solutions (flux distributions).

In order to use genome-scale models for simulating cellular behaviour, it is necessary to further constrain the set of fluxes. This can be done by measuring a set of fluxes. But, as in the case of genome-scale models, since the number of degrees of freedom is normally very large, a large number of fluxes need to be measured so as to find a solution to the equation system. A more commonly applied method is therefore Flux Balance Analysis (Varma and Palsson, 1994), wherein the underestimated equation system is formulated as an optimisation problem. This method allows the determination of a particular solution – an optimal flux distribution, by stating a suitable optimisation criterion (objective function) and employing linear programming methods (Bertsimas and Tsitsiklis, 1997). A number of physiologically meaningful optimisation criteria have been used, such as maximal growth rate (Edwards *et al.*, 2001) or maximal formation rate of a particular product (Varma and Palsson, 1993a).

Network-based pathway analysis methods have also been used in connection with genome scale-models for the analysis of metabolic networks, namely Extreme Pathway Analysis and Elementary Flux Mode Analysis (Schilling and Palsson, 2000; Papin *et al.*, 2002; Price *et al.*, 2002; Schilling *et al.*, 2002). Herein, convex analysis (Rockafellar, 1970) is employed to define a unique set of systemic pathways that provide information about the topology of the network. From these systemic pathways, it is possible to describe all possible steady state flux distributions and characteristics of the metabolic network (Schuster and Hilgetag, 1994; Schuster *et al.*, 2000).

2.4 Applications of genome-scale models

Genome scale-models have been used in connection with flux-based modelling approaches in a number of studies, namely for the understanding of cellular objectives and prediction of optimal phenotypic behaviour, for simulating the phenotypic effects of genetic manipulations (e.g. gene addition and deletion studies, network robustness studies), for assessing pathway redundancy, and for gaining insight into regulation (e.g. identification of systemically correlated reactions). In the following, emphasis will be put on fungal networks (see section 3 - Fungal models), while detailed descriptions on methods and applications of genome-scale models concerning other systems may be found in the recent reviews by Patil *et al.* (2004) and Price *et al.* (2003).

2.5 Modelling considerations

Simulation of cellular behaviour using the described modelling framework should be carried out cautiously, as a number of issues arise in connection with the scope and validity of the genome scale-models and flux-based methods.

On one hand, the reconstructed metabolic network may not correspond to the real network, since all the metabolic genes are not necessarily taken into consideration. Thus, there may be metabolic reactions/functions missing in the model. Conversely, some of the metabolic reactions assumed to be present in the reconstructed network do not necessarily take place in the cell. Reactions included in the reconstruction based on indirect physiological evidence, insufficient biochemical knowledge (e.g. unknown cofactor requirements) or wrong annotations are likely to fall into this situation.

Furthermore, these models rely only on information about reaction stoichiometry, and they do not take into account any regulatory mechanisms. Consequently, they are based on the assumption that all genes are constitutively expressed, and hence the model represents all possible phenotypes the cell may express. This may lead to wrong predictions, if the metabolic model is not complemented with additional experimental information (see section 4 - Integrative Analysis: Future outlook) or the optimization problem is not properly defined.

In stoichiometric models, growth or biomass production is generally regarded as a drain of biosynthetic precursors required to produce cellular components, and the demands on these compounds are estimated based on biomass composition. In general, a single overall equation denoting growth is considered in the models, not taking into account the variation of the cellular composition with the growth rate or medium composition. The

sensitivity of the biomass yield to perturbations in the biosynthetic demands has been assessed in different studies and some authors concluded that the biomass yield was not overly sensitive to changes in biosynthetic requirements (Varma and Palsson, 1993b), whereas others emphasized the importance of incorporating changes in biomass composition with specific growth rate in flux estimation (Pramanik and Keasling, 1997).

Simulation of growth of auxotrophic strains, wherein the uptake of more than one carbon source has to be considered, is also a problem that is difficult to address using this modelling framework, particularly when auxotrophic requirements are unknown.

In addition, the assumption of a quasi-stationary level of all metabolites restricts the applicability of these modelling approaches, not allowing the description of dynamic behaviour of cells.

When Flux Balance Analysis is applied to genome-scale models, wherein linear programming methods are used to determine an optimal flux distribution, an additional assumption has to be considered, namely that in the process of evolution certain criteria have been maximized, such as the specific growth rate. It has been demonstrated that wild-type strains of *E. coli* have evolved towards optimal growth (Edwards *et al.*, 2001; Ibarra *et al.*, 2002; Burgard and Maranas, 2003). However, this assumption might not hold for genetically modified organisms. In fact, it has been shown that knock-out mutant strains grow suboptimally, at least initially, using the metabolic networks in a very similar way to reference strains (Segre *et al.*, 2002). Moreover, optimal solutions found in this way are not necessarily unique, i.e., there may exist multiple optimal solutions wherein the same objective can be achieved through different flux distributions. Mahadevan and Schilling (2003) investigated the effects of alternate optimal solutions on the predicted range of flux distributions in constraint-based genome-scale metabolic models, and concluded that the extent of flux variability was highly dependent on environmental conditions and network composition. Their work highlights the need for developing approaches capable of discriminating biologically significant flux distributions based on additional experimental data, such as expression data.

In the light of the above, application of stoichiometric modelling approaches should be accompanied by a critical evaluation of the model validity and predictive power. In spite of the limitations mentioned, genome-scale models are extremely valuable, since they represent structured information about the metabolism in a given cellular system and thereby assist in focusing on central questions concerning cellular function (Wiechert, 2002). Furthermore, they constitute a starting point for integration of data of different types of sources (genome, transcriptome,

proteome, metabolome, fluxome), required for the understanding of overall cellular function (see section 4 - Integrative Analysis: Future outlook).

3. FUNGAL MODELS

Fungi are characterised by remarkable biochemical versatility, manifested in the production of a wide range of acids and degradative enzymes that sustain their absorptive mode of nutrition, as well as a large number of low molecular weight primary and secondary metabolites. Many of these metabolites have industrial and pharmaceutical applications (http://gene.genetics.uga.edu/white_papers/fgi.html).

Even though 24 fungal genome sequences have been published to date (http://www.ncbi.nlm.nih.gov), only one genome-scale fungal model has been reported so far, namely for the yeast *Saccharomyces cerevisiae* (Förster *et al.*, 2003). In addition to being the first eukaryote to have its genome completely sequenced (Goffeau *et al.*, 1996), studies on the budding yeast have benefited from the availability of powerful genetic, biochemical and molecular biological tools that have facilitated the assignment of function to a great majority of its genome. This has assisted in unravelling the encoded metabolism and subsequently in the development of the first genome-scale model for an eukaryotic organism.

However, yeast is not an adequate model for analysing the overall cellular function of filamentous fungi. The most detailed stoichiometric model for a filamentous fungus currently available was developed for the industrially relevant organism *Aspergillus niger* (David *et al.*, 2003). The metabolic reconstruction was based on genomic data existing in the literature (in the absence of a publicly accessible genomic sequence), and complemented with biochemical and physiological information. The reconstructed network provides a detailed description of the central carbon metabolism of *A. niger*, namely of the metabolism of carbohydrates, organic acids, polyols and other alcohols, and aminosugars, as well as the oxidative phosphorylation in the electron transport chain.

The genome of *A. niger* is almost three times larger than that of *S. cerevisiae* (34.5 Mb in *A. niger* and 12.1 Mb in yeast (GOLD database, Integrated Genomics)). However, this ratio seems to be lower with respect to the number of ORFs identified (14,000 ORFs in *A. niger* (GOLD database, Integrated Genomics) compared with 5,773 ORFs in yeast (Nielsen, 2003)). While the genomic information existing for *A. niger* is much more limited than for yeast (owing to fewer studies accomplished for the filamentous fungus and to the fact that its genomic sequence and annotation are in possession of private companies), many of the genes in the yeast genome are

still questionable (Salzberg, 2003). Hence any statistics presented might not correspond to reality, and conclusions based on these numbers should be drawn with precaution. Nevertheless, it can be speculated that the larger genome of *A. niger* and higher number of ORFs endow this organism with more extensive metabolic capabilities. In fact, physiological evidence supports this statement, if one thinks about the wide variety of carbon compounds that can be used by *A. niger* as the sole carbon source for growth, and the range of products derived from its metabolism (particularly with respect to secondary metabolites).

3.1 Network properties

In *S. cerevisiae*, more than 60% of the identified ORFs have been characterised (Saccharomyces Genome Database), whereas only about 40% of the ORFs identified in *A. niger* have been annotated by DSM. The annotated genes in the genomes of these fungi have been classified into functional categories and some statistics are presented in Table 3, along with the characteristics of the metabolic networks reconstructed for *S. cerevisiae* and *A. niger*.

In the reconstructed metabolic networks for *S. cerevisiae* and *A. niger*, intracellular compartmentation is considered and consequently reactions and metabolites are distributed among the extracellular medium and the intracellular compartments, namely cytosol and mitochondria (as well as glyoxysomes, in the metabolic network reconstructed for the filamentous fungus). Thus, besides biochemical conversions, the metabolic network also includes transport processes between the different compartments and between the cell and the environment. Even though there are a vast number of studies on transport in fungi, it is still a long way to go before there is complete understanding of the principles and mechanisms of metabolite transport. In many cases, even the stoichiometry of transport processes is still unknown, which represents an additional obstacle to the development of detailed metabolic models.

3.2 Functional properties

Biochemical conversions comprising the metabolic reconstructions have been classified into six main classes of enzymes, according to the type of transformation in question (David *et al.*, 2003; Förster *et al.*, 2003). The involvement of each class of enzymes in the reconstructed carbohydrate metabolism of *S. cerevisiae* and *A. niger* has been assessed and a comparison is presented in Figure 3.

Table 3. Characteristics of the genome and the metabolic networks reconstructed for the fungi *S. cerevisiae* and *A. niger*.

Organism	Dataset	Genomic characteristics		Metabolic reconstruction	
		No. of metabolic ORFs (% of total no. ORFs)	No. ORFs in metabolism of carbohydrates (% of no. of metabolic ORFs)	No. of metabolic reactions	No. of metabolites
Saccharomyces cerevisiae	MIPS[a]	1156 (20%)	414 (36%)		
	Metabolic reconstruction[d]	708[b] (12%)	189[b] (16%)	1175	584
Aspergillus niger	DSM	3111 (22%)	209 (7%)		
	Metabolic reconstruction[e]	20[c] (0.1%)	20[c] (0.6%)	355	284

[a] Munich Information Center for Protein Sequences.
[b] based on genomic data.
[c] based on literature surveys.
[d] (Förster *et al.*, 2003).
[e] (David *et al.*, 2003).

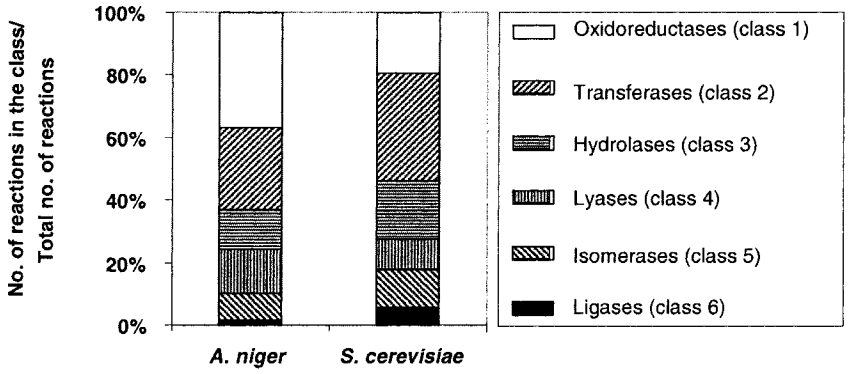


Figure 3. Comparison of relative contributions of different enzyme classes in the reconstructed carbohydrate metabolism of *S. cerevisiae* (125 reactions) and *A. niger* (172 reactions).

Oxidoreduction reactions, which are catalysed by oxidoreductases (class 1), represent the predominant group of biochemical transformations in the carbohydrate metabolism of *A. niger* (37%), whereas in yeast the major part of the reactions are catalysed by transferases (class 2), which account for 34% of the total number of reactions participating in the metabolism of carbohydrates. In other words, reactions involving the transfer of a group from one compound to another (transferases) are the second most abundant enzymes in the carbohydrate metabolism of *A. niger*, while for yeast this position is occupied by oxidoreductases (class 1) together with hydrolases (class 3). Lyases (class 4) and isomerases (class 5) have contributions in the range of 8-14% to the carbohydrate metabolism in both fungal networks, as does hydrolases (class 3) for *A. niger*. Ligases (class 6) are the least involved class in the carbohydrate metabolism of both fungi. Thus, the relative contributions of the different classes of enzymes to the reconstructed carbohydrate metabolism of *S. cerevisiae* and *A. niger* seem to follow the same trend, since, in both fungi, oxidoreductases and transferases correspond to more than half of the reactions, while ligases represent the least abundant enzymes. The major difference seems to be the high occurrence of reactions catalysed by hydrolases in yeast, when compared to the carbohydrate metabolism of *A. niger*.

Additionally, the substrate specificity of the different groups of enzymes included in the metabolic reconstructions has been evaluated based on the ratio of the number of reactions to the number of enzymes in each category. In *A. niger*, transferases appear to be the enzymes with the lowest substrate specificity, followed by oxidoreductases, isomerases and lyases, whereas hydrolases and ligases seem to have high substrate specificities, each of them catalysing only one reaction (David *et al.*, 2003). Isomerases and transferases are the less substrate specific enzymes in *S. cerevisiae* (Förster *et al.*, 2003).

3.3 Topological properties

The structures of the reconstructed metabolic networks of *S. cerevisiae* and *A. niger*, in particular concerning the carbohydrate metabolism, were addressed quantitatively, and it was observed that both follow a power law distribution (Figure 4), and hence embody scale-free networks (Jeong *et al.*, 2000).

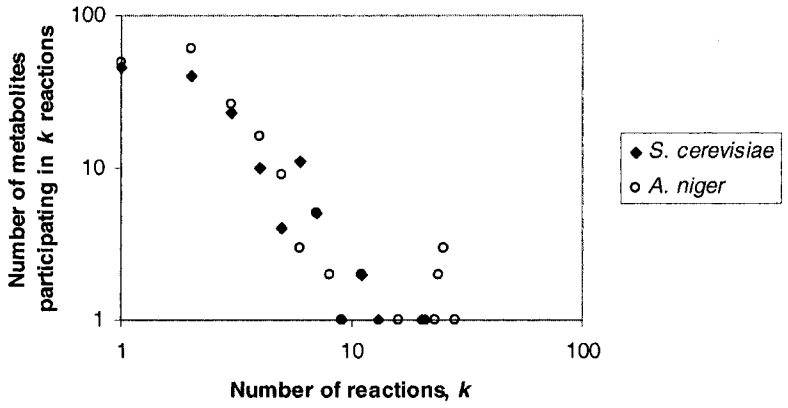


Figure 4. Comparison of connectivity distributions concerning the reconstructed carbohydrate metabolism of *S. cerevisiae* (125 reactions) and *A. niger* (172 reactions).

This class of networks has been shown to be characterised by robustness and tolerance to random failures. However, if the most connected substrates in the network are subject to perturbations, the effects will be reflected all over the network (Albert *et al.*, 2000; Jeong *et al.*, 2000).

Table 4. List of the 10 most connected metabolites in the reconstructed carbohydrate metabolism of *S. cerevisiae* and *A. niger*.

Saccharomyces cerevisiae		*Aspergillus niger*	
Metabolite	**No. of reactions**	**Metabolite**	**No. of reactions**
ATP	21	ATP	25
ADP	20	NAD^+	25
D-Fructose 6-phosphate	13	NADH	25
NAD^+	11	$NADP^+$	24
NADH	11	NADPH	24
D-Galactose	9	ADP	23
D-Glucose	7	Orthophosphate	16
D-Glucose 6-phosphate	7	D-Fructose 6-phosphate	11
Orthophosphate	7	CO_2	11
CO_2	7	Pyruvate	9

The substrates participating in the carbohydrate metabolism of both fungi were ranked according to their frequency of participation in reactions. The ranking of the most connected substrates is almost identical in *S. cerevisiae* and *A. niger*, with the most metabolites participating in most reactions being cofactors (Table 4).

Furthermore, the number of metabolites participating in each reaction of the reconstructed carbohydrate metabolism of *S. cerevisiae* and *A. niger* was assessed (Figure 5). For both fungal networks, the majority of the reactions encompass four metabolites and are catalysed by oxidoreductases or transferases, which involve the transfer of electrons or groups respectively from one compound to another.

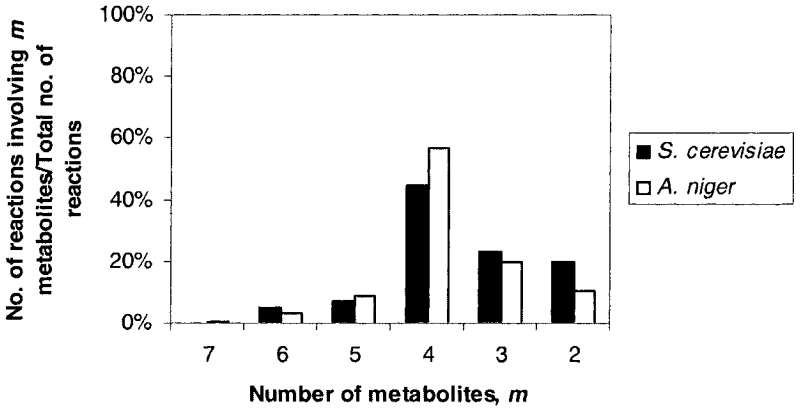


Figure 5. Comparison of the number of reactions as a function of the number of metabolites involved, in the reconstructed carbohydrate metabolism of *S. cerevisiae* (125 reactions) and *A. niger* (172 reactions).

3.4 Reaction deletion analysis

In order to study the essentiality of the biochemical reactions in the reconstructed carbohydrate metabolism of *S. cerevisiae* and *A. niger*, each individual reaction was deleted from the metabolic network and optimal growth on glucose was simulated for the corresponding mutant. As shown in Figure 6, for both fungi, only a small number of biochemical reactions are essential for growth on glucose, reflecting the flexibility of the metabolic networks to meet the biosynthetic requirements, as well as the fact that many of the reactions are not active during growth on this carbon source.

In general, deletion of reactions essential for growth of *A. niger* on glucose has no effect on the growth of yeast on glucose, and *vice versa*. In particular, it is observed that if any of the reactions participating in the biosynthetic pathway of chitin are removed from the metabolic network of

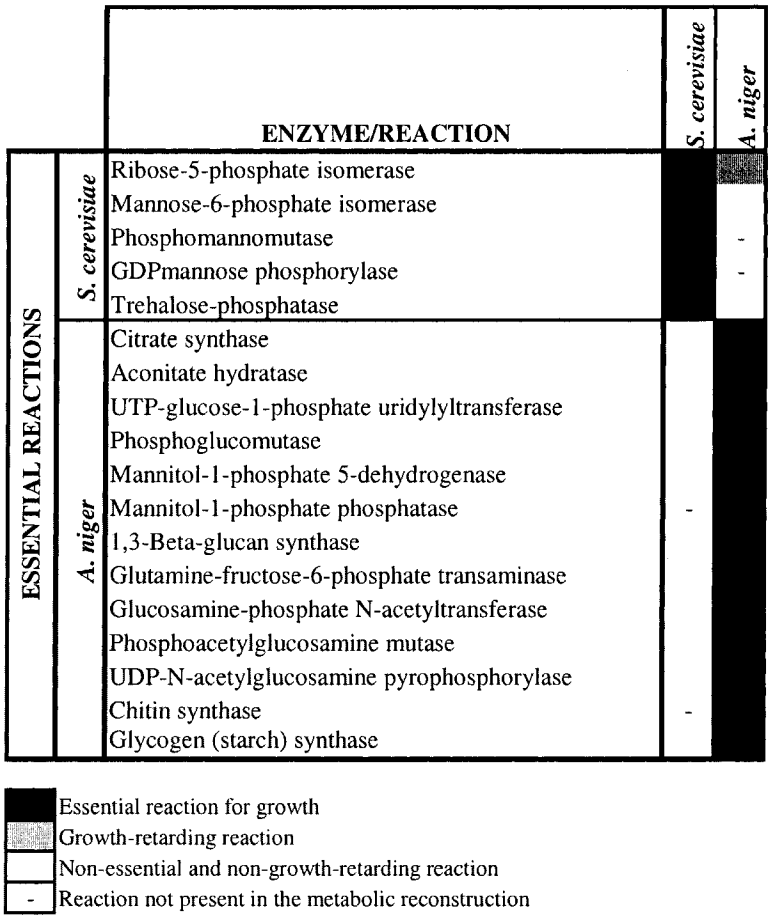


Figure 6. Essential reactions in the reconstructed carbohydrate metabolism of *S. cerevisiae* and *A. niger* for growth on glucose.

A. niger, it will not be able to grow since this is the only pathway leading to the formation of chitin, which is a component of its cell wall. However, as chitin does not make part of the cellular composition of yeast (or if present, is at a very low extent), deletion of these reactions does not have an effect on the growth of this microorganism. In contrast, removal of reactions involved in the biosynthesis of mannan from the metabolic network reconstructed for *S. cerevisiae* prevents it to grow, whereas it has no effect on the growth of *A. niger*. A similar explanation may be given for this observation; specifically, this polysaccharide does not enter in the biomass composition of the filamentous fungus. On the other hand, removal, from the metabolic

network, of reactions that are growth-retarding for one of the organisms has either a retarding effect or no effect on the growth of the other organism (results not shown).

The effects of inactivating specific reactions on the metabolic capabilities of *A. niger* to produce several metabolites have been assessed and illustrated with the case study of succinate production (David *et al.*, 2003). Deletions leading to high product formation at optimal growth were computed and optimal single and double deletion mutants identified. The results suggested that the genes encoding the enzymes in question might be potential targets for metabolic engineering.

4. INTEGRATIVE ANALYSIS: FUTURE OUTLOOK

Cellular behaviour is the end result of complex interactions occurring among the different components of the biological system. Therefore, in order to understand the principles that govern cellular function or to design strains efficiently with improved properties, an integrative approach should be undertaken, wherein the focus is on the system properties, rather than on the properties of the individual parts - an approach referred to as systems biology (Palsson, 2000; Ideker, 2001).

Genome-scale models can play an important role in the integration of genome-scale data that depict the systemic properties of the cell (*e.g.* genome, transcriptome, proteome, and metabolome). The iterative process of model building provides a suitable framework for integrating biological data supplied by high-throughput analytical methods (see Figure 1). In this way, models broaden in scope and validity, leading to improved predictions of phenotypes.

The genome-scale models developed to date are still at an embryonic stage, as they rely basically on the stoichiometry of reactions and thus provide essentially structural information about metabolism. Nevertheless, these models constitute a natural starting-point for incorporating other features of the genome determining cellular behaviour, such as regulation. The incorporation of "omics" data other than genomics requires the development of more advanced modelling frameworks. Several efforts have been carried out recently to combine stoichiometric models with additional data from different levels of cellular processes, namely metabolite profiles (Förster *et al.*, 2002) and gene expression data (Covert *et al.*, 2001b; Åkesson *et al.*, unpublished). Moreover, stoichiometric approaches have been extended with additional features for improving predictions, such as the use of FBA in combination with Boolean operators (Covert *et al.* 2001b; Covert and Palsson, 2002, 2003) or the use of dynamic FBA (Mahadevan *et*

al., 2002), in order to account for regulation, as well as the use of FBA in combination with Energy Balance Analysis (EBA) (Beard *et al.*, 2002), so as to incorporate thermodynamics (for reviews see Kauffman *et al.* (2003) and Patil *et al.* (2004)).

Improvement of the prediction performance of genome-scale models will benefit both fields of functional genomics and metabolic engineering, allowing a better analysis and interpretation of the different "omics" data, and a more efficient design of engineered strains with improved properties respectively.

Rapid progress is being made towards analysing cells as complete systems, wherein genome-scale data are being generated by high-throughput analytical methods, and genome-scale models as well as simulation methods are evolving in order to incorporate and analyse simultaneously all available biological knowledge, which will ultimately make possible the simulation of whole-cell behaviour.

ACKNOWLEDGMENTS

The authors thank Mats Åkesson (Novo Nordisk A/S) and Vijayendran Raghevendran (DTU) for valuable comments on the manuscript. Financial support was provided in part by Fundação para a Ciência e a Tecnologia, Portugal, through a research fellowship for H. David.

REFERENCES

Albert R, Jeong H and Barabasi AL. Error and attack tolerance of complex networks. *Nature.*, 406: 378-382 (2000).

Beard DA, Liang SD and Qian H. Energy balance for analysis of complex metabolic networks. *Biophys. J.*, 83: 79-86 (2002).

Bertsimas D and Tsitsiklis JN. *Introduction to Linear Optimization*, Athena Scientific, Belmont (1997).

Bonarius HPJ, Schmid G and Tramper J. Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. *Trends Biotechnol.*, 15: 308-314 (1997).

Burgard AP and Maranas CD. An optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnol. Bioeng.*, 82: 670-677 (2003).

Christensen B and Nielsen J. Metabolic network analysis: A powerful tool in metabolic engineering. *Adv. Biochem. Eng. Biotechnol.*, 66: 209-231 (2000).

Covert MW, Schilling CH, Famili I, Edwards JS, Goryanin II, Selkov E and Palsson BO. Metabolic modeling of microbial strains *in silico*. *Trends Biochem. Sci.*, 26: 179-186 (2001a).

Covert MW, Schilling CH and Palsson BO. Regulation of gene expression in flux balance models of metabolism. *J. Theor. Biol.*, 213: 73-88 (2001b).

Covert MW and Palsson BO. Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *J. Biol. Chem.*, 277: 28058-28064 (2002).

Covert MW and Palsson BO. Constraints-based models: Regulation of gene expression reduces the steady state solution space. *J. Theor. Biol.*, 221: 309-325 (2003).

David H, Åkesson M and Nielsen J. Reconstruction of the central carbon metabolism of *Aspergillus niger*. *Eur. J. Biochem.*, 270: 4243-4253 (2003).

Edwards JS and Palsson BO. Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *J. Biol. Chem.*, 274: 17410-17416 (1999).

Edwards JS and Palsson BO. The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. *Proc. Natl. Acad Sci. USA*, 97: 5528-5533 (2000).

Edwards JS, Ibarra RU and Palsson BO. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.*, 19: 125-130 (2001).

Förster J, Gombert AK and Nielsen J. Metabolome analysis combined with *in silico* pathway analysis as a tool for functional analysis. *Biotechnol. Bioeng.*, 79: 703-712 (2002).

Förster J, Famili I, Fu P, Palsson BO and Nielsen J. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.*, 13: 244-253 (2003).

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H and Oliver SG. Life with 6000 genes. *Science*, 274: 546-567 (1996).

Gombert AK and Nielsen J. Mathematical modelling of metabolism. *Curr. Opin. Biotechnol.*, 11: 180-186 (2000).

Goodwin, B. C. *Oscillatory Organization in Cells, A Dynamic Theory of Cellular Control Processes*, New York (1963).

Ibarra RU, Edwards JS and Palsson BO. *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature.*, 420: 186-189 (2002).

Ideker T, Galitski T and Hood L. A new approach to decoding life: Systems Biology, *Annu Rev Genomics Hum Genet.*, 2: 343-372 (2001).

Jeong H, Tombor B, Albert R, Oltvai ZN and Barabasi AL. The large-scale organization of metabolic networks. *Nature*. 407: 651-654 (2000).

Kauffman KJ, Prakash P and Edwards JS. Advances in flux balance analysis. *Curr. Opin. Biotechnol.*, 14: 491-496 (2003).

Mahadevan R and Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.*, 5: 264-276 (2003).

Nielsen J. Metabolic Engineering. *Appl. Microbiol. Biotechnol.*, 55: 263-283 (2001).

Nielsen J. It is all about metabolic fluxes. *J. Bacteriol.*, 185: 7031-7035 (2003).

Palsson BO. The challenges of *in silico* biology. *Nat. Biotechnol.*, 18: 1147-1150 (2000).

Papin JA, Price ND, Edwards JS and Palsson BO. The genome-scale metabolic extreme pathway structure in *Haemophilus influenzae* shows significant network redundancy. *J. Theor. Biol.*, 215: 67-82 (2002).

Patil KR, Åkesson M and Nielsen J. Use of genome-scale microbial models for metabolic engineering. *Curr. Opin. Biotechnol.*, 15: 1-6 (2004).

Pramanik J and Keasling JD. Stoichiometric model of *Escherichia coli* metabolism: Incorporation of growth-rate dependent biomass composition and mechanistic energy Requirements. *Biotechnol. Bioeng.*, 56: 398-421 (1997).

Price ND, Papin JA and Palsson BO. Determination of redundancy and systems properties of the metabolic network of *Helicobacter pylori* using genome-scale extreme pathway analysis. *Genome Res.*, 12: 760-769 (2002).

Price ND, Papin JA, Schilling CH and Palsson BO. Genome-scale microbial *in silico* models: The constraints-based approach. *Trends Biotechnol.*, 21: 162-169 (2003).

Rockafellar RT, *Convex Analysis*, Princeton University Press, Princeton, New Jersey (1970).

Salzberg SL. Genomics: Yeast rises again. *Nature.*, 423: 233-234 (2003).

Schilling CH, Edwards JS and Palsson BO. Toward metabolic phenomics: Analysis of genomic data using flux balances. *Biotechnol. Prog.*, 15: 288-295 (1999a).

Schilling CH, Schuster S, Palsson BO and Heinrich R. Metabolic pathway analysis: Basic concepts and scientific applications in the post-genomic era. *Biotechnol. Prog.*, 15: 296-303 (1999b).

Schilling CH, Letscher D and Palsson BO. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.*, 203: 229-248 (2000).

Schilling CH and Palsson BO. Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. *J. Theor. Biol.*, 203: 249-283 (2000).

Schilling CH, Covert MW, Famili I, Church GM, Edwards JS and Palsson BO. Genome-scale metabolic model of *Helicobacter pylori* 26695. *J. Bacteriol.*, 184: 4582-4593 (2002).

Schuster S and Hilgetag C. On elementary flux modes in biochemical reaction systems at steady state, *J Biol Syst.* 2: 165-182 (1994).

Schuster, S., Fell, D. A., and Dandekar, T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.*, 18: 326-332 (2000).

Segre D, Vitkup D and Church GM. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. USA*, 99: 15112-15117 (2002).

Stephanopoulos G, Aristidou AA and Nielsen J. *Metabolic Engineering - Principles and Methodologies*, Academic Press, San Diego (1998).

Stephanopoulos G. Metabolic fluxes and metabolic engineering. *Metab Eng.*, 1: 1-11 (1999).

Szyperski T. ^{13}C-NMR, MS and metabolic flux balancing in biotechnology research. *Q. Rev. Biophys.*, 31: 41-106 (1998).

Tomita M. Whole-cell simulation: A grand challenge of the 21st century. *Trends Biotechnol.*, 19: 205-210 (2001).

Varma A and Palsson BO. Metabolic capabilities of *Escherichia coli*: I. Synthesis of biosynthetic precursors and cofactors. *J. Theor. Biol.*, 165: 477-502 (1993a).

Varma A and Palsson BO. Metabolic capabilities of *Escherichia coli*: II. Optimal growth patterns. *J. Theor. Biol.*, 165: 503-522 (1993b).

Varma A and Palsson BO. Metabolic flux balancing: Basic concepts, scientific and practical use. *Bio/Technology*, 12: 994-998 (1994).

Wiechert W and de Graaf AA. *In vivo* stationary flux analysis by ^{13}C labeling experiments, *Adv Biochem Eng Biotechnol.*, 54: 109-154 (1996).

Wiechert W. ^{13}C metabolic flux analysis, *Metab Eng.*, 3: 195-206 (2001).

Wiechert W. Modeling and simulation: Tools for metabolic engineering. *J. Biotechnol.*, 94: 37-63 (2002).

Chapter 13

DETAILED KINETIC MODELS USING METABOLOMICS DATA SETS

Construction, validation and analysis

Jacky L. Snoep[1,2], Johann M. Rohwer[1]

[1]Triple-J group for Molecular Cell Physiology, Department of Biochemistry, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa [2]Molecular Cell Physiology, Vrije Universiteit, Amsterdam, The Netherlands

1. INTRODUCTION

In the last decade the field of Systems Biology has expanded dramatically. Largely driven by rapid developments in the "-omics" fields, Systems Biology is an integrative approach, trying to combine datasets generated in the different disciplines, with the aim to understand systems as a function of their components. In functional genomics approaches, a lot of progress has been made in describing the cellular components, and predictions can be made on the presence of metabolic pathways on the basis of gene sequence homologies. Such descriptions of metabolic networks have reached a high degree of completeness, and although not all gene products have been mapped, it is only a matter of time before we will get descriptions of all components in a cell. Clearly, such descriptions will be important, providing a map of the networks in the cell, but the information on the interactions between the components is rather limited, i.e., only stoichiometric conversions of metabolites are listed, and no kinetic information can be extracted from the genomics analyses.

Although important information can be obtained from such structural models (see Section 2), such models can only make predictions on flux ratios (not on absolute flux values) and cannot predict metabolite concentrations. More detailed information on interactions between the cellular components is necessary, including kinetic information, to address

these latter issues. In contrast to structural models, which have been made on a genome scale (Schilling *et al.*, 2002; Edwards and Palsson, 2000; Forster *et al.*, 2003), kinetic models have only been made for relatively small systems.

Here, we will address the specific problems associated with building detailed kinetic models (Section 3) and propose a strategy to circumvent these problems (Section 7). We will specifically focus on using metabolomics data sets in building kinetic models, using NMR spectroscopy as a tool for measuring *in vivo* kinetics, and as a tool for model validation. We apply most of the approaches to glycolysis, a well studied system of moderate size. Clearly we cannot be complete in our treatment of modelling approaches. We merely want to indicate a link between the fields of metabolomics and structural and kinetic modelling approaches.

Both structural and kinetic analysis methods are largely dependent on computer models, due to the many interactions or their non-linear character, or both. Ultimately such models would give accurate descriptions of all components in the system and become highly detailed and as a consequence be difficult to work with. Several well-developed systems approaches exist for biochemical networks, e.g. biochemical systems theory (Savageau, 1971; Savageau, 1976) and metabolic control analysis (MCA, see Section 6). We will use the latter framework to illustrate its use both for experimental and modeling approaches, yielding a quantitative higher level description while maintaining the ability to relate systemic properties to characteristics of the individual components. In this chapter we focus on two aspects related to detailed kinetic models: 1) how to build and validate these models, using metabolomics datasets and kinetic information of the system, and 2) how to use these highly detailed models, where we will focus on MCA. First, we will discuss some aspects of structural modeling approaches, focusing on steady-state methods that are also important for kinetic models.

2. STRUCTURAL MODELING OF BIOLOGICAL SYSTEMS

Structural or stoichiometric models focus on the network structure, i.e. what metabolites are linked via what reactions, and largely ignore the nature of these links. The complete information on the network structure can be stored in a so-called stoichiometry matrix. This matrix links the different reactions in a system to the metabolites and is a compact notation form to keep track which reactions consume or produce how much of which metabolites. When starting to model a system, one would typically first write down all the reactions occurring in the system (Table 1). These reaction

stoichiometries can be converted in a stoichiometry matrix (eq. 1), which in addition to being a compact formulation of the network structure, also opens the powerful toolbox of linear algebra. The strength of structural models becomes immediately evident when analyzing so called steady-state conditions. In steady state, all variables (metabolite concentrations) are constant in time, i.e. all reactions must balance and the differential equations are equal to zero. All solutions that obey this steady-state constraint are contained within the null space of the stoichiometry matrix, and can be calculated using standard linear algebra routines. The solution will be in the form of flux ratios, for the reactions listed in Table 1. It can be deduced that v[GLT], v[GLK], v[PGI], v[PFK], v[ALD] and v[TPI] must have the same activity to balance the formation and consumption of CLCi, G6P, F6P, F16P, DHAP, and v[GAPDH], v[PGK], v[PGM], v[ENO], v[PYK], must have twice the activity of v[GLT] to balance the metabolites in the second half of glycolysis.

Table 1. Reaction stoichiometry of glycolysis

Reaction label	Reaction stoichiometry
v[GLT]	GLCo = GLCi
v[GLK]	GLCi + ATP = G6P + ADP
v[PGI]	G6P = F6P
v[PFK]	F6P + ATP = F16P + ADP
v[ALD]	F16P = GAP + DHAP
v[TPI]	DHAP = GAP
v[GAPDH]	GAP + NAD + P = BPG + NADH
v[PGK]	BPG + ADP = P3G + ATP
v[PGM]	P3G = P2G
v[ENO]	P2G = PEP
v[PYK]	PEP + ADP = PYR + ATP
v[PFL]	PYR + CoA = Form + AcCoA
v[PTA]	AcCoA + P = ACP + CoA
v[ACdh]	AcCoA + NADH = Acet + NAD + CoA
v[ACK]	ACP + ADP = Ac + ATP
v[ATP]	ATP = ADP + P
v[ADH]	Acet + NADH = EtOH + NAD
v[LDH]	PYR + NADH = Lac + NAD

[1] GLCo=external glucose, GLCi=internal glucose, G6P=glucose-6-phosphate, F6P=fructose-6-phosphate, F16P=fructose-1,6-bisphosphate, GAP=glyceraldehyde-3-phosphate, DHAP=dihydroxyacetone phosphate, BPG=1,3-bisphophoglycerate, P3G=3-phosphoglycerate, P2G=2-phosphoglycerate, PEP=phospho-enol-pyruvate, PYR=pyruvate, CoA=coenzyme-A, AcCoA=acetyl-coenzyme-A, ACP=acetyl-phosphate, Acet=acetaldehyde, Ac=acetate, EtOH=ethanol, Lac=lactate

For any system to reach a steady state not equal to chemical equilibrium it is necessary for the system to be open, i.e., there must be mass transfer to and from the system. In our small system, GLCo must be added to the

system and Lac, Form, and EtOH must be removed from the system. An experimental set-up where a steady state can be reached is a chemostat, here a continuous supply of medium and removal of effluent leads to a situation where all variables are constant. In batch cultures at best a quasi steady state can be reached where all internal variables are constant while the external metabolites such as GLCo, Lac, Form, and EtOH are changing.

$$\begin{pmatrix} \dot{AcCoA} \\ \dot{Acet} \\ \dot{ACP} \\ \dot{ADP} \\ \dot{ATP} \\ \dot{BPG} \\ \dot{CoA} \\ \dot{DHAP} \\ \dot{F1\acute{6}P} \\ \dot{F\acute{6}P} \\ \dot{G\acute{6}P} \\ \dot{GAP} \\ \dot{GLCi} \\ \dot{NAD} \\ \dot{NADH} \\ \dot{P} \\ \dot{P2G} \\ \dot{P3G} \\ \dot{PEP} \\ \dot{PYR} \end{pmatrix} = \begin{pmatrix} 0&0&0&0&0&0&0&0&0&0&0&1&-1&-1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&-1&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&1&0&-1&0&0&0 \\ 0&1&0&1&0&0&0&-1&0&0&-1&0&0&0&-1&1&0&0 \\ 0&-1&0&-1&0&0&0&1&0&0&1&0&0&0&1&-1&0&0 \\ 0&0&0&0&0&0&1&-1&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&-1&1&1&0&0&0&0 \\ 0&0&0&0&1&-1&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&1&-1&0&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&1&-1&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&1&-1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&1&-1&0&0&0&0&0&0&0&0&0&0&0 \\ 1&-1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&-1&0&0&0&0&0&0&1&0&0&1&1 \\ 0&0&0&0&0&0&1&0&0&0&0&0&0&-1&0&0&-1&-1 \\ 0&0&0&0&0&0&-1&0&0&0&0&0&0&-1&0&0&1&0&0 \\ 0&0&0&0&0&0&0&0&1&-1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&1&-1&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&1&-1&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&1&-1&0&0&0&0&0&-1 \end{pmatrix} \cdot \begin{pmatrix} vGLT \\ vGLK \\ vPGI \\ vPFK \\ vALD \\ vTPI \\ vGAPDH \\ vPGK \\ vPGM \\ vENO \\ vPYK \\ vPFL \\ vPTA \\ vACdh \\ vACK \\ vATP \\ vADH \\ vLDH \end{pmatrix} \tag{1}$$

Balancing the reactions appears trivial for simple systems but rapidly becomes more complex for bigger systems. Even for our small system, with a limited number of branches at the level of pyruvate it might not have been immediately clear that acetate and ethanol are stoichiometrically linked and must be produced in a 1:1 ratio in steady state. In Eq. 2, we show the result of a standard linear algebra routine to express fluxes in a system as a function of independent fluxes, again illustrated for the reaction scheme given in Table 1. Thus, measuring the ethanol and lactate production rates is sufficient to calculate all fluxes in the system.

Typically, external concentrations such as EtOH and Lac can be readily determined experimentally, and by simply measuring the change in time of these metabolites one can calculate all internal fluxes, assuming that the system is in steady state. In addition one could measure the other external fluxes, such as glucose consumption, and acetate and formate production to check on the validity of the system delineation and steady-state assumption. In steady-state chemostat cultures, external concentrations do not change in time but uptake and production rates can easily be calculated from their steady state concentrations, dilution rate and biomass concentration.

$$\begin{pmatrix} \text{vADH} \\ \text{vLDH} \\ \text{vGLT} \\ \text{vGLK} \\ \text{vPGI} \\ \text{vPFK} \\ \text{vALD} \\ \text{vTPI} \\ \text{vGAPDH} \\ \text{vPGK} \\ \text{vPGM} \\ \text{vENO} \\ \text{vPYK} \\ \text{vPFL} \\ \text{vPTA} \\ \text{vACdh} \\ \text{vACK} \\ \text{vATP} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0.5 \\ 1 & 0.5 \\ 1 & 0.5 \\ 1 & 0.5 \\ 1 & 0.5 \\ 1 & 0.5 \\ 2 & 1 \\ 2 & 1 \\ 2 & 1 \\ 2 & 1 \\ 2 & 1 \\ 2 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 3 & 1 \end{pmatrix} \cdot \begin{pmatrix} \text{vADH} \\ \text{vLDH} \end{pmatrix} \tag{2}$$

Again starting from the stoichiometry matrix and applying similar linear algebra methods that were used to find a set of independent fluxes, it is also possible to find dependencies amongst metabolites. Most clearly such dependencies are visible in so-called conserved moieties that have only two members, such as NADH and NAD. Whenever NAD is used, NADH is produced, therefore the one metabolite can be expressed as a function of the other, e.g. NADH = sum - NAD, where sum is the total amount of NADH and NAD. For our reaction network as given in Table 1.1 four conservation relationships exist: NAD+NADH, ATP+ADP, CoA+AcCoA and PEP + ACP + ATP + 2 BPG + DHAP + 2 F16P + F6P + G6P + GAP + P + P2G + P3G. The latter group consists of all phosphorylated intermediates in the system.

For our simple system most of the above has always been part of microbial physiology studies where, typically, carbon and redox balances are determined. Thus, in a homolactic fermentation lactate production must equal twice the glucose consumption for both the carbon and redox (i.e. NADH/NAD) balance to be closed.

The contribution of flux analysis or flux balance analysis [e.g. (Varma and Palsson, 1984; Bonarius *et al.*, 1997)] has been a formalization of these analyses using robust mathematical methods. Ideally one would choose all independent fluxes to be related to external metabolites as these are easy to measure. However, in more complicated systems with parallel routes this is impossible, and then more detailed, intracellular measurements must be made. Robust methods have been developed for such measurements, for instance using isotope labelling techniques in combination with GC-MS [e.g. (Wiechert, 2001; Fischer and Sauer, 2003)]. These methods can distinguish between metabolites labelled at different positions dependent on the

metabolic pathways used. As such they can sometimes be used to study under-determined systems, but they are also very important as an independent validation of the flux balance measurements. Such an independent validation can be of particular importance for flux balance experiments where often large fluxes must be subtracted from one another to calculate a small resultant flux, leading to large errors in the estimation of the latter.

2.1 Elementary mode analysis

In addition to being an important experimental tool structural analysis methods have also been important in more theoretical approaches. The mathematical analysis to find steady state flux relations has also been used to delineate pathways without preconceptions from a given metabolic network. Initial work by Clarke to define "extreme currents" (Clarke, 1981) has more recently been used to define "extreme pathways" (Schilling *et al.*, 2000) and "elementary modes" [e.g. (Schuster *et al.*, 2000)]. A detailed comparison of these methods is beyond the scope of this chapter and the reader is referred to some recent reviews (Schuster, 2004; Papin *et al.*, 2003). Here we only illustrate the concept of elementary modes by applying it to our model system. An elementary mode is defined as a non-decomposable subset of reactions that fulfills the steady-state criteria. For our system, two such elementary modes exist, consisting of the following steps:

- v[GLT] v[GLK] v[PGI] v[PFK] v[ALD] v[TPI] (2 v[GAPDH]) (2 v[PGK]) (2 v[PGM]) (2 v[ENO]) (2 v[PYK]) (2 v[PFL]) v[ACdh] v[ADH] v[PTA] v[ACK] (3 v[ATP])
- v[GLT] v[GLK] v[PGI] v[PFK] v[ALD] v[TPI] (2 v[GAPDH]) (2 v[PGK]) (2 v[PGM]) (2 v[ENO]) (2 v[PYK]) (2 v[LDH]) (2 v[ATP])

The strength of this approach is that these "minimal pathways" can often be given a physiological interpretation. For instance, the two elementary models given above have the following overall reaction stoichiometry: GLCo = 2 Form + EtOH + Ac, and GLCo = 2 Lac, i.e. they are the well known mixed acid and homolactic types of fermentation, observed in many types of bacteria.

Structural models play an important role in biotechnological application to estimate maximal yield values. When a structural model has a sufficient degree of completeness of the system under study, maximal yield values can be calculated (i.e. yields are flux ratios and can be deduced from the stoichiometry matrix).

Although in principle structural models do not contain kinetic information, as is the case in the flux balance methods described above, in

many cases some kinetic or thermodynamic constraints are incorporated. For instance, in the elementary mode analysis reversibility of reactions (a thermodynamic constraint) is used to restrict the number of elementary modes to physiologically feasible ones. Another example where kinetic constraints are also used in addition to thermodynamic constraints is the constraint-based flux balancing method [e.g. (Reed and Palsson, 2003)]. Here maximal flux values are included in calculations of, for instance, maximal growth rates of *Escherichia coli* on a number of substrates. These methods optimize for an objective function within the steady-state solution space, implementing additional kinetic constraints [e.g. (Ibarra *et al.*, 2002)].

3. KINETIC MODELLING OF BIOLOGICAL SYSTEMS

Kinetic models are the ideal tool for the integration of genomics data sets. Quantitative descriptions of mRNA, enzyme and metabolite concentrations by themselves will not lead to understanding of a system. A kinetic model could be an important tool to a) integrate the different data sets, and b) come to an understanding of the system.

Kinetic models come in many different forms. It is not our aim to give an extensive overview of kinetic modelling, excellent review articles and text books exist (Heinrich *et al.*, 1977; Heinrich and Schuster, 1996; Stephanopoulos *et al.*, 1998), and more recently (Wiechert, 2002), including a critical assessment of assumptions and validation techniques. We will restrict ourselves to models using ordinary differential equations (ODEs), and focus on models with a strong mechanistic representation in the biological system. Using such models it is our ultimate aim to come to a quantitative understanding of cellular behavior in terms of experimentally determined enzyme characteristics.

We distinguish between two classes of kinetic models, a) detailed mechanistic models that include realistic enzyme kinetic rate equations and b) phenomenological models that are used to describe data sets but of which the kinetic constants cannot be related to physical entities in the system. Our focus on the first type of models is purely related to our aim to come to an understanding of the system on the basis of the characteristics of the underlying components. We see the importance of the second type of models, especially in studies where a precise description is more important than the ability to relate components of the model to the real system. The phenomenological models usually strive for the simplest model with as few parameters as possible that still describes the essential characteristics of the system. Often these models estimate the kinetic parameters of the model

using fitting routines on system data sets. As explained below we suggest a very different strategy for building mechanistic kinetic models. We will first explain this strategy in general terms in this section, and then make it explicit for a number of case studies where NMR technologies are used in the next section. An extreme form of these types of models aims at making computer replicas of the real system, so-called Silicon Cell models, is treated in Section 7.

Building a mechanistic kinetic model starts in exactly the same way as building a structural model. From a set of reactions as given in Table 1 one builds a set of ODEs as given in Eq. 1. Whereas structural models express the ODE in terms of reaction rates, kinetic models make these reaction rates explicit using kinetic rate equations. Importantly, these rate equations are functions of the metabolite concentrations, time and a set of kinetic parameters. For the structural models we focused on steady-state flux relations that are the solution where all the differential equations are equal to zero. Whereas in structural models the steady-state solution is given as flux relations, in kinetic models the solution is given in metabolite concentrations (and absolute flux values). In addition kinetic models can also simulate dynamic behaviour in time courses.

The rate equation lies at the heart of the kinetic model. In line with our mechanistic approach, the differential equations are built up from reaction steps in the real systems, and as such the rate equations should have a direct relation to the catalyst of that step. Here an immediate advantage of the mechanistic models becomes apparent; in metabolism these catalysts are enzymes and in principle we should be able to isolate these enzymes from the system, characterise them and build our rate equation. Clearly this approach has been followed by enzymologists for over a century, and we should be able to use much of the information that has been accumulated in such isolated enzyme studies. Although such an approach has been very successful in modelling electrical circuits, it has been less so in modelling of biological systems for a number of reasons: first, enzymologists consider the enzyme to be their system and do not necessarily have an interest in the system from which the enzyme was isolated. As a consequence they will study the enzyme under optimal conditions for enzyme activity, and these conditions are not necessarily physiological. In addition, enzymologists often tend to limit themselves to initial rate kinetics in the absence of products. Clearly in networks of enzymes, there is always product present and product sensitivity must be part of the rate equation used in kinetic models. Of a more fundamental nature is the difference between *in vitro* measurements (on the isolated enzyme) and *in vivo* measurements (intracellularly). Whereas the enzyme kinetic mechanism might not depend

on the incubation conditions of the enzyme, the kinetic parameter values may well.

Ideally, enzyme kinetic parameters would be determined under *in vivo* conditions. The number of non-invasive techniques via which one can measure enzyme kinetic parameters *in vivo* is very limited, but see Section 4. Importantly the enzyme, although in its natural environment is characterized as an isolated component. However, very few studies exist in which scientists have measured *in vivo* enzyme kinetics, and the majority of the kinetic models rely on *in vitro* enzyme kinetic measurements. Although it will be hard to simulate intracellular conditions precisely *in vitro*, an attempt should be made to use at least physiological pH, temperature, and ionic strength values. Under these conditions the enzymes should be characterized with respect to sensitivities for substrates, products and effectors and a rate equation (preferably on the basis of a known enzyme kinetic mechanism but otherwise a random order, equilibrium, generic rate equation) can be fitted to the data points to estimate the kinetic parameters. An example of such a generic rate equation for PGK is given in eq. 3:

$$vPGK = \frac{V_{max}\frac{ADP \cdot BPG}{KmADP \cdot K_BPG}(1 - \frac{\Gamma}{K_{eq}})}{(1 + \frac{ADP}{KmADP} + \frac{ATP}{KmATP})(1 + \frac{BPG}{KmBPG} + \frac{P3G}{KmP3G})} \tag{3}$$

The equation serves to illustrate the large number of parameters that will need to be determined (see below and also Section 4). Note that this rate equation, although generic for a two substrate-two product reaction, is derived from a kinetic mechanism, i.e. random order, equilibrium binding. The importance of reversibility and product sensitivity of rate equations should be stressed as they have profound effect on the system behaviour, and in principle all reactions are reversible and product sensitive (albeit that the K_{eq} and binding constant for product might be high). Note that in the rate equation the thermodynamic term, $(1-\Gamma/K_{eq})$ is separated from the kinetic terms and incorporated such that the Haldane relation is always obeyed.

Once the rate equations have been constructed the model can be used in computer simulations. Many different software packages are used, ranging from programming environments such as Fortran, C and Python, via general mathematical programs such as Mathematica (http://www.wolfram.com) and MatLab (http://www.mathworks.com) to dedicated simulation packages designed for biochemical systems simulations such as, for example, Gepasi (Mendes, 1997), Scamp (Sauro, 1991), Jarnac (Sauro, 2000), Copasi (http://www.copasi.org) and PySCeS Olivier et al., 2004). All of these

programs have good numerical integration routines such as LSODA that can handle sets of stiff differential equations. Most simulations can be run on simple desktop computers, although large systems might need stronger computers.

4. DETERMINATION OF KINETIC PARAMETERS *in vivo* AND *in situ*

One of the most important requirements for building silicon-cell type models is the availability of kinetic data for the enzymes in the considered system. Historically, these data have been obtained from enzyme kinetic experiments on purified or partially purified enzyme preparations. This has led to the disadvantage that the different kinetic parameters have often been determined under different conditions of, for example, pH, temperature, buffer and ionic strength for the various enzymes, or for the same enzyme studied in different laboratories. This makes assembly of the kinetic data into an unified model of a pathway difficult, since many kinetic constants are known to change with experimental conditions, as outlined above. A further complication stems from the fact that conditions that are optimal for assaying an enzyme *in vitro* (e.g. pH), may not be physiologically representative.

Discrepancies between model and experiment could in principle result from enzymes behaving differently on a kinetic level *in vivo* than in a dilute solution, normally employed for kinetic assays *in vitro*. Differences in behaviour could result from metabolic channelling (direct transfer of an intermediate from one enzyme to the next without equilibrating with the bulk solution) (Agius and Sherratt, 1997; Srere, 1987; Srivastava and Bernhard, 1987) or from macromolecular crowding effects that alter enzyme kinetic properties (Garner, 1997; Garner and Burg, 1994; Zimmerman and Minton, 1993). In view of this, it would be advantageous to determine enzyme kinetic parameters directly inside the living cell in order to overcome the limitations listed above.

This section thus addresses the determination of kinetic parameters *in vivo* and *in situ*. We shall focus on the use of nuclear magnetic resonance (NMR) spectroscopy as an analytical technique, since it makes online *in vivo* measurement of intracellular concentrations possible, in contrast to classical approaches (e.g. enzyme-linked assays) or high-throughput metabolomic techniques (e.g. LC-MS or GC-MS) that require prior sample extraction.

4.1 *In vivo* enzyme kinetics by NMR

The approach of using NMR spectroscopy for *in vivo* characterisation of cellular metabolism was pioneered by Shulman and co-workers in the late 1970s (den Hollander *et al.*, 1979). They performed NMR spectroscopy on a suspension of *Saccharomyces cerevisiae* cells that had been fed ^{13}C-labelled glucose. The low natural abundance of this isotope of carbon made direct detection of those metabolites possible that were specifically derived from the added glucose and were present in sufficiently high concentrations. In this way, extracellular glucose, glycerol and ethanol, as well as intracellular fructose bisphosphate could be quantified over a time-course of 15 minutes. The authors used the data to fit glucose uptake kinetics to a Michaelis-Menten model and to obtain information about the aldolase-triose phosphate isomerase triangle.

Subsequently, the field of *in vivo* NMR has expanded significantly, and there are many reports in the literature about the use of this technique for direct on-line determination of the intracellular concentrations of metabolites. The application to kinetic characterisation of enzymes has, however, been more limited. We will restrict ourselves to the determination of kinetic parameters that are relevant to silicon-cell type models (such as enzyme K_M values). Of course, NMR spectroscopy has also been used extensively in the study of enzyme mechanisms and the determination of mechanistic rate constants—those details fall beyond the scope of this chapter.

A classical example of the application of this approach is provided by the work of Mulquiney in the group of Kuchel (Mulquiney *et al.*, 1999), who kinetically characterised the enzyme 2,3-bisphosphoglycerate (2,3-BPG) synthase/phosphatase in human erythrocytes *in vivo*. The work was part of a larger project to build a detailed kinetic model of erythrocyte metabolism (Mulquiney and Kuchel, 1999b). The authors incubated erythrocyte suspensions with ^{13}C-labelled glucose and measured glucose, lactate and 2,3-BPG with ^{13}C-NMR, as well as 2,3-BPG, inorganic phosphate and ATP with ^{31}P-NMR. Furthermore, starvation of erythrocytes in the absence of glucose was used to monitor enzyme activity at low 2,3-BPG levels. Through a process of iterative adjustment, the enzyme kinetic parameters were fitted to the experimentally obtained metabolic time courses. The kinetic characetrisation included details on the pH-dependence of the enzyme and moreover found significant differences between kinetic parameters *in vivo* and those that had earlier been obtained *in vitro* (Mulquiney *et al.*, 1999):

- *in vivo*, 3-phosphoglycerate and 2-phosphoglycerate are much weaker inhibitors of the phosphatase reaction than *in vitro*;

- the K_M for 2,3-BPG *in vivo* is significantly higher than measured *in vitro*;
- the V_{max} for the phosphatase *in vivo* is about twice that measured *in vitro*; and
- 2-phosphoglycollate does not play a role in the activation of the phosphatase *in vivo*.

NMR spectroscopy has also been used to obtain parameters for constructing models using phenomenological kinetic equations. While such an approach can yield a kinetic model that describes a system more or less satisfactorily, it does, however, have the disadvantage that the equations used to describe the enzyme reactions are not based on mechanism and may consequently break down under conditions different from those for which the parameters were determined. Nevertheless, two examples will be mentioned briefly here. In the first example, Santos and co-workers (Neves *et al.*, 1999) studied glycolytic kinetics in *Lactococcus lactis* using *in vivo* ^{13}C and ^{31}P NMR spectroscopy in a circulating system. Combined with NMR analysis of phosphorylated metabolites in extracts, this allowed the authors to construct a kinetic model using a general equation for each of the steps considered. This model could predict qualitatively the shifts from anaerobic to aerobic glucose metabolism.

In a second, recent example (Martini *et al.*, 2004), ^{13}C NMR was used to construct a mathematical model that specifically describes the response of the yeast *S. cerevisiae* to ethanol stress. An aggregated approach was followed by merely describing glucose degradation and ethanol production as overall rates depending on the concentrations of those compounds and on the number of active yeast cells in the system. While the model could reproduce experimental time courses of glucose concentrations during different fermentation conditions, it was not based on the functional characteristics of the underlying enzyme activities and is hence limited in its applicability towards answering more general or extended questions.

This section has given some examples of how NMR spectroscopy can be used to determine kinetic parameters of enzyme reactions or aggregates of these reactions *in vivo*, and how these parameters can then be combined in kinetic models. The next section will address how these kinetic models can be validated to ascertain whether they indeed describe experimental reality adequately.

5. MODEL VALIDATION

A good modelling practice is to perform a sensitivity analysis and model validation after the model construction. In a sensitivity analysis all model parameters are perturbed and the effect on model behavior, for instance the steady state result, determined. Such an analysis is important to pin-point influential parameters, i.e. those parameters that have a large effect on the model outcome if they are changed. In an iterative approach one could go back to the *in vitro* determination of those parameters and maybe perform more extensive experiments. There are several ways of model validation; it is crucial that the validation should be performed independently from the model construction, i.e., the data sets used for model construction should be independent from the data set used for validation. In our modelling approach validation is of critical importance and we adhere strictly to using a completely different data set for model construction and validation. The model is constructed using *in vitro* data and validated with *in vivo* data, i.e., using experimental data determined on the whole system. For a model focussing on steady state behavior such a validation set would be steady-state metabolite concentrations and fluxes. On comparison of the model predictions and the experimental data it is important to view possible differences in line with the sensitivity analysis, i.e., whether the validation data set can be described perfectly within a 5 % experimental error of the kinetic parameter set. If so, there is no reason for rejection of the model and if a higher accuracy of the model is important, the kinetic parameters can be determined more accurately in a new set of experiments on the isolated enzyme. If the model cannot describe the validation data set within a 5 % experimental error, the different reaction steps must be re-evaluated. Importantly this can be done for each step in isolation. From the validation data set for each of the reaction step the steady state substrate, product and effector concentrations and the flux through the reaction can be extracted. Subsequently for each of the rate equations in the model it can be verified whether upon insertion of the measured steady state metabolite concentrations, the rate of the enzyme is equal to the steady state flux through the enzyme. Again, for differences between the measured and predicted values it should be checked whether they are within the experimental error of the parameter estimation, if not, then the rate equation of that specific enzyme should be tested with more extensive enzyme kinetic measurements.

5.1 Examples

A number of kinetic models have been built using the mechanistic approach as outlined above. We discuss two of these models that were specifically constructed to test whether one can make reliable kinetic models on the basis of *in vitro* enzyme kinetic data (Bakker *et al.*, 1997; Teusink *et al.*, 2000). For the first system, glycolysis in the bloodstream form of the parasite *Trypanosoma brucei*, the authors had the availability of a precise set of kinetic data determined in one laboratory. Thus, the *in vitro* data formed a consistent set, determined under comparable conditions and measured in the same strain. The model was constructed using this *in vitro* kinetic data set, and no additional parameters were fitted. Model validation was performed on isolated intact parasites and on parasites living in the bloodstream of rats. The model described the steady state *in vivo* system quite accurately, all metabolite concentrations were within one order of magnitude of the experimentally determined values and the fluxes were described reasonably accurately. Additional validations, such as descriptions of both the aerobic and anaerobic state, and testing of model predictions on control of the transport step were also successful (Bakker *et al.*, 1997). The model has subsequently been refined (Helfert *et al.*, 2001) and used for the development of a rational strategy for drug design (Bakker *et al.*, 2000; Eisenthal and Cornish-Bowden, 1998).

The construction of a detailed kinetic model for the best studied metabolic pathway, glycolysis in the yeast *S. cerevisiae*, might serve as a second example (Teusink *et al.*, 2000). Although much more research had been done on the isolated glycolytic enzymes of yeast than for *T. brucei*, no consistent set of kinetic data was available for *S. cerevisiae*. Measurements were performed in separate laboratories, each using their own conditions, and enzymes were isolated from several yeast strains grown under different conditions. For instance, for the enzyme phosphofructokinase (PFK), one of the model enzymes to study allosteric regulation, no consistent data set was available for the effects of the different effectors, let alone a rate equation describing these effects. As a consequence the authors were forced to measure kinetic parameters for many of the different enzymes, at least the V_{max} values, but often also binding constants, and combine these with kinetic data from the literature. For PFK many more kinetic experiments were performed and these data points were combined with literature data to come to a consistent data set to which a single rate equation was fitted.

The rate equations built on the basis of *in vitro* enzyme kinetic data were combined in a kinetic model for anaerobic yeast glycolysis. No fitting of kinetic parameters that could be determined *in vitro* was done on the complete system. However, for the ATP hydrolysis reaction, which

describes the kinetics of a large number of reactions and which was not measured *in vitro*, the kinetic parameters were fitted on measurements on the complete system. The authors restricted themselves to include detailed kinetic information on the core enzymes of glycolysis, i.e., the steps from glucose to ethanol. However, when validating their model in the complete system, it became apparent that trehalose, glycogen, glycerol and succinate were also formed. Subsequently these branches were included in the model, but no detailed kinetic information was available and these branch activities were inserted as fixed flux values equal to the experimentally determined values. As such the model became limited to describing a single steady state. Even for this limited scope the model description of the *in vivo* steady state behavior was not perfect. Whereas the glycolytic flux was accurately described, some of the metabolite concentrations predicted were off by more than a factor of 5, although all were within the same order of magnitude.

In addition to being the best studied metabolic system, yeast glycolysis is probably also the system for which the most kinetic models have been made. These models range from core models, used to illustrate a principle such as glycolytic oscillations (e.g. (Goldbeter and Lefever, 1972; Wolf *et al.*, 2000)), to models using more of an engineering approach (Galazzo and Bailey, 1990), to detailed kinetic models (e.g. (Teusink *et al.*, 2000; Rizzi *et al.*, 1997; Hynne *et al.*, 2001)). We here just illustrated our modelling approach using the model developed by (Teusink *et al.*, 2000). It should be realised that each model is developed to answer a certain question, and therefore many different models can be made for the same system, the model described in Teusink *et al.* (2000) is closest to our approach and we therefore chose this model.

5.2 Validation by NMR spectroscopy

As outlined above, the basis of model validation is the comparison to independent experimental data on fluxes and metabolite concentrations. Since fluxes can be calculated from changes in concentrations over time, an accurate determination of metabolite concentrations is crucial for kinetic model validation. NMR spectroscopy is well suited to this type of measurement, since it allows for the direct on-line quantification of metabolite concentrations *in vivo*. Because there is no need for sampling, extraction and separate analysis, metabolite data for whole time-courses can be acquired in a single experimental run. An example of such a time course is given in Figure 1.

It must be pointed out that NMR spectroscopy is not a "high-throughput" technique in the sense of providing data on all or most of the metabolites in a living cell (i.e. the metabolome). The power of techniques such as LC-MS or

GC-MS in providing such data is summarized elsewhere in this volume (Chapter 7). A further restriction of NMR is that it is generally limited to quantifying those metabolites *in vivo* that are relatively abundant. However, distinct advantages of NMR spectroscopy lie in its direct and non-invasive measurement and in its time resolution because of the online determination. Especially when used in combination with specific isotopic labelling of substrates (e.g. ^{13}C), NMR can provide powerful answers to problems of model validation. Ultimately the approach chosen for a particular problem will depend on the specific questions addressed.

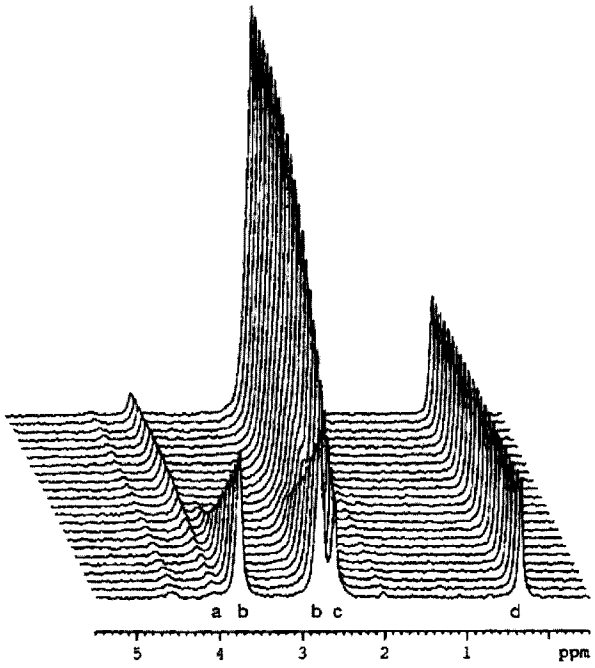


Figure 1. Typical *in vivo* ^{31}P-NMR time-course that can be used for model validation. Human erythrocytes were suspended in buffer at 37 °C in the absence of glucose. The time interval between successive spectra is 30 min. Peak assignments: a, AMP; b, 2,3-bisphosphoglycerate; c, inorganic phosphate; d, triethylphosphate (internal standard).

In the following, we shall briefly describe three case studies where NMR has been used successfully in the validation of kinetic models. The first is the kinetic model of erythrocyte metabolism (Mulquiney and Kuchel, 1999b) already mentioned above (see Figure. 2). In that specific case, NMR was

used both in the determination of kinetic parameters for model construction and in the validation of the model. In fact, the authors followed an iterative approach of model validation and parameter refinement. Typical NMR data

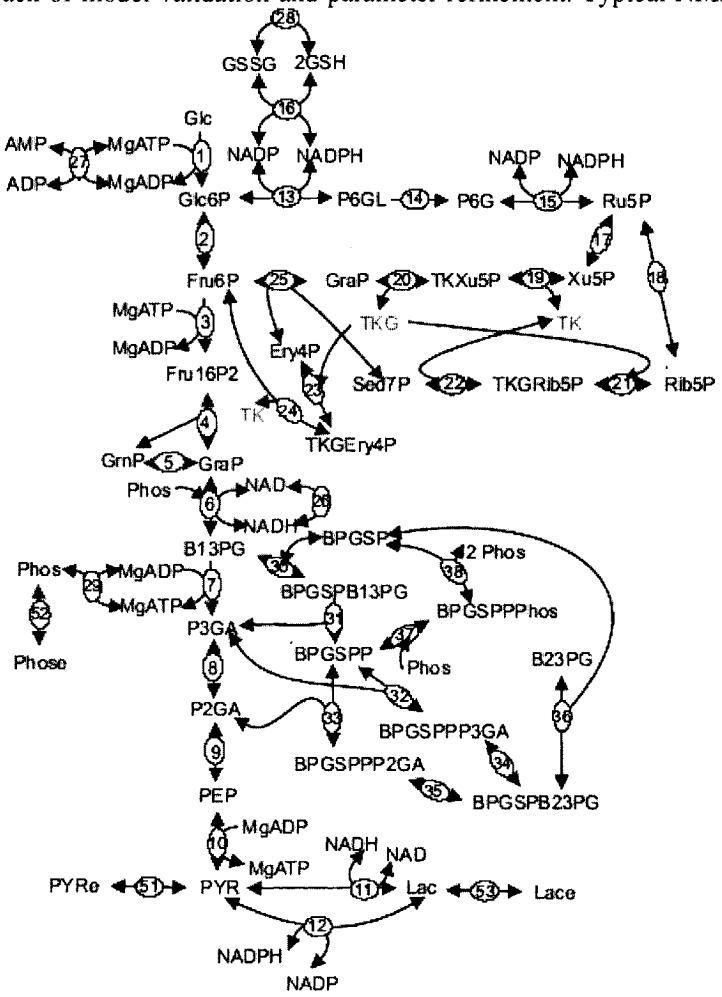


Figure 2. Reaction scheme of the erythrocyte kinetic model in (Mulquiney and Kuchel, 1999b). The abbreviation BPGSP refers to the enzyme 2,3-bisphosphoglycerate (2,3-BPG) synthase/phosphatase that was characterised by the authors using in vivo NMR kinetics (see above). The other abbreviations are defined in the original paper.

that were used for this included time courses of 2,3-BPG or glucose concentrations, or of pH (Mulquiney *et al.*, 1999). These were combined with metabolite data from the literature obtained from other means such as

enzymatic assays (Mulquiney and Kuchel, 1999b). One specific conclusion that arose from this work was that the kinetic parameters for the enzyme phosphofructokinase in the model were uncertain, since *in vitro* data were only available for the rat enzyme and not the human one. The NMR analysis revealed significant kinetic differences between the human enzyme and the published rat data. The study could not resolve whether the differences were the result of interspecies differences or of different mechanisms in situ (Mulquiney and Kuchel, 1999b). The approach is general, and in earlier work, Kuchel's group (Thorburn and Kuchel, 1985) had already employed *in vivo* proton NMR determinations of glutathione in erythrocytes to validate a kinetic model of the hexose-monophosphate shunt.

Glycolysis, being the "oldest" metabolic pathway, has been studied in many organisms and tissue types. It stands to reason that the glycolytic enzymes belong to those that have been characterised in the greatest detail. This in turn has facilitated the development of numerous glycolytic kinetic models. In addition to the erythrocyte model described above, a model for mammalian skeletal muscle (Lambeth and Kushmerick, 2002) is available. The approach followed by Lambeth and coworkers was to collect kinetic information on the enzymes from the literature and assemble this into a model. This model was subsequently validated in independent experiments by performing *in vivo* ^{31}P NMR spectroscopy on the hind limb muscles of mice. The experimental time courses of phosphocreatine and inorganic phosphate thus obtained matched the simulations and thus validated the model predictions (Lambeth *et al.*, 2002).

The fact that the ^{31}P isotope has a natural abundance of 100%, has led to numerous NMR applications measuring phosphorylated intermediates, which has greatly facilitated the quantification of metabolic intermediates of cellular free-energy metabolism. In one such a study, ^{31}PNMR was used to determine phosphocreatine kinetics in skeletal muscle of creatine kinase knockout mice and to compare these to the wild type (Roman *et al.*, 2002). The authors measured the *in vivo* concentrations of phosphocreatine and ATP in the gastrocnemius muscles of mice upon twitch stimulation and used the data to validate a kinetic model of ATP production by oxidative phosphorylation, muscle ATPase, adenylate kinase and creatine kinase.

This section has dealt with the validation of kinetic models. In the next section, we introduce another tool for analysing the behaviour, control and regulation of cellular systems, viz. the high-level formalism of metabolic control analysis.

6. METABOLIC CONTROL ANALYSIS

Metabolic control analysis [MCA, reviewed in (Fell, 1992; Liao and Delgado, 1993; Fell, 1996; Heinrich and Schuster, 1996)] is a quantitative framework for analysing the steady-state behaviour of metabolic pathways. It quantifies the dependence of the system variables (typically, fluxes and concentrations) on the parameters (e.g., enzyme concentrations, external effector concentrations, or kinetic properties of enzymes), which determine the steady state that prevails in the system. The contributions of MCA are twofold: first, it allows for the quantification of flux or concentration control in terms of a coefficient that can assume a precise value, leading to a more differentiated description than the all-or-nothing notion of a "rate-limiting-step"; and second, it allows the systemic properties to be understood in terms of the local properties of the components (i.e., enzymes) through a set of mathematical laws/relations.

6.1 General definitions

Two types of coefficients defined in MCA are control coefficients, which describe the dependence of steady-state variables on changes in the activities of the system reactions, and elasticity coefficients, which characterize the local dependence of the reaction rates on their substrates, products or effectors in isolation. Elasticity coefficients (or simply, elasticities) are thus local properties, which can be determined from studying individual reactions in isolation (e.g., by enzyme-kinetic analysis). Control coefficients, on the other hand, are global properties, which characterise the steady-state behaviour of the entire system.

An elasticity coefficient quantifies the effect of any molecular species or parameter that affects a step directly on the local rate through that step in isolation, and is defined mathematically as follows:

$$\varepsilon_s^v = \frac{\partial \ln v}{\partial \ln s} \tag{4}$$

where v is the local rate of any step (e.g., enzyme-catalysed reaction) inthe system and s the concentration of any molecular species or parameter that affects the unit step directly.

Control coefficients are defined mathematically as follows [see (Heinrich *et al.*, 1977; Schuster and Heinrich, 1992; Kacser and Burns, 1973)]:

$$C_{v_i}^{y} = \frac{(\partial \ln y / \partial \ln p)_{ss}}{(\partial \ln v_i / \partial \ln p)_{\text{step } i}} \tag{5}$$

where y is a steady-state system variable (flux or metabolite concentration), and p is any parameter that acts specifically only on step i. The subscript ss indicates that the entire system relaxes to a new steady state after a change in p, and subscript step i indicates that only the change in local rate v_i of step i is considered at constant reactant, product and effector concentrations. Importantly, this definition is parameter independent and can be conceptualised as the steady-state response in y to a change in the local activity of step i. If y refers to a flux, the coefficient is termed a flux-control coefficient (denoted C^{J}_{vi}), whereas if y refers to a metabolite concentration, the coefficient is termed a concentration-control coefficient (denoted C^{S}_{vi}).

The most powerful result of metabolic control analysis is that it relates the control and elasticity coefficients through a series of summation and connectivity relationships (Kacser and Burns, 1973; Heinrich and Rapoport, 1974; Westerhoff and Chen, 1984). These relations can be used to express the control coefficients in terms of the elasticities. In theory, it should therefore be possible to describe the control of the steady-state behaviour of an entire system of coupled enzyme-catalysed reactions from knowing the behaviour of every individual enzyme in isolation (i.e., its dependence on product, substrate and effector concentrations).

The remainder of this section will deal with the determination of control and elasticity coefficients in experimental systems, focusing first on their direct experimental determination, and then on their calculation using kinetic models.

6.2 Direct experimental determination of control coefficients

To determine control coefficients in a biochemical system, different experimental approaches can be followed (see (Fell, 1992; Liao and Delgado, 1993; Fell, 1996; Heinrich and Schuster, 1996) for review). The intracellular concentration of an enzyme can be modulated directly by genetic means to change the enzyme activity, for example by cloning the gene that codes for the enzyme behind an inducible promoter. Alternatively, enzymes may be titrated with specific inhibitors, and the control coefficients may be estimated from the definition in Eq. 5 (the inhibitor concentration then equates to the parameter p). In a third approach, control coefficients

may be calculated from elasticities using the summation and connectivity relationships described above.

Here, we shall briefly discuss two examples where NMR spectroscopy has been used in the context of *in vivo* metabolite concentration measurements to determine control coefficients in a cellular system. In the first example, Jeneson and co-workers used *in vivo* ^{31}P-NMR spectroscopy to determine the control coefficents in ATP free-energy metabolism in contracting skeletal muscle (Jeneson *et al.*, 2000). They divided the network into three steps that form a branch around the central intermediate ATP: an ATP-producing branch (mitochondria), and two ATP-consuming branches (actinomyosin in the muscle filaments and sarcoplasmic calcium ATPase). By measuring the steady-state concentrations of cytosolic ATP and ADP, as well as the ATP metabolic flux, with NMR, the authors were able to calculate the elasticities of each of the blocks for ATP from the respective kinetic functions. These elasticities were then used to calculate the control coefficients using the formalism of MCA. The steady state was varied by increasing the ATP load on the system through increased muscle contraction frequencies. This provided an elegant demonstration of how the control on ATP flux shifted from the demand (actinomyosin) to the supply (mitochondria) as the contraction frequency was increased, and proved that the muscle ATPase system is inherently homeostatically buffered.

A second example that deserves mention is work from Shulman's group (Chase *et al.*, 2001) measuring flux control in the glycogen synthesis pathway of rat gastrocnemius muscle by *in vivo* NMR spectroscopy. The approach was similar to that of Jeneson *et al.* (2000). The system was divided into three blocks around glucose-6-phosphate: glucose transport and phosphorylation, glycogen synthesis, and glycolysis. The concentration of glucose-6-phosphate and the glycogen synthesis flux were measured by ^{13}C and ^{31}P *in vivo* NMR spectroscopy. Different metabolic states were achieved by insulin stimulation. From the data, the *in vivo* elasticities of each of the blocks for the intermediate glucose-6-phosphate were calculated using the MCA theorems. The authors concluded that most of the flux control on glycogen synthesis lies in the glucose transport and hexokinase (Chase *et al.*, 2001).

6.3 Control analysis using kinetic models

If the enzymes of the investigated system are characterised in terms of their kinetic rate laws and a kinetic model of the system is available, control coefficients may be estimated directly from this model. This approach has the great advantage that, once a model is available, it is trivially easy to manipulate the activity of each of the steps and calculate the control

coefficients and elasticities *in silico*. This approach also makes it possible to determine the control coefficients of those steps that may not be amenable to direct experimental manipulation (because no inhibitor is available, or enzyme levels cannot be varied by genetic means, etc.). Moreover, it is less time consuming than an experimental control analysis.

Of course, this approach comes with a caveat: because the control analysis reflects the model reality, extrapolations to experimental reality can only be made if the model has been properly validated and is thus deemed to be a realistic representation of the experimental system (see also Section 5).

This approach has been applied extensively, e.g. to erythrocyte metabolism in the groups of Heinrich (1990) and Kuchel (Mulquiney and Kuchel, 1999a); the latter model has already been described above, and the *in silico* control analysis yielded several new insights into the regulation of 2,3-BPG metabolism. Specifically, the model allowed changes in the flux through the 2,3-BPG shunt and the steady-state 2,3-BPG concentrations upon different energy demands to be calculated (this would have been very hard if not impossible to do experimentally) and the routes of control and regulation identified and quantified (Mulquiney and Kuchel, 1999a). The glycolytic model of skeletal muscle described above (Lambeth and Kushmerick, 2002) has also been subject to MCA, leading to the important conclusion that under most conditions, the majority of the control on the glycolytic flux resided in the demand for ATP, i.e. "outside" of the glycolytic pathway itself [see also (Hofmeyr and Cornish-Bowden, 2000)].

In another example, Rohwer and Botha (2001) have applied this approach to a kinetic model of sucrose metabolism and accumulation in the sugarcane plant. They constructed and validated a model describing the uptake of glucose and fructose by sugarcane culm cells, the metabolism and conversion to sucrose, and the utilisation of carbon in glycolysis. One of the central results of this work was that the model reproduced experimental data showing a "futile cycle" of continuous sucrose synthesis and degradation. It was hypothesised that this futile cycling would lead to a decrease in the amount of sucrose accumulated, which would have an impact on, amongst others, agricultural yields of sucrose. Importantly, the model allowed the control of this futile cycle to be quantified in terms of MCA, leading to suggestions for experimental manipulations that could possibly reduce futile cycling of sucrose and increase sucrose accumulation (Rohwer and Botha, 2001). In terms of biotechnological manipulation, an experimental strategy directed by such simulation results is more promising than a mere random, trial-and-error approach.

7. THE SILICON CELL: LINKING THE MODULES

The Silicon Cell project (http://www.siliconcell.net) has the same approach to building kinetic models as we have advocated here, using experimentally determined rate laws and parameter values to calculate whether the systems behavior can be described on the basis of our knowledge of the individual components and their interactions. The ultimate aim of the project is to come to realistic models on a cellular level. The approach as we have described in this chapter is illustrated for relatively small systems and we here would like to discuss how such models on parts of the system can be combined. Clearly, any attempt to make a detailed model of a system containing several thousands of reaction steps is doomed to fail. Instead, we propose a modular approach, and although we do not know whether such an approach will be successful, we can increase the chances of success dramatically by at least trying.

Detailed kinetic models, certainly models of the mechanistic type, contain a large number of kinetic parameters and each parameter value will contain an experimental error. To prevent a potential accumulation of these experimental errors we suggest the building of detailed kinetic models of subsets of the system. Such smaller models should be validated separately and, when "approved", stored in a database. Models of interacting modules can be combined and again validated. In such a way a gradually growing silicon cell can be constructed from validated smaller models.

An advantage of our mechanistic approach to modelling and the restriction to include only parameter values that are determined on the "isolated" components, is that the parameter values and the individual rate equations can be used in extended versions of that specific model system. Of course the kinetic parameter values are limited to the specific conditions for which they have been determined, notably V_{max} values are very dependent on the specific environmental conditions. However, when these same conditions are applied, models on subsystems can be combined without having to adjust the enzyme kinetic parameters. This is an immediate consequence of the mechanistic approach, the kinetic parameters have a physical interpretation, e.g. a binding constant for a substrate, and this constant does not depend on the way in which we define our model. This is an important difference with the phenomenological models where the kinetic parameters of a subset of the system are fitted on a dataset obtained on the complete system. Whenever a reaction is added to such a model all parameter values will have to be fitted again; this is not a very constructive exercise in a modular approach to build a complete model of a cellular system.

For such a modular approach to be successful in building detailed kinetic models on a cellular level, a coordinated group effort will be essential.

Building kinetic models even of moderate size is very time consuming, and a typical cell consisting of several thousands of enzymes cannot be modelled in a single group. For groups to collaborate in such a modelling effort, standardisation of experiments (e.g. protocols for cultivating the organisms, *in vitro* enzyme kinetic measurements, for model validation) and of model construction (e.g. model format) will be important to be able to combine the different models on parts of metabolism. A first task will be to divide the cell in manageable parts for which detailed kinetic models can be constructed and validated. This is not a trivial exercise and we would like to stress the importance of model validation in the definition of such modules; it should be possible to validate each module on its own.

Each module that has been validated can be stored in a database of models; such a database can be made public (for instance if the model has been published) or access can be restricted to the members of the collaborating team. It is important that the models should be stored in a format that can be used in many different software packages (for instance SBML - http://www.sbml.org) but also in a format that allows easy grouping of modules. In addition, the models should be accessible to all members of the research team, also if they are non-expert modelers. We have developed a model database that is accessible via the internet, via a very friendly user-interface [http://jjj.biochem.sun.ac.za and mirror sites in Europe http://jjj.bio.vu.nl, and the United States http://jjj.vbi.vt.edu, (Olivier and Snoep, 2004)]. Note that all models discussed in this chapter are listed in the database. The models can be interrogated in your browser, using a server-client set-up, employing Java applets on the user site and web Mathematica (http://www.wolfram.com) on the server site. This set-up is used by the yeast systems biology network, a group of yeast researchers, experimentalists, and modelers that use systems biology tools to understand rules and principles of the dynamic operation of cellular systems, with yeast as a model system (http://www.gmm.gu.se/YSBN/). Another initiative, using *E. coli* as a model system is the international *E. coli* alliance (IECA) (http://www.uni-giessen.de/~gx1052/IECA/ieca.html).

REFERENCES

Agius L and Sherratt HSA. (eds.) *Channelling in Intermediary Metabolism*, London. Portland Press (1997).

Bakker BM, Michels PAM, Opperdoes FR and Westerhoff HV. Glycolysis in bloodstream form *Trypanosoma brucei* can be understood in terms of the kinetics of the glycolytic enzymes. *J. Biol. Chem.*, 272: 3207–3215 (1997).

Bakker BM, Westerhoff HV, Opperdoes FR and Michels PA. Metabolic control analysis of glycolysis in trypanosomes as an approach to improve selectivity and effectiveness of drugs. *Mol Biochem Parasitol.*, 106: 1–10 (2000).

Bonarius HPJ, Schmid G and Tramper J. Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. *Trends biotechnol.*, 15: 308–314 (1997).

Chase JR, Rothman DL and Shulman RG. Flux control in the rat gastrocnemius glycogen synthesis pathway by in vivo $^{13}C/^{31}P$ NMR spectroscopy. *Am. J. Physiol. Endocrinol. Metab.*, 280: E598–607 (2001).

Clarke BL. Complete set of steady states for the general stoichiometric dynamical system. *J. Chem. Phys.*, 75: 4970–4979 (1981).

den Hollander JA, Brown TR, Ugurbil K and Shulman RG. ^{13}C Nuclear Magnetic Resonance studies of anaerobic glycolysis in suspensions of yeast cells. *Proc. Natl. Acad. Sci. USA*, 76: 6096–6100 (1979).

Edwards JS and Palsson BO. The *Escherichia coli* MG1655 *in silico* metabolic genotype: Its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. USA*, 97: 5528–5533 (2000).

Eisenthal R and Cornish-Bowden A. Prospects for antiparasitic drugs. the case of *Trypanosoma brucei*, the causative agent of African sleeping sickness. *J. Biol. Chem.*, 273: 5500–5505 (1998).

Fell DA. Metabolic control analysis: a survey of its theoretical and experimental development. *Biochem. J.*, 286: 313–330 (1992).

Fell DA. *Understanding the Control of Metabolism*. Portland Press, London (1996).

Fischer E and Sauer U. Metabolic flux profiling of *Escherichia coli* mutants in central carbon metabolism using GC-MS. *Eur. J. Biochem.*, 270: 880–891 (2003).

Forster J, Famili I, Fu P, Palsson BO and Nielsen J. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res*, 13: 244–253 (2003).

Galazzo JL and Bailey JE. Fermentation pathway kinetics and metaboic flux control in suspended and immobilized *Saccharomyces cerevisiae*. *Enzyme Microb. Technol.*, 12: 162–172 (1990).

Garner MM. The consequences of macromolecular crowding for metabolic channelling. In Agius, L. and Sherratt, H. S. A., (eds), *Channelling in Intermediary Metabolism*, pp 41–52, London. Portland Press (1997).

Garner MM and Burg MB. Macromolecular crowding and confinement in cells exposed to hypertonicity. *Am. J. Physiol.*, 266: C877–C892 (1994).

Goldbeter A and Lefever R. Dissipative structures for an allosteric model. *Biophys J*, 12: 1302–1315 (1972).

Heinrich R, Rapoport SM and Rapoport TA. Metabolic regulation and mathematical models. *Progr. Biophys. Molec. Biol.*, 32: 1–82 (1977).

Heinrich R and Rapoport TA. A linear steady-state treatment of enzymatic chains. general properties, control and effector strength. *Eur. J. Biochem.*, 42: 89–95 (1974).

Heinrich R. Metabolic control analysis: principles and application to the erythrocyte. In Cornish-Bowden, Athel and C´ardenas, M. Luz, (eds.), *Control of Metabolic Processes*, pp 329–342, New York. Plenum Press (1990).

Heinrich R and Schuster S. *The Regulation of Cellular Systems*. Chapman and Hall, New York (1996).

Helfert S, Estevez AM, Bakker B, Michels P and Clayton C. Roles of triosephosphate isomerase and aerobic metabolism in *Trypanosoma brucei*. *Biochem. J.*, 357: 117–125 (2001).

Hofmeyr J-HS and Cornish-Bowden A. Regulating the cellular economy of supply and demand. *FEBS Lett.*, 476: 47–51 (2000).

Hynne F, Dano S and Sorensen PG. Full-scale model of glycolysis in *Saccharomyces cerevisiae*. *Biophys. Chem.*, 94: 121–163 (2001).

Ibarra RU, Edwards JS and Palsson BO. *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature*, 420: 186–189 (2002).

Jeneson JA, Westerhoff HV and Kushmerick MJ. A metabolic control analysis of kinetic controls in ATP free energy metabolism in contracting skeletal muscle. *Am. J. Physiol. Cell. Physiol.*, 279: C813–832 (2000).

Kacser H and Burns JA. The control of flux. *Symp. Soc. Exp. Biol.*, 27: 65–104 (1973).

Lambeth MJ and Kushmerick MJ. A computational model for glycogenolysis in skeletal muscle. *Ann. Biomed. Eng.*, 30: 808–827 (2002).

Lambeth MJ, Kushmerick MJ, Marcinek DJ and Conley KE. Basal glycogenolysis in mouse skeletal muscle: *in vitro* model predicts *in vivo* fluxes. *Mol. Biol. Rep.*, 29: 135–139 (2002).

Liao JC and Delgado J. Advances in metabolic control analysis. *Biotechnol. Prog.*, 9: 221–233 (1993).

Martini S, Ricci M, Bonechi C, Trabalzini L, Santucci A and Rossi C. *In vivo* ^{13}C-NMR and modelling study of metabolic yield response to ethanol stress in a wild-type strain of *Saccharomyces cerevisiae*. *FEBS Lett.*, 564: 63–68 (2004).

Mendes P. Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.*, 22: 361-363 (1997)

Mulquiney PJ and Kuchel PW. Using the β/α peak-height ratio of ATP in ^{31}P NMR spectra to measure free [Mg^{2+}]: theoretical and practical problems. *NMR In Biomedicine*, 12: 217–220 (1999a).

Mulquiney PJ, Bubb WA and Kuchel PW. Model of 2,3-bisphosphoglycerate metabolism in the human erythrocyte based on detailed enzyme kinetic equations: *in vivo* kinetic characterization of 2,3-bisphosphoglycerate synthase/phosphatase using ^{13}C and ^{31}P NMR. *Biochem. J.*, 342: 567–580 (1999).

Mulquiney PJ and Kuchel PW. Model of 2,3-bisphosphoglycerate metabolism in the human erythrocyte based on detailed enzyme kinetic equations: equations and parameter refinement. *Biochem. J.*, 342: 581–596 (1999b).

Neves AR, Ramos A, Nunes MC, Kleerebezem M, Hugenholtz J, de Vos WM, Almeida J and Santos H. *In vivo* nuclear magnetic resonance studies of glycolytic kinetics in *Lactococcus lactis*. *Biotechnol Bioeng*, 64: 200–212 (1999).

Olivier BG, Rohwer JM, Hofmeyr JHS. Modelling cellular systems with PySCeS. *Bioinformatics* Advanced Access Publication September 2004 bti046.

Olivier BG and Snoep JL. Web-based kinetic modelling using JWS Online. *Bioinformatics*, 20: 2143–2144 (2004).

Papin JA, Price ND, Wiback SJ, Fell DA and Palsson BO. Metabolic pathways in the postgenome era. *Trends Biochem. Sci.*, 28: 250–258 (2003).

Reed JL and Palsson BO. Thirteen years of building constraintbased *in silico* models of *Escherichia coli*. *J. Bacteriol.*, 185: 2692–2699 (2003).

Rizzi M, Baltes M, Theobald U and Reuss M. *In vivo* analysis of metabolic dynamics in *Saccharomyces cerevisiae*: 2. mathematical model. *Biotechnol. Bioeng.*, 55: 592–608 (1997).

Rohwer JM and Botha FC. Analysis of sucrose accumulation in the sugar cane culm on the basis of in vitro kinetic data. *Biochem. J.*, 358: 437–445 (2001).

Roman BB, Meyer RA and Wiseman RW. Phosphocreatinekinetics at the onset of contractions in skeletal muscle of MM creatine kinase knockout mice. *Am. J. Physiol. Cell. Physiol.*, 283: C1776–1783 (2002).

Sauro HM. SCAMP: A general-purpose simulator and metabolic control analysis program. *CABIOS*, 9: 441-450 (1991).

Sauro HM. JARNAC: A system for interactive metabolic analysis. In: Hofmeyr JHSH, Rohwer JM, Snoep JL (eds) *Animating the cellular map*: Proceedings of the 9th international meeting on biothermokinetics. Stellenbosch University Press, Stellenbosch, pp: 221-228 (2000).

Savageau MA. *Biochemical Systems Analysis: a Study of Function and Design in Molecular Biology*. Addison-Wesley, Reading, Massachusetts (1976).

Savageau MA. Concepts relating the behavior of biochemical systems to their underlying molecular properties. *Arch. Biochem. Biophys.*, 145: 612–621 (1971).

Schilling CH, Covert MW, Famili I, Church GM, Edwards JS and Palsson BO. Genome-scale metabolic model of *Helicobacter pylori* 26695. *J. Bacteriol.*, 184: 4582–4593 (2002).

Schilling CH, Letscher D and Palsson BO. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.*, 203: 229–248 (2000).

Schuster S. Metabolic pathway analysis in biotechnology. In Kholodenko, B.N. and Westerhoff, H.V., (eds.), *Metabolic engineering in the post genomic era*. Horizon Bioscience, Wymondham, UK (2004).

Schuster S, Fell DA and Dandekar T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.*, 18: 326–332 (2000).

Schuster S and Heinrich R. The definitions of metabolic control analysis revisited. *BioSystems*, 27: 1–15 (1992).

Srere PA. Complexes of sequential metabolic enzymes. *Annu. Rev. Biochem.*, 56: 89–124 (1987).

Srivastava DK and Bernhard SA. Biophysical chemistry of metabolic reaction sequences in concentrated enzyme solution and in the cell. *Annu. Rev. Biophys. Biophys. Chem.*, 16: 175–204 (1987).

Stephanopoulos GN, Aristidou AA and Nielsen J. *Metabolic Engineering-Principles and Methodologies*. Academic Press, New York (1998).

Teusink B, Passarge J, Reijenga CA, Esgalhado E, van der Weijden CC, Schepper M, Walsh MC, Bakker BM, van Dam K, Westerhoff HV and Snoep JL. Can yeast glycolysis be understood in terms of *in vitro* kinetics of the constituent enzymes? Testing biochemistry. *Eur. J. Biochem.*, 267: 5313–5329 (2000).

Thorburn DR and Kuchel PW. Regulation of the humanerythrocyte hexose-monophosphate shunt under conditions of oxidative stress. A study using NMR spectroscopy, a kinetic isotope effect, a reconstituted system and computer simulation. *Eur. J. Biochem.*, 150: 371–386 (1985).

Varma A and Palsson BO. Metabolic flux balancing: Basic concepts, scientific and practical use. *Bio/Technology*, 12: 994–998 (1984).

Westerhoff HV and Chen Y-D. How do enzyme activities control metabolite concentrations? an additional theorem in the theory of metabolic control. *Eur. J. Biochem.*, 142: 425–430 (1984).

Wiechert W. ^{13}C metabolic flux analysis. *Metab. Eng.*, 3: 195–206 (2001).

Wiechert W. Modeling and simulation: tools for metabolic engineering. *J. Biotechnol.*, 94: 37–63 (2002).

Wolf J, Passarge J, Somsen OJG, Snoep JL, Heinrich R and Westerhoff HV. Transduction of intracellular and intercellular dynamics in yeast glycolytic oscillations. *Biophysical J.*, 78: 1145–1153 (2000).

Zimmerman SB and Minton AP. Macromolecular crowding: Biochemical, biophysical, and physiological consequences. *Annu. Rev. Biophys. Biomol. Struct.*, 22: 27–65 (1993).

Chapter 14

METABOLIC NETWORKS

Structure and utilization

Eivind Almaas[1], Zoltán N. Oltvai[2] and Albert-László Barabási[1]
[1]*Center for Network Research and Department of Physics, University of Notre Dame, Notre Dame, IN 46556,* [2]*Department of Pathology, Northwestern University, Chicago, IL 60611*

1. INTRODUCTION

During the last century we have witnessed a dramatic progress in the natural sciences. Most of the developments can be directly related to the reductionist approach, which presumes that the often complex behavior of a system can be predicted and understood from the detailed knowledge of the system's (often identical) elementary constituents. However, it is by now clear that our ability to understand simple fundamental laws governing the individual "building blocks" is a far cry from being able to predict the overall behavior of a complex system (Anderson, 1972). Furthermore, for most complex systems there exists a considerable variation in the nature of both the elementary building blocks and their interactions requiring novel methods capable of analyzing and predicting their large-scale behavior.

In the last few years network approaches have shown great promise as a new tool to analyze and understand complex systems (Strogatz, 2001; Albert and Barabasi, 2002; Dorogovtsev and Mendes, 2003; Bornholdt and Schuster, 2003; Pastor-Satorras and Vespignani, 2004). For example, technological information systems like the internet and the world-wide web are naturally modeled as networks, where the nodes are routers (Faloutsos *et al.*, 1999; Vázquez *et al.*, 2002) or web-pages (Albert *et al.*, 1999; Lawrence and Giles, 1999; Broder *et al.*, 2000) and the links are physical wires or URL's respectively. Human society is also naturally described within the framework of network analysis, with people as nodes and the links between the nodes being either friendships (Milgram, 1967), collaborations (Kochen,

1989; Wasserman and Faust, 1994), sexual contacts (Liljeros *et al.*, 2001), or co-authorship of scientific papers (Redner, 1998; Newman, 2001) to name just a few possibilities. It seems that the closer we look at the world surrounding us, the more we find entangled and interacting webs, and that to describe them we need to understand the architecture of the various networks, nature and technology offers us.

In the biological sciences we can represent systems as disparate as food webs in ecology and biochemical interactions in molecular biology as networks. In particular, the complex interactions of the various types of intracellular molecules offer a wide range of structures whose salient features are well captured by a network concept. Important examples include the many interactions between genes, proteins and metabolites. The development of high-throughput measurement tools in molecular biology during the last several years has made available a huge amount of genomic- and postgenomic data. For example, in the fields of transcriptomics and proteomics there is now a plethora of data on protein levels under various conditions, and genome wide analysis of gene expression at the mRNA level is now routine (Pandey and Mann, 2000; Caron *et al.*, 2001; Burge, 2001). Furthermore, protein-protein interaction maps have been generated for a variety of organisms including viruses (Flajolet *et al.*, 2000), prokaryotes, like *Helicobacter pylori* (Rain *et al.*, 2001) and eukaryotes, like *Saccharomyces cerevisiae* (Ito *et al.*, 2000; 2001; Schwikowski *et al.*, 2000; Uetz *et al.*, 2000; Gavin *et al.*, 2002; Ho *et al.*, 2002; Jeong *et al.*, 2001) and *Caenorrhabditis elegans* (Walhout *et al.*, 2000). In this chapter we will discuss recent results and developments in the study and characterization of metabolic networks.

2. STRUCTURE AND CHARACTERIZATION OF NETWORKS

We begin by discussing some of the tools of network analysis, and the key properties they reveal of metabolic networks. It is important to realize that the representation of different systems as networks has uncovered surprising similarities, many of which are intimately tied to power laws. Although the details of the networks, e.g., the explicit nature of a node and the nature of its interactions with other nodes, are frequently quite unique, the overall statistical features of different networks can be very similar. The simplest statistical measure of a network property is the average number of nearest neighbors of a node, also called the average degree $\langle k \rangle$. A natural refinement of this property, revealing deeper insights into an organism's metabolic network organization, is the distribution of nearest

neighbors $P(k)$. For a surprisingly large number of networks, the degree distribution is best characterized by the power law functional form (Barabási and Albert, 1999) (Figure 1a),

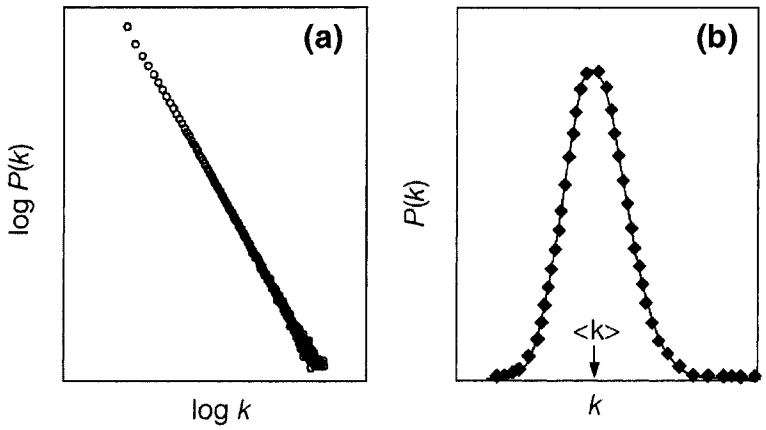


Figure 1. Characterizing degree distributions. For the power-law degree distribution (a), there exists no typical node, while for distributions with a single peak (see (b)) most nodes are well represented by the degree of the average (typical) node.

$$P(k) \sim k^{-\alpha}. \tag{1}$$

However, if the degree distribution is instead single-peaked (e.g., Poisson or Gaussian) as in Figure 1b, the majority of the nodes would be well described by the average degree, and hence, the properties of a "typical" node. In contrast, for networks with a power-law degree distribution, the majority of the nodes have only one or two neighbors while coexisting with many nodes with hundreds and a few even with thousands of nearest neighbors. For the power law networks there exists no typical node, and thus they are often referred to as "scale-free". In Figure 2a-c respectively, we show the degree distributions of the metabolic networks of the three notably disparate organisms of *Archaeoglobus fulgidus* (archae), *Escherichia coli* (eubacterium) and *C. elegans* (eukaryote), all adhering to a power law functional form (Jeong *et al.*, 2000). The hypothesis that the scale-free network structure probably is a universal feature of metabolic networks is further strengthened (Figure 2d) by the presence of a power law when averaging over 43 organisms (Jeong *et al.*, 2000).

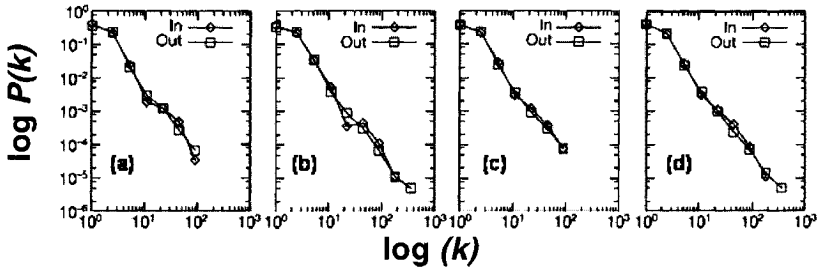


Figure 2. Degree distributions of metabolic networks. The degree distribution displays a power law in both the in- and the out degrees for (a) *A. fulgidus* (archae), (b) *E. coli* (eubacterium), (c) *C. elegans* (eukaryote), and (d) when averaged over 43 organisms (Jeong *et al.*, 2000).

In order to investigate the local network structure, we use the clustering of a node C_i, which measures the degree to which the neighborhood of a node resembles a complete subgraph (Watts and Strogatz, 1998). The clustering of a node i can also be thought of as the probability that two neighboring nodes are also neighbors of each other (i.e., form a triangle). For a node i with degree k_i the clustering is defined as

$$C_i = \frac{2n_i}{k_i(k_i - 1)}, \tag{2}$$

representing the ratio of the number of actual connections (n_i) between the neighbors of node i to the number of possible connections. For a node, which is part of a fully interlinked cluster (all the nodes are connected to each other), $C_i = 1$, while $C_i = 0$ for a node which acts as a bridge between different clusters or parts of the network. Accordingly, the overall clustering coefficient of a network with N nodes, given by

$$\langle C \rangle = \sum C_i / N,$$

represents a measure of a network's potential modularity. By studying the clustering of nodes with a given degree k, information about the actual modular organization of a network can be uncovered (Ravasz *et al.*, 2002; Ravasz and Barabasi, 2003; Dorogovtsev *et al.*, 2002; Vázquez *et al.*, 2002): For all metabolic networks available, $C(k)$ behaves like the power law

$$C(k) \sim k^{-\delta}, \tag{3}$$

suggesting the existence of a hierarchy of nodes with different degrees of modularity (as measured by the clustering coefficient) overlapping in an iterative manner (Ravasz *et al.*, 2002). In Figure 3a-c we show the clustering as function of k for the organisms *Aquidex aeolicus* (archaea), *E. coli* (eubacterium) and *S. cerevisiae* (eukaryote), respectively. In Figure 3d $C(k)$ is averaged over 43 organisms, displaying a robust power-law behavior.

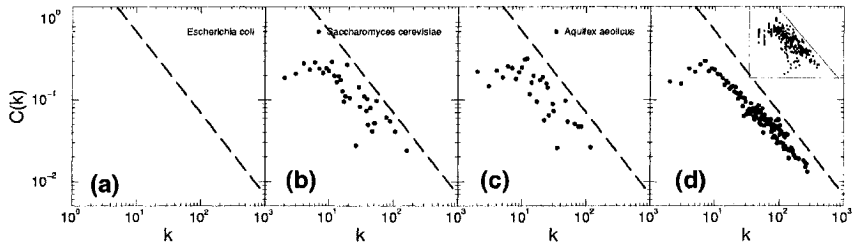


Figure 3. The clustering of metabolic networks. The average clustering as function of node degree k for (a) *A. aeolicus* (archaea), (b) *E. coli* (eubacterium), (c) *S. cerevisiae* (eukaryote), and (d) averaged over 43 organisms, displays a power law behavior (Ravasz *et al.*, 2002). The dashed lines represent C(k)~1/k. The inset in (d) displays all the 43 organisms together.

3. IMPORTANT NETWORK MODELS

Several network models are currently available to model a multitude of aspects of real networks. In the following, we will focus only on three such models, namely the random network model, the scale-free model and the hierarchical model. These models are then compared to the features actually observed in metabolic networks.

3.1 Random network models

While graph theory initially focused on regular graphs, since the 1950's the properties of large networks with no apparent design principles were presumed to be well described by random graphs (Bollobas, 1985). The random graph model represents the simplest and most straightforward realization of a complex network. According to the Erdos-Renyi (ER) model of random networks (Erdos and Renyi, 1960), we start with N nodes and connect every pair of nodes with probability p, creating a graph with approximately $pN(N-1)/2$ randomly distributed edges (Figure 4a). For

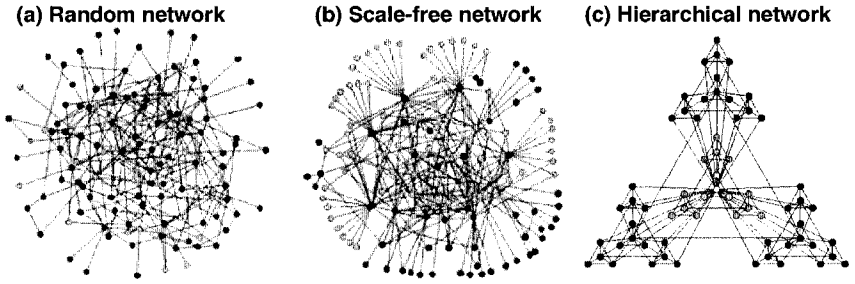


Figure 4. Graphical representation of three network models: (a) The ER (random) model, (b) the BA (scale-free) model and (c) the hierarchical model. Panel (c) demonstrates the iterative construction of a hierarchical network by starting from a fully connected cluster of four nodes (light gray). This cluster is then copied three times (gray) while connecting the peripheral nodes of the replicas to the central node of the starting cluster. We end up with a 64-node scale-free hierarchical network by once more repeating this replication and connection process (dark gray nodes). In panels (a) and (b) we emphasize the difference between the ER and the BA networks by shading the five nodes with the highest number of links gray and their first neighbors light gray. For the scale-free network we reach more than 60% of the nodes using the five largest hubs, while for the random network only 27% of the nodes are directly accessible from the five most connected nodes, demonstrating the heterogeneous nature of scale-free networks. Note that the networks in (a) and (b) consist of the same number of nodes and links.

this model the degrees follow a Poisson distribution (Figure 5a). Consequently, the typical node is well described by the average degree $\langle k \rangle$ of the network. Furthermore, for this "democratic" network model, the clustering is independent of the node degree k (Figure 5d). Hence, comparing with Figures 2 and 3 we conclude that the ER model does not accurately capture the topological properties of metabolic networks.

3.2 Scale-free network model

In the network model of Barabási and Albert (BA) (Figure 4b), the emergence of a power-law degree distribution is attributed to two crucial mechanisms, both absent from the classical random network model (Barabási and Albert, 1999). First, networks grow through the addition of new nodes linking to nodes already present in the system. Second, in most real networks there is a high probability that a new node links to an existing node with a large number of connections, a mechanism often referred to as preferential attachment. These two principles are implemented as follows: starting from a small core graph consisting of m_0 nodes, a new node with m links ($m \leq m_0$) is added at each time step and connected to the already

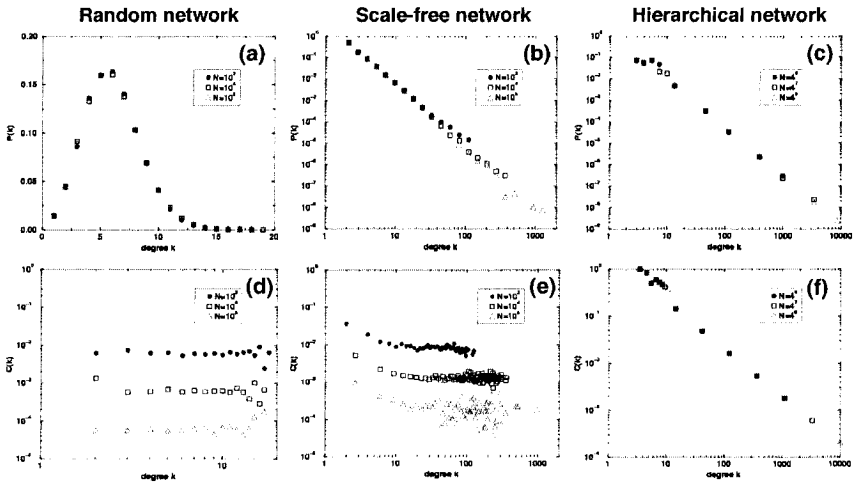


Figure 5. Properties of the three network models. (a) The ER model gives a Poisson degree distribution *P(k)* (the probability that a randomly selected node has exactly *k* links), being strongly peaked around the average degree $\langle k \rangle$ and decaying exponentially for large *k*. For the scale-free (b) and the hierarchical (c) network models the degree distributions instead decay according to the power-law $P(k) \sim k^{-\gamma}$. The average clustering coefficient for nodes with exactly *k* neighbors, *C(k)*, is independent of *k* for both the ER (d) and the scale-free (e) network model, while in contrast (f) $C(k) \sim k^{-1}$ for the hierarchical model (cf. Figure 3).

existing nodes. Each of the *m* new links are then preferentially attached to a node *i* (with k_i neighbors) chosen according to the probability

$$\Pi_i = k_i / \sum_j k_j \,. \tag{4}$$

The simultaneous combination of these two network growth rules gives rise to the observed power-law degree distribution (Figure 5b). In contrast to a random network, the probability that a node is highly connected is statistically significant in a scale-free network; hence many network properties are determined by a relatively small number of highly connected nodes, frequently called "hubs". In Figure 4a and b we show an example of the effect of the hubs on the network structure by coloring the five nodes with largest degrees are red and their nearest neighbors green. While in the ER network only 27% of the nodes are reached by the five most connected ones, more than 60% of the nodes in the scale-free network are covered, demonstrating the key role played by the hubs. Additionally, the hub's dominance of the network topology cause the scale-free networks to be

highly tolerant to random failures (perturbations) while being extremely sensitive to targeted attacks (Albert *et al.*, 2000). Comparing the properties of the BA network model with those of the ER model, we note that while the clustering of the BA network is larger, $C(k)$ however is approximately constant (Figure 5e), suggesting that a hierarchical structure is absent.

3.3 Hierarchical network model

Many real networks are expected to be fundamentally modular, meaning that the network can be partitioned into a collection of modules where each module performs an identifiable task, separable from the function(s) of other modules (Hartwell *et al.*, 1999; Lauffenburger, 2000; Rao and Arkin, 2001). Thus, we expect a seamless combination of the scale-free property with such potential modularity. In order to account for the modularity as reflected in the power-law behavior of $C(k)$ (Figure 3) and a scale-free degree distribution (Figure 2), we can assume that clusters combine in an iterative manner, generating a hierarchical network (Ravasz *et al.*, 2002; Vázquez *et al.*, 2002) (Figure 4c). Such a network emerges from a repeated duplication and integration process of clustered nodes (Ravasz *et al.*, 2002), which in principle can be repeated indefinitely. This process is depicted in Figure 4c, where by starting from a small cluster of four densely linked nodes (blue), one next generates three replicas (green) of this hypothetical initial module and connect the three external nodes of the replicated clusters to the central node of the old cluster. The centers of the replicas are also connected to each other, thus obtaining a large 16-node module. Subsequently, we again generate three replicas (red) of this 16-node module, and connect the replicas as described above, obtaining a new module now consisting of 64 nodes. This (deterministic) hierarchical network model seamlessly integrates a scale-free topology with an inherent modular structure by generating a network that has a power law degree distribution (Figure 5c) with degree exponent $\gamma = 1 + \ln 4 / \ln 3 \approx 2.26$ and a clustering coefficient $C(k)$ which proves to be dependent on k^{-1} (Figure 5f). However, it is of importance to note that modularity does not imply clear-cut sub-networks linked in well-defined ways (Ravasz *et al.*, 2002; Holme *et al.*, 2003). Indeed, the boundaries of modules are often considerably blurred and bridged by highly connected nodes (hubs) which interconnect modules.

4. TOPOLOGIC MODULARITY OF METABOLIC NETWORKS

What are the topological modules, and which function, if any, is associated with them? Ravasz *et al.* (2002) studied the topological overlap matrix, defined as

$$O_T(i,j) = \frac{J(i,j)}{\min(k_i, k_j)}, \tag{5}$$

where $J(i,j)$ is the number of neighbors nodes i and j have in common. If i and j are directly connected, $J(i,j)$ is incremented by one. Since $O_T(i,j)$ estimates the degree to which nodes i and j are members of a tightly interconnected cluster, it can be used as a basis for a hierarchical grouping of the metabolites of a metabolic network (Ravasz *et al.*, 2002). The result of such a grouping is shown in Figure 6. We can observe that the modules are considerably blurred and interwoven, and there exists no unambiguous distinction between two clusters. However, by juxtaposing the modules resulting from purely topological considerations with functional information, we observe the emergence of succinctly defined clusters of metabolites (Ravasz *et al.*, 2002). Moreover, when essentiality information about the various enzymes that connect two substrate to each other (Gerdes *et al.*, 2003) is overlayed on the hierarchical tree gained from the overlap analysis (Figure 7, top), the clustering based approach alone groups the essential enzymes into a few clearly delineated modules. Furthermore, when considering their evolutionary retention (i.e., the propensity to be conserved as an ortholog), a striking correlation between enzyme essentiality and evolutionary conservation is evident (Figure 7) (Gerdes *et al.*, 2003).

5. METABOLIC NETWORK UTILIZATION

It is important to realize that despite their successes, purely topological approaches have intrinsic limitations. Since the activity of the various metabolic reactions or regulatory interactions differs widely, some being highly active under most growth conditions while others are switched on only for rare environmental circumstances, it is necessary to include this information in a network description. Therefore, a biologically relevant understanding of metabolic- and other biochemical reaction networks

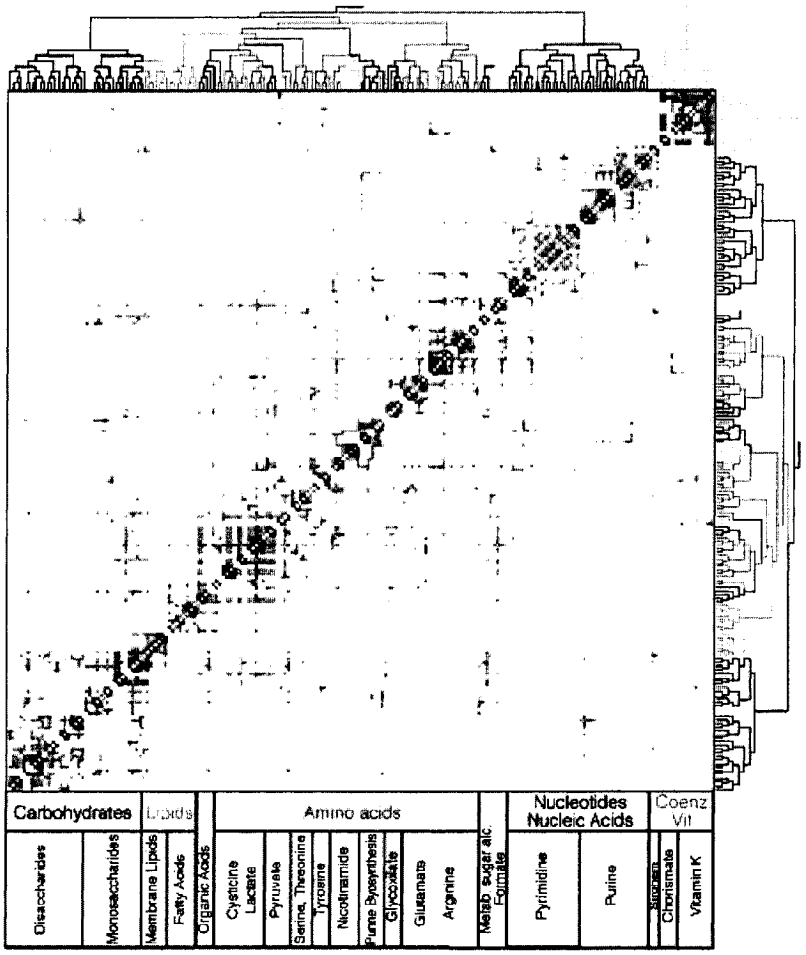


Figure 6. Topological modules. Hierarchically clustering the metabolic network of *E. coli* according to the topological overlap matrix (Ravasz *et al.*, 2002) uncovers modules, as indicated by the color coding of the various branches of the tree according to the dominant biochemical classification of their substrates. The color coding of the matrix indicates the degree of topological overlap, from black ($O_T(i,j) \approx 1$) through grey ($O_T(i,j) \approx 0.5$) to dark grey ($O_T(i,j) \approx 0.05$). Elements with zero overlap are not colored (white).

requires us to consider the intensity (i.e., strength), the direction (when applicable), and the temporal aspects of the interactions. While so far we know little about the temporal aspects of the various metabolic reactions, recent results have added insights on how the strength of the interactions (i.e., fluxes) of the metabolic reactions are organized (Almaas *et al.*, 2004).

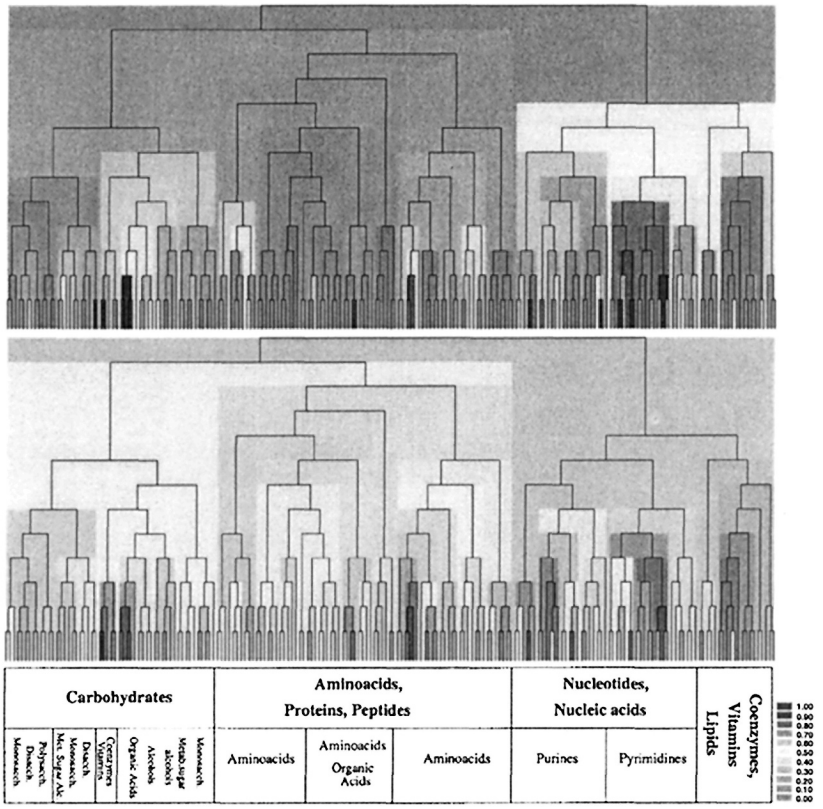


Figure 7. Topological modules, enzyme essentiality and evolutionary retention. The branches of the hierarchical tree (Ravasz *et al.*, 2002) obtained from the topological overlap (see Figure 6) are shaded according to essentiality (top panel) and evolutionary retention (ERI) of a metabolic enzyme (bottom panel) (Gerdes *et al.*, 2003). Enzymes with high levels of essentiality and ERI largely overlap, indicating a strong correlation between metabolic network topology and enzyme importance.

A natural measurement of interaction strength for a metabolic network is given by the flux of the metabolic reactions, representing the amount of substrate being converted to a product within unit time. Recent metabolic flux-balance approaches (FBA) (Edwards and Palsson, 2000; Edwards *et al.*, 2001; 2002; Ibarra *et al.*, 2002; Segre *et al.*, 2002) make it feasible to calculate the flux for each reaction. This has markedly improved our ability to generate quantitative predictions on the relative importance of the various reactions, leading to experimentally testable hypotheses. The much utilized

FBA approach can be stated as follows: Starting from a stoichiometric matrix representation of the *E. coli* K12 MG1655 metabolic network, which contains 537 metabolites and 739 reactions (Edwards and Palsson, 2000; Edwards *et al.*, 2001; 2002; Ibarra *et al.*, 2002), the steady state concentrations of all metabolites satisfy the relation

$$\frac{d}{dt}[A_i] = \sum_j S_{ij} v_j = 0, \tag{6}$$

where S_{ij} is the stoichiometric coefficient of metabolite A_i in reaction j and v_j is the flux of reaction j. We adhere to the convention of $S_{ij} < 0$ ($S_{ij} > 0$) if metabolite A_i is a substrate (product) in reaction j, and we constrain all fluxes to be positive by dividing each reversible reaction into two "forward" reactions with positive fluxes. Any vector of positive fluxes $\{v_j\}$ which satisfies Eq. (6) corresponds to a stoichiometrically allowed state of the metabolic network, and hence, a potential state of operation of the cell.

Assuming that cellular metabolism is in a steady state and optimized for the maximal growth rate (Edwards *et al.*, 2001; Ibarra *et al.*, 2002), FBA allows us to calculate the flux for each reaction using linear optimization. This provides a measure of each reaction's relative activity (Almaas *et al.*, 2004). In a manner similar to that of the degree distribution, the flux distribution of *E. coli* displays a strong overall inhomogeneity: reactions with fluxes spanning several orders of magnitude coexist under the same conditions (Figure 8a). This is captured by the flux distribution for *E. coli* which follows a power law, where the probability that a reaction has flux v is given by $P(v) \sim (v + v_0)^{-\alpha}$. The flux exponent is predicted to be $\alpha = 1.5$ by FBA methods (Almaas *et al.*, 2004). In a recent experiment (Emmerling *et al.*, 2002) the strength of the various fluxes of the central metabolism was measured, revealing the power-law flux dependence $P(v) \sim v^{-\alpha}$ with $\alpha \cong 1$ (Figure 8b) (Almaas *et al.*, 2004). This power law behavior indicates that the vast majority of the metabolic reactions have quite small fluxes, while coexisting with a few reactions with very large flux values.

The observed flux distribution is compatible with two quite different potential *local* flux structures. A homogeneous local organization would imply that all reactions producing (consuming) a given metabolite, have comparable flux values. On the other hand, a more delocalized "hot backbone" is expected if the local flux organization is heterogeneous, such that each metabolite has a dominant source (consuming) reaction.

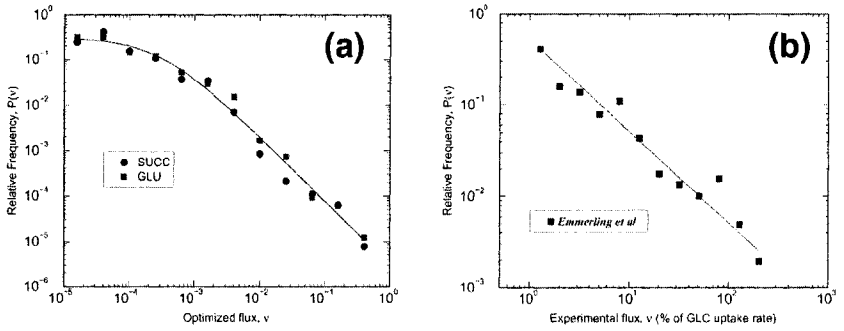


Figure 8. Flux distribution for the metabolism of *E. coli*. (a) Flux distribution when maximizing the biomass production on succinate (circle) and glutamate (square) rich uptake substrates. The solid line corresponds to the power law fit $P(v) \sim (v+v_0)^{-\alpha}$ with $v_0 = 0.0003$ and $\alpha = 1.5$. (b) The distribution of experimentally determined fluxes (see Emmerling *et al.*, 2002) from the central metabolism of *E. coli* also displays power-law behavior which is best fit to $P(v) \sim v^{-\alpha}$ with $\alpha = 1$.

To distinguish between these two scenarios for each metabolite i produced (consumed) by k reactions, we define the measure (Barthelemy *et al.*, 2003; Derrida and Flyvbjerg, 1987)

$$Y(k,i) = \sum_{j=1}^{k} \left(\frac{\hat{v}_{ij}}{\sum_{l=1}^{k} \hat{v}_{il}} \right)^2 , \tag{7}$$

where $\hat{v}_{ij}$ is the mass carried by reaction j which produces (consumes) metabolite i. If all reactions producing (consuming) metabolite i have comparable $\hat{v}_{ij}$ values, $Y(k,i)$ scales as $1/k$. If, however, a single reaction's activity dominates Eq. (7), we expect $Y(k,i) \sim 1$, i.e., $Y(k,i)$ is independent of k. For the *E. coli* metabolism optimized for succinate and glutamate uptake (Figure 9) we find that both the *in* and *out* degrees follow the power law $Y(k,i) \sim k^{-0.27}$, representing an intermediate behavior between the two extreme cases (Almaas *et al.*, 2004). This indicates that the large-scale inhomogeneity observed in the overall flux distribution is increasingly valid at the level of the individual metabolites as well: the more reactions consume (produce) a given metabolite, the more likely it is that a single reaction carries the majority of the flux.

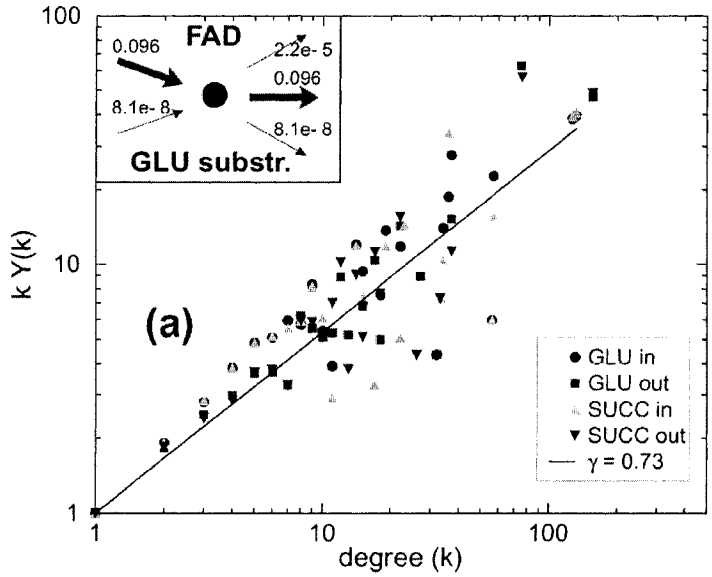


Figure 9. Characterizing the local inhomogeneity of the metabolic flux distribution. The measured $kY(k)$ (see Eq. (7)) shown as function of k for incoming and outgoing reactions for fluxes calculated on both succinate and glutamate rich substrates, averaged over all metabolites, indicating $Y(k) \sim k^{-0.27}$, as the straight line in the figure has slope $\gamma = 0.73$. Inset: The non-zero mass flows $\hat{v}_{ij}$ producing (consuming) flavin adenine dinucleotide (FAD) on a glutamate rich substrate.

6. UTILIZATION AND REGULATION OF METABOLIC REACTIONS

The local flux inhomogeneity described above suggests that we can identify a single reaction dominating the production or consumption of most metabolites. Henceforth, we can construct a simple algorithm which systematically removes, for each metabolite, all reactions but the one providing the largest incoming and outgoing flux contribution. When the largest outgoing flux of metabolite A is identical to the largest incoming flux of metabolite B the *high flux backbone* (HFB) of the metabolism can be uncovered, whose identity is specific to the given growth condition. In Figure 10 we show an example of the HFB for *E. coli* on a minimal medium with succinate as the only carbon source. The HFB mostly consists of reactions linked together, forming a giant component with a star-like topology which includes almost all metabolites produced under the given growth condition. Only a few pathways are disconnected: while these

pathways are parts of the HFB, their end product serves only as the second most important source for some other HFB metabolite. It is interesting to note that groups of individual HFB reactions for the most part overlap with the traditional, biochemistry-based partitioning of cellular metabolism: e. g.,

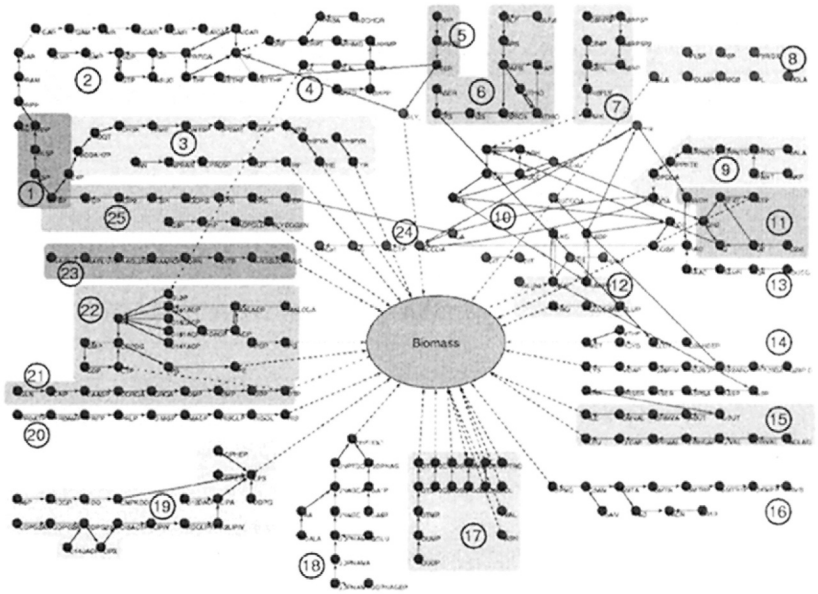


Figure 10. The High Flux Backbone (HFB) of *E. coli* in succinate-rich minimal media. We connect two metabolites A and B with a directed link pointing from A to B only if the reaction with maximal flux consuming A is the reaction with maximal flux producing B. The shading of the metabolites (vertices) and the reactions (edges) indicate a comparison with the HFB of a glutamate rich substrate. Metabolites in black have at least one neighbor in common for the two cases, while those in gray have none. Reactions are thin if they are identical in the two cases, gray if a different reaction connects the same neighbor pair and thick if this is a new neighbor pair. Thus, the gray nodes and links highlight changes in the wiring diagram while changing from succinate to glutamate rich conditions. The numbers identify the various biochemical pathways: (1) Pentose Phospate, (2) Purine Biosynthesis, (3) Aromatic Amino Acids, (4) Folate Biosynthesis, (5) Serine Biosynthesis, (6) Cysteine Biosynthesis, (7) Riboflavin Biosynthesis, (8) Vitamin B6 Biosynthesis, (9) Coenzyme A Biosynthesis, (10) TCA Cycle, (11) Respiration, (12) Glutamate Biosynthesis, (13) NAD Biosynthesis, (14) Threonine, Lysine and Methionine Biosynthesis, (15) Branched Chain Amino Acid Biosynthesis, (16) Spermidine Biosynthesis, (17) Salvage Pathways, (18) Murein Biosynthesis, (19) Cell Envelope Biosynthesis, (20) Histidine Biosynthesis, (21) Pyrimidine Biosynthesis, (22) Membrane Lipid Biosynthesis, (23) Arginine Biosynthesis, (24) Pyruvate Metabolism and (25) Glycolysis.

all metabolites of the citric-acid cycle of *E. coli* are recovered, and so are a considerable fraction of other important pathways, such as those being involved in histidine-, murein- and purine biosynthesis, to mention a few. However, while the detailed nature of the HFB depends on the particular growth conditions, the HFB in general captures the subset of reactions that dominate the activity of the metabolism for this condition. As such, it offers a complementary approach to elementary flux mode analyses (Dandekar *et al.*, 1999; Schuster *et al.*, 2000; Stelling *et al.*, 2002), which successfully determine the available modes of operation for smaller metabolic sub-networks, but whose application to the full *E. coli* metabolism has not yet been possible.

As the flux of the individual metabolic reactions depends on the growth conditions, we need to investigate the sensitivity of the HFB to changes in the environment. In Figure 11, we plot the relationship between the individual fluxes for the two external conditions of using either glucose or glutamate as the carbon source. Surprisingly, only the reactions in the high flux region undergo noticeable flux changes, while the reactions within the intermediate and low flux regions remain practically unaltered (the small shift is caused by increased biomass production in glucose- as compared to glutamate-rich media). We can group the observed flux changes into two categories: First, certain pathways are turned off completely (type I reactions) having zero flux under one growth condition and high flux in the other. These reactions are shown as symbols along the horizontal and vertical axis in Figure 11. In contrast, other reactions remain active but display orders of magnitude shifts in fluxes under the two different growth conditions (type II reactions). With two exceptions, these drastic type II changes are limited to the HFB reactions. The same phenomenon is predicted when we inspect the transitions between various random uptake conditions (Almaas *et al.*, 2004).

To test the generality of this finding, we simulated the effect of various growth conditions by randomly choosing 50% of the potential input substrates and measuring in each input configuration the flux for each reaction. For each reaction the average flux (v), as well as the standard deviation (σ) around this average, was determined by averaging over 5000 random input conditions. It is evident that the $\sigma \sim v$ curve of the small flux reactions all closely follow a straight line with unit slope, supporting the suggestion that small fluxes remain essentially unaltered as the external conditions change (Figure 12). For the high flux reactions, however, there are noticeable deviations from this line, indicating significant flux variations from one external condition to the other. A closer inspection of the flux distribution shows that the reactions along the straight line all have a clear unimodal flux distribution (Figure 13), indicating that shifts in growth

conditions lead to only small changes of their flux values. In contrast, the reactions deviating from the straight line display a bi- or trimodal distribution, indicating that under different growth conditions they exhibit several discrete and quite distinct flux values (Figure 13). Therefore, Figures 11-13 offer valuable insights on how *E. coli* responds to changes in growth conditions: It (de)activates certain metabolic reactions among the HFB metabolites in novel ways without altering the identity of the major pathways that participate in the backbone, resulting in major discrete changes in the fluxes of the HFB reactions. As the metabolic reactions of the HFB are all enzyme-catalyzed, the finding also suggests that the activity of the enzymes exist at distinct modes. Yet, regulatory mechanisms (allosteric, post-translational or transcriptional), responsible for shifting the enzyme activity from one mode to another, are not included in this framework.

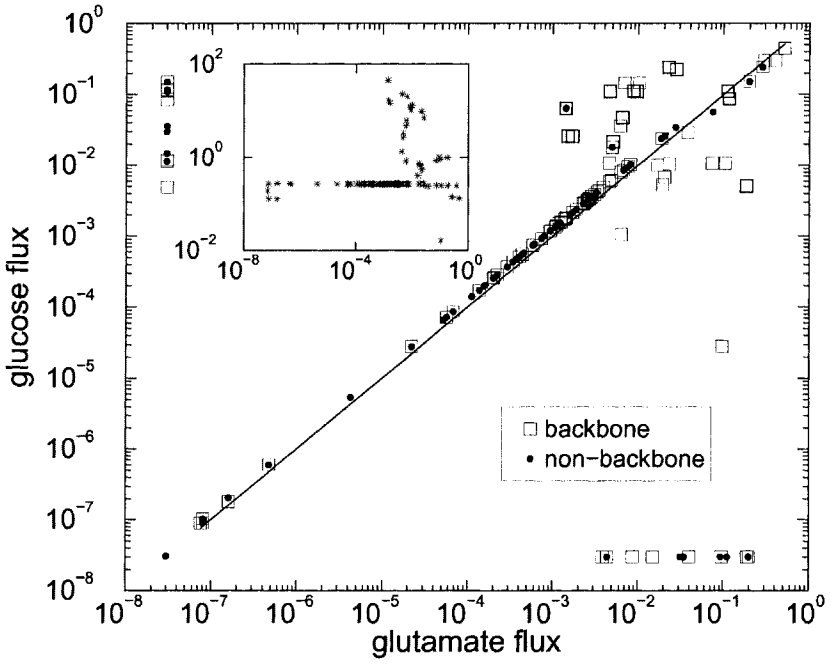


Figure 11. Flux change of individual reactions. When departing from glutamate to glucose rich conditions, some reactions are turned on in only one of the conditions (shown close to the coordinate axes). Reactions which partake of the flux backbone for either of the substrates are squares, the remaining reactions are marked by dots and reactions that change directionality under the two growth conditions are thick squares.

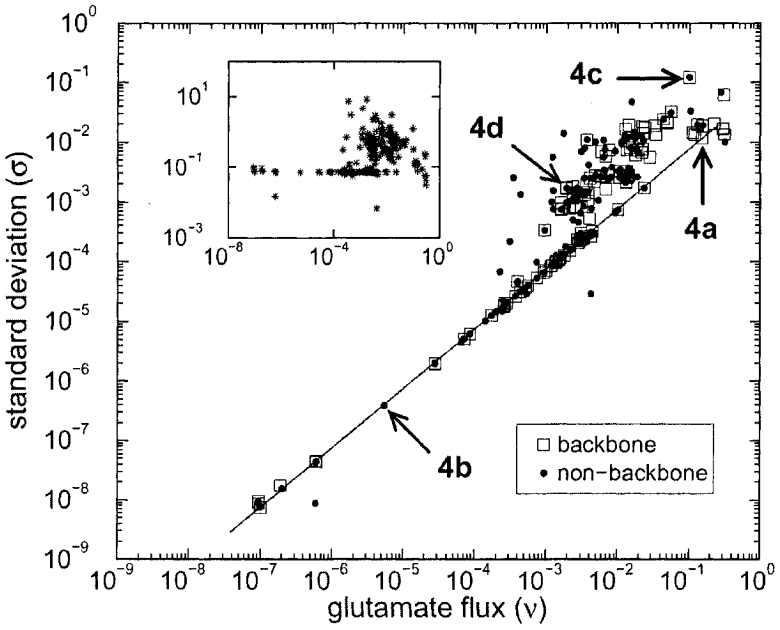


Figure 12. Fluctuations in metabolic fluxes. Absolute value of glutamate flux v_i for reaction i averaged over 50% randomly chosen inputs averaged over 5000 samples, plotted against the standard deviation of that same reaction. The straight line is $y=a\,x$ for reference purpose, with α=0.075. The inset displays the relative flux fluctuation σ_i/v_i per reaction.

7. CONCLUSIONS

During the last few years, it has become evident that power laws are abundant in nature, affecting both the evolution and the utilization of real networks. The power-law degree distribution has become the trademark of scale-free networks and can be explained by invoking the principle of network growth and preferential attachment. In the utilization of complex networks, it is important to realize that most links represent disparate connection strengths or transportation thresholds. For the metabolic network of *E. coli* we have implemented a flux-balance approach and calculated the distribution of link weights (fluxes), which (reflecting the scale-free network

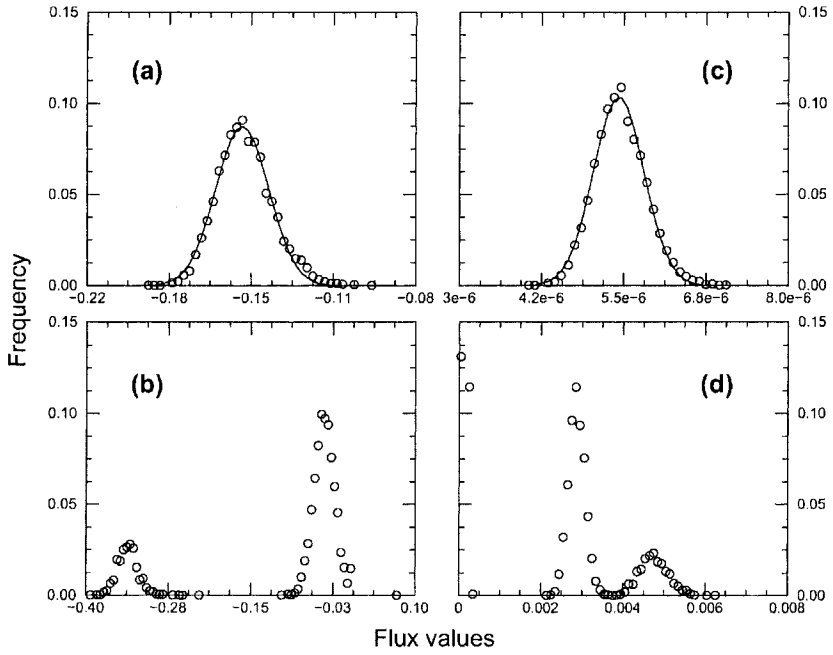


Figure 13. Effect of growth conditions on individual fluxes. Shown is the flux distribution for four select *E. coli* reactions in a 50% random environment. (a) Triosphosphate, isomerase; (b) carbon dioxide transport; (c) NAD kinase; (d) guanosine kinase. Reactions on the σ-ν curve have Gaussian distributions (see (a) and (c)) while reactions off this curve have multimodal distributions (see (b) and (d)) with several discrete flux values. Solid curves correspond to Gaussians derived using the calculated ν and σ values of -0.15 and 0.012 (a) and 5.4e-6 and 3.9e-7 (c).

topology) displays a robust power-law which is independent of any environmental perturbations. Furthermore, this global inhomogeneity in the link strengths is also present at the level of the individual metabolites, allowing us to uncover automatically the high flux backbone of the metabolism. This offers novel insights into the metabolic network's response to changes in the external environment. Defining the nature and the degree of changes under different growth conditions, as well as identifying the regulatory needs and challenges the cell needs to overcome to control these changes, could provide significant insights into metabolic organization and offer valuable inputs for metabolic engineering in the near future.

REFERENCES

Albert R and Barabási AL. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74: 47-97 (2002).

Albert R, Jeong H and Barabási AL. Diameter of the World-Wide Web. *Nature*, 401: 130-1 (1999).

Albert R, Jeong H and Barabási AL. Attack and error tolerance of complex networks. *Nature*, 406: 378-82 (2000).

Almaas E, Kovacs B, Vicsek T, Oltvai ZN and Barabási AL. Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature*, 427: 839-843 (2004).

Anderson PW. More is different. *Science*, 177: 393-6 (1972).

Barabási AL and Albert R. Emergence of scaling in random networks. *Science*, 286: 509-12 (1999).

Barthelemy M, Gondran B and Guichard E. Spatial structure of the Internet traffic. *Physica A*, 319: 633-42 (2003).

Bollobas B. *Random Graphs*. Academic Press, London (1985).

Bornholdt S and Schuster HG. *Handbook of graphs and networks: From the genome to the Internet*. Wiley-VCH, Berlin, Germany (2003).

Broder A, Kumar R, Maghoul F, Raghavan P, Rajalopagan S, Stata R, Tomkins A and Wiener J. Graph structure in the web. *Comput. Netw.*, 33: 309-20 (2000).

Burge CB. Chipping away at the transcriptome. *Nature Genet.*, 27: 232-4 (2001).

Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA, Heisterkamp S, van Kampen A and Versteeg R. The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. *Science*, 291: 1289-92 (2001).

Dandekar T, Schuster S, Snel B, Huynen M and Bork P. Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem. J.*, 343: 115–124 (1999).

Derrida B and Flyvbjerg H. Statistical properties of randomly broken objects and of multivalley structures in disordered-systems. *J. Phys. A: Math. Gen.*, 20: 5273-88 (1987).

Dorogovtsev, S.N., Goltsev, A.V. and Mendes, J.F.F.. Pseudofractal scale-free web. *Phys. Rev. E*, 65: 066122 (2002).

Dorogovtsev SN and Mendes JFF. *Evolution of networks: From biological nets to the Internet and WWW*. Oxford University Press, Oxford (2003).

Edwards JS, Ibarra RU and Palsson BO. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.*, 19: 125-30 (2001).

Edwards JS and Palsson BO. The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. USA*, 97: 5528-33 (2000).

Edwards JS, Ramakrishna R and Palsson BO. Characterizing the metabolic phenotype: A phenotype phase plane analysis. *Biotechnol. Bioeng.*, 77: 27-36 (2002).

Emmerling M, Dauner M, Ponti A, Fiaux J, Hochuli M, Szyperski T, Wuthrich K, Bailey JE and Sauer U. Metabolic flux responses to pyruvate kinase knockout in *Escherichia coli*. *J. Bacteriol.*, 184: 152-64 (2002).

Erdos P and Renyi A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5: 17-61 (1960).

Faloutsos M, Faloutsos P and Faloutsos C. On power-law relationships of the Internet topology. *Comput. Commun. Rev.*, 29: 251-62 (1999).

Flajolet M, Rotondo G, Daviet L, Bergametti F, Inchauspe G, Tiollais P, Transy C and Legrain P. A genomic approach to the hepatitis C virus. *Gene*, 242: 369-79 (2000).

Gavin AC *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415; 141-7 (2002).

Gerdes SY *et al.* Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.*, 185: 5673-84 (2003).

Hartwell LH, Hopfield JJ, Leibler S and Murray AW. From molecular to modular cell biology. *Nature*, 402: C47-52 (1999).

Ho Y *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415: 180-3 (2002).

Holme P, Huss M and Jeong H. Subnetwork hierarchies of biochemical pathways. *Bioinformatics*. 19, p532-9 (2003).

Ibarra RU, Edwards JS and Palsson BO. *Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, 420: 186-9 (2002).

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M and Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci.*, 98: 4569-74 (2001).

Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S and Sakaki Y. Towards 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.*, 97: 1143-47 (2000).

Jeong H, Mason SP, Barabási AL and Oltvai ZN. Lethality and centrality in protein networks. *Nature*, 411: 41-2 (2001).

Jeong H, Tombor B, Albert R, Oltvai ZN and Barabási AL. The large-scale organization of metabolic networks. *Nature*, 407: 651-4 (2000).

Kochen M. (ed.). *The small-world*. ISBN: 0893914797 Ablex Pub., Norwood, N.J. (1989).

Lauffenburger D. Cell signaling pathways as control modules: Complexity for simplicity. *Proc. Natl. Acad. Sci.*, 97: 5031-33 (2000).

Lawrence S and Giles CL. Accessibility of information on the web. *Nature*, 400: 107-9 (1999).

Liljeros F, Edling CR, Amaral LAN, Stanley HE, Aberg Y. The web of human sexual contacts. *Nature*, 411: 907-8 (2001).

Milgram S. The small-world problem. *Psychology Today*, 2: 60-7 (1967).

Montoya JM and Sole RV. Small-world patterns in food webs. *J. Theor. Biol.*, 214: 405-12 (2002).

Newman MEJ. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA*, 98: 404-9 (2001).

Pandey A and Mann M. Proteomics to study genes and genomes. *Nature*, 405: 837-46 (2000).

Pastor-Satorras R and Vespignani A. *Evolution and structure of the Internet: A statistical physics approach*. Cambridge University Press, Cambridge (2004).

Rain J-C, Selig L, DeReuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schächter V, Chemama Y, Labigne A and Legrain P. The protein-protein interaction map of *Helicobacter pylori*. *Nature*, 409: 211-15 (2001).

Rao CV and Arkin AP. Control motifs for intracellular regulatory networks. *Annu. Rev. Biomed. Eng.*, 3: 391 (2001).

Ravasz E and Barabási A-L. Hierarchical organization in complex networks. *Phys. Rev. E*, 67: 026112 (2003).

Ravasz E, Somera AL, Mongru DA, Oltvai ZN and Barabási A-L. Hierarchical organization of modularity in metabolic networks. *Science*, 297: 1551-5 (2002).

Redner S. How popular is your paper? An empirical study of the citation distribution. *Eur. Phys. J. B*, 4: 131-134 (1998).

Schuster S, Fell DA and Dandekar T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechn.*, 18: 326–332 (2000).

Schwikowski B, Uetz P and Fields S. A network of protein-protein interactions in yeast. *Nat. Biotechnol.*, 18: 1257-61 (2000).

Segre D, Vitkup D and Church GM. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci.*, 99: 15112-7 (2002).

Stelling J, Klamt S, Bettenbrock K, Schuster S and Gilles ED. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420: 190–193 (2002).

Strogatz SH. Exploring complex networks. *Nature*, 410: 268-76 (2001).

Uetz P *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403: 623-27 (2000).

Vázquez A, Pastor-Satorras R and Vespignani A. Large-scale topological and dynamical properties of the Internet. *Phys. Rev. E*, 65: 066130 (2002).

Walhout A, Sordella R, Lu X, Hartley J, Temple G, Brasch M, Thierry-Mieg N and Vidal M. Protein interaction mapping in *C. elegans* using proteins involved in vulva development. *Science*, 287: 116-22 (2000).

Wasserman S and Faust K. *Social Network Analysis: Methods and Applications.* Cambridge University Press, Cambridge (1994).

Watts DJ and Strogatz SH. Collective dynamics of small-world networks. *Nature*, 393: 440-2 (1998).

Chapter 15

METABOLIC NETWORKS FROM A SYSTEMS PERSPECTIVE

From experiment to biological interpretation

Wolfram Weckwerth[1], Ralf Steuer[2]
[1]*Max-Planck-Institute of Molecular Plant Physiology, 14424 Potsdam, Germany,* [2]*University, Potsdam, Nonlinear Dynamics Group, Am Neuen Palais 10, 14469 Potsdam, Germany*

1. INTRODUCTION

Recently, we introduced a novel concept for the analysis of metabolite *in vivo* dynamics based on the differential comprehensive identification and quantification of metabolite profiles (Weckwerth *et al.*, 2001, 2004a; Weckwerth 2003). Using a metabolite connectivity matrix it is possible to define key-points at which behaviour is changed in metabolic networks (Weckwerth *et al.*, 2004a). Most importantly, the differences are defined from a systems perspective and not for isolated parts of the biochemical system. Using this approach, novel hypotheses are generated ranging from gene function to pleiotropic effects. To interpret the biological significance of observed changes meaningfully, we developed an integrative profiling approach that complements highly complex connectivity networks with data on protein expression, transcript levels, and environmental data (see Figure 1) (Weckwerth *et al.*, 2004b). The aim of these studies is to provide a global view of *in vivo* biological system dynamics in the context of developmental state, environment, or gene alteration.

Integrative data matrices enable the search for co-regulated biochemical components (Weckwerth *et al.*, 2004b) and the *de novo* identification of regulatory hubs in complex networks. Like the efforts of many other groups described in this book, these studies are groundbreaking attempts at understanding organisms as systems, systems that are more than the sum of linear metabolic pathways. In parallel, the analyses are complementary to

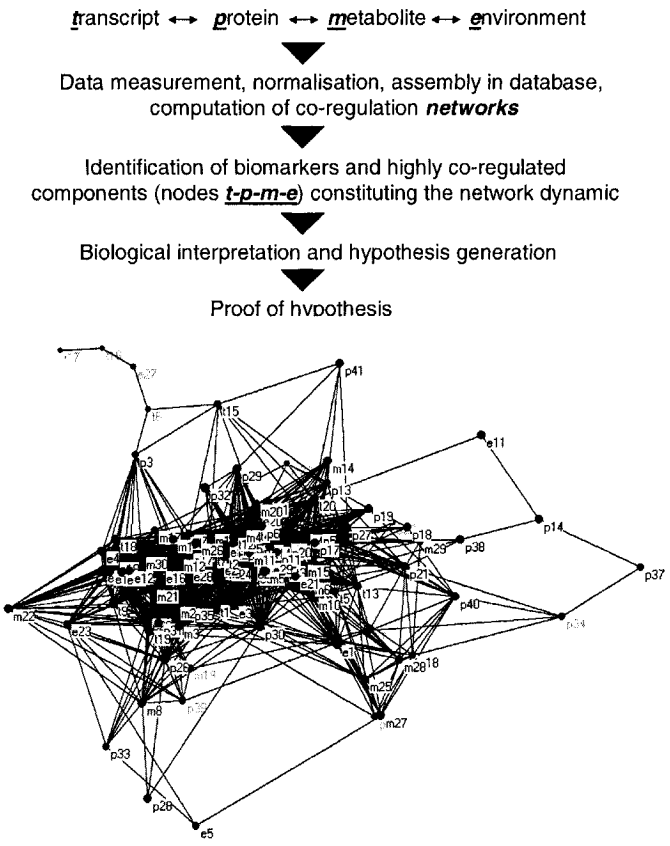


Figure 1. Overall process scheme for the application of "omics-data" In the lower panel, a network is shown exemplifying the interaction between transcripts (prefix t), proteins (prefix p), metabolites (prefix m) and environment (prefix e). The nodes (t,p,m,e) are the components and the edges reveal their distance (for further details see text and Weckwerth, 2003, Weckwerth *et al.*, 2004b).

classical knowledge-driven studies such as investigations of specific pathways based on the chemical structure of the substrates, products, and intermediates (Weckwerth *et al.*, 2000, Schuster *et al.*, 2002). Determining metabolite levels as a measure of metabolic fine and coarse control of pathways has a long tradition in biochemistry (apRees, 1980, Stitt *et al.*, 1988). These measurements enable the detection of diurnal rhythms, enzyme regulation, and serve as clues to understand pathway organization. Results

from such studies have highlighted that biological variability must be minimized to remove confounding parameters and to fix a biological system exactly to the state where the tested hypothesis is effectual. However, it can also be effective to exploit biological variability for multivariate systems analysis (Nicholson *et al.*, 1999, Fiehn *et al.*, 2000, Goodacre *et al.*, 2004, Weckwerth *et al.*, 2004b). Changes at the metabolite level are closely related to the microenvironment of a biological system. Metabolic reaction chains are able to sense environmental stimuli within milliseconds resulting high metabolic fluctuations. It is possible to exploit this biological "noise" to investigate pathway structures or the regulation of gene networks (Arkin *et al.*, 1997, Rao *et al.*, 2002, Steuer *et al.*, 2003). Thus, the measurement and interpretation of *in vivo* dynamics at a systems level represents one of the greatest opportunities (and challenges) to biochemists, especially as a mean to elucidate gene function.

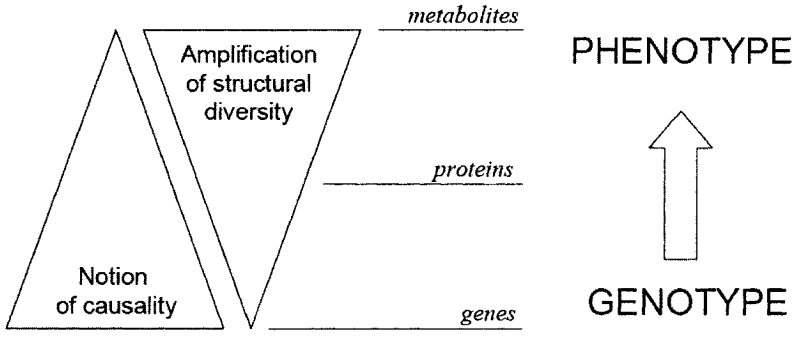


Figure 2. Causality in complex biochemical networks.

The complexity of genome organization - structural diversity, gene duplication and redundancy - inherently implies that molecular phenotypes are not phenomena that can be understood in the context of single gene expressions, but rather as the output of gene interaction networks (Wagner, 1996). The concept of "synthetic lethality" is of considerable interest in this context; the flexibility of genetic interactions results in robust biochemical networks (Sharom *et al.*, 2004). Consequently, interaction networks are best determined by multiparallel measurement of gene and protein expression, and metabolite levels. These interactions can be viewed as correlation networks. However, correlations *per se* contain no information of causality (Wagner, 1997). Nevertheless, correlation of gene and protein expression

analysis and the resulting metabolic phenotype correspond well to our understanding of causality, particularly in discussions of genotype-phenotype relationships (see figure 2).

From the statements above it is evident that co-regulation and causal connectivity can be defined if variables of different levels are analyzed in an integrative data matrix (see Figure 1).

The comprehensive profiling of biological samples requires both statistical and novel data-mining tools to reveal significant correlations. It is further enhanced by profound studies on theoretical metabolic networks (Kacser *et al.*, 1995, Schuster *et al.*, 2000, Papin *et al.*, 2003, Ravasz and Barabasi, 2003, Steuer *et al.*, 2003). Most of these approaches can be divided into the following classes: (i) Studies on network topology and properties based on theoretical reaction pathways and/or regulatory gene networks, (ii) measuring biochemical networks such as protein association, gene, protein and metabolite correlation and co-regulation, and finally (iii) combining experimental data with theoretical modeling. At the moment, there is a clear effort to complement experimental data with some comments or modeling studies on the proposed system structure. System structures are defined with reference to gene annotation or, pathway-, gene-, and protein-databases. Comprehensive invasive investigations such as two hybrid studies and mass spectrometry-based protein-protein association analysis are also used. The modeling of metabolic pathways is complicated by inherently complex cellular and regulatory structures and our gaps in knowledge concerning genome organization. Not all pathways and enzymatic reactions are currently known and it is likely to take years to elucidate functions of unknown and putative proteins in genomes. As a consequence, the models are fragmentary. The presence and absence of pathways under various conditions has to be considered as a major question (Marcotte, 2001, Ihmels *et al.*, 2004).

Thus, many modeling approaches are conclusive for accessible systems like *Escherichia coli* and yeast but not easily applied in more complex systems like plants or mammals. However, the hope is that results from these studies can be extrapolated to more complicated systems (Oliver *et al.*, 1998, Castrillo and Oliver, 2004). This is a reasonable supposition since gene functions can be extrapolated based on sequence homology and conserved protein domain structures.

2. INTEGRATIVE BIOCHEMICAL PROFILING - METABOLITES AND PROTEINS

Omic technologies are able to measure many variables simultaneously in a biological sample (Weckwerth, 2003, Weckwerth *et al.*, 2004b). These measurements represent snapshots of the system enabling the methodical search for correlations between the variables and thus descriptions of the system. These technologies enable protein identification and quantification, mRNA quantification using microarrays, and metabolite measurements using classical methodology such as GCMS, LCMS, NMR, LCUV, etc. The systematic description of living systems requires a substantial sample throughput in parallel with comprehensive analysis of as many constituents as possible. In this context, metabolomics is a promising technique. A global view on *in vivo* dynamics of metabolic networks is achieved with metabolic fingerprinting and metabonomics. These approaches allow high sample throughput but decreased dynamic range and deconvolution of individual components. Here, the reader's attention is directed to excellent reviews covering this topic, including NMR, direct infusion mass spectrometry, and/or IR spectroscopy (Nicholson *et al.*, 1999, Nicholson *et al.*, 2002, Castrillo and Oliver, 2004, Goodacre *et al.*, 2004). A lower sample throughput but unambiguous identification and quantification of individual compounds in a complex sample can be achieved with GCMS and LCMS technology. Owing to major steps forward in these hyphenated technologies, it is possible to adapt specific problems to specific instruments and novel developments in the performance of mass analyzers (see table 1). For GCMS analysis the coupling to TOF mass analyzers is an emerging field. For LCMS, target profiling is usually done with triple quadrupole instruments whereas non-targeted metabolomic approaches require the most sensitive full scan mode combined with peak deconvolution (see Table 1).

A very promising hyphenation technique is capillary electrophoresis (CE) coupled to mass analyzers. This technique is discussed elsewhere in this book (Chapter 6). It is important to note that each type of technology has a bias towards certain compound classes depending on ionisation techniques, detector capabilities, chromatography, etc. One has to decide which technique to apply to a specific question. For metabolomics, GCMS has evolved as an important technology (Sauter *et al.*, 1991, Fiehn *et al.*, 2000, Roessner *et al.*, 2000, Weckwerth *et al.*, 2001).

Very recently, the coupling of GC to a TOF mass detector extended the well-established GC-quadrupole and GC-ion trap technology. The TOF detector has two features, one is mass accuracy and the other is high sensitivity in full scan mode. Mass accuracy is inversely related to sensitivity. High sensitivity in the full scan mode is achieved by time-array

detection using integrated transient recorder technology (ICRTM) (Watson *et al.*, 1990, Leonard and Sacks, 1999, Veriotti and Sacks, 2001). This technology divides the TOF detector into small mass windows, which accelerates data transfer to the computer resulting in high scan speeds of up to 500 full spectra/sec. In comparison to conventional GC-quadrupole MS, this high scan speed enables fast chromatography. Additionally, the signal to

Table 1. Mass analyzer and performance.

Mass analyser	Ionization technique	Chromatography	General Properties	Speediness, sensitivity, and mass accuracy
Quadrupole	ESI, EI, FI, APCI, APPI, MALDI	GC, CE, LC	full scan	Scan speed slow
Triple Quadrupole	ESI, APCI, APPI, MALDI	GC, CE, LC	full scan, MS^2, SIM, SRM, MRM	Full scan slow and insensitive, MRM very fast and sensitive, Exact masses with internal calibration
Triple Quadrupole linear trap	ESI, APCI, APPI, MALDI	CE, LC	Full scan, MS^2, SRM, MRM, MS^n	Full scan medium, MS^n possible.
Ion trap	ESI, APCI, APPI, MALDI	CE, LC	Full scan, MS^2, SIM, MS^n	as for above
Linear ion trap	ESI, APCI, APPI, MALDI	CE, LC	Full scan, SIM, MS^2, MS^n	Very fast full scan, rest as for above
ToF	ESI, EI, FI, APCI, APPI, MALDI	GC, CE, LC	Full scan, source fragmentation	Most sensitive full scan, exact masses with internal calibration
Quadrupole ToF	ESI, APCI, APPI, MALDI	CE, LC	Full scan, MS^2	Most sensitive full scan, exact masses with internal calibration
FTICR	ESI, EI, FI, APCI, APPI, MALDI	GC, CE, MS	Full scan, MS^2, MS^n	Exact masses without internal calibration

ToF = time of flight, FTICR = Fourier Transform Ion Cyclotron Resonance, ESI = electrospray ionisation, EI = electron impact, FI = field ionisation, APCI = atmospheric pressure chemical ionisation, APPI = photoionisation, MALDI = matrix assisted laser desorption ionisation, LC = liquid chromatography, GC = gas chromatography, CE = capillary chromatography, SIM = single ion monitoring, SRM = single reaction monitoring, MRM = multiple reaction monitoring.

noise ratio is increased making the search for low abundant analytes in complex samples possible. These features together provide an improvement over conventional GCMS analysis with respect to the analysis of complex samples as in the metabolomic approach (Weckwerth et al. 2001, Weckwerth et al. 2004a).

Most typically, one has to cope with a high dynamic range of abundance and co-elution of analytes. Thus, accurate deconvolution of chromatogram peaks demands high quality spectra and peak shapes. Recently, we exploited GCTOF analysis of complex plant tissue samples for the distinction of a silent plant phenotype from its wild type using network connectivity analysis (Weckwerth *et al*, 2004a, see also below). Using the full potential of spectral deconvolution, it was possible to extract more than 1000 compounds from the data. However, this process is only semi-automated and due to the necessary manual interpretation, it is time-consuming. In Figure 3, the potential of peak deconvolution in complex sample analysis is exemplified.

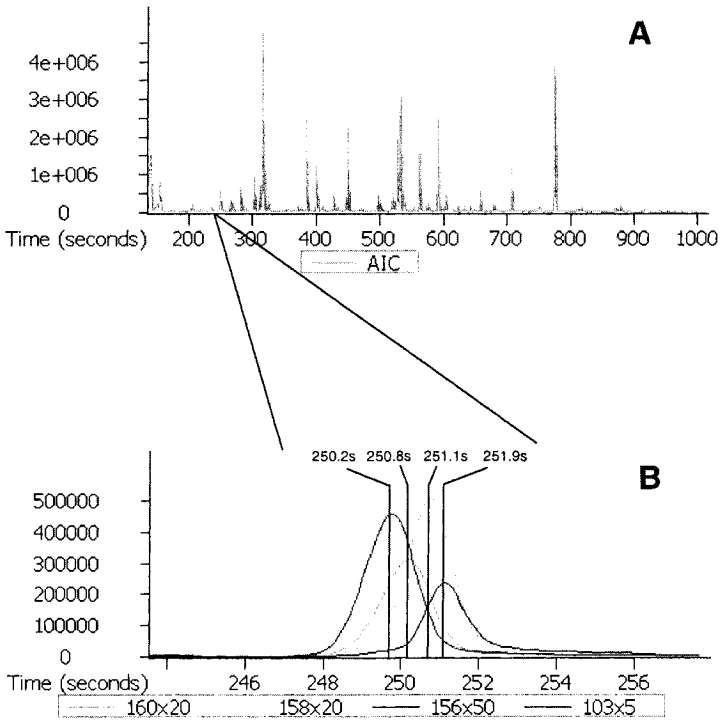


Figure 3. Peak deconvolution in complex samples using GCTOF analysis. (A) Analytical ion chromatogram of a complex plant leaf tissue extract. (B) Different unique masses used for spectral compound identification separated only by 0:3 - 0:8s.

According to the scheme in Figure 1, it is advantageous to inject whole extracts of a plant sample without pre-fractionation of the polar and hydrophobic phase. This is demonstrated in a study where we investigated the application of the integrative extraction method (Figure2) to plant leaf tissue (Weckwerth *et al.*, 2004a). Consequently, all typical metabolite representatives are found in such a chromatogram. The integrated protein/metabolite data matrix enabled the correlation analysis between metabolites and proteins and revealed differential biochemical networks between two *Arabidopsis thaliana* accessions. An interesting finding was the coregulation of L-ascobate peroxidase and inositol pointing to a relationship between ascorbate metabolism and myo-inositol (Weckwerth *et al.*, 2004b). This pathway is only known in animals but was recently evidenced for plants too (Lorence *et al.*, 2004). This is a nice example of how integrative data sets can reveal novel hypotheses.

A major limitation of GCMS is its inability to handle high molecular metabolites larger than for instance tri- to tetra-saccharides, organic diphosphates, or co-factors. Furthermore, it is difficult if not impossible to elucidate unknown structures of metabolites using GCMS alone, although many efforts are under way that combine GCMS with comprehensive spectral libraries and multivariate clustering tools. From this it is clear that data acquisition using a single technology like GCMS can not fulfill the requirements of metabolomic approaches, i.e. comprehensiveness, selectivity, and sensitivity. Alternative technologies have to be combined. LCMS, the most important complementary technology, is a hyphenated technique, established in the late 1980s that combines the high separation power of HPLC with structural information on the components present in complex mixtures. A key development here was electrospray ionisation (ESI) as an interface transferring analyte molecules in solution into gas phase, suitable for mass analysis (Dole *et al.*, 1968, Yamashita and Fenn, 1984). Combined with high-end mass spectrometers, there is no mass range restriction like in GCMS and even complete proteins can be analyzed using this technique (VerBerkmoes *et al.*, 2002). Most importantly, the analytes are not necessarily derivatised, thus providing the parent ion mass as a protonated molecule $[M+H]^+$ or sodiated adduct (e.g. $[M+Na]^+$), in contrast to GCMS and electron impact (EI) ionisation.

Further structural information can be gained by collision-induced decomposition (CID) (Jennings, 2000). In order to obtain fragmentation of parent ions produced by ESI they are isolated and accelerated inside the mass spectrometer using quadrupole mass filters (e.g. triple quadrupole instruments) so as to collide with molecules of the bath gas, usually helium or argon. The resulting fragment spectrum (MS/MS) of an isolated parent ion is then interpreted and can provide important structural information.

Depending on the mass analyzer used, several MS/MS per second can be performed "on the fly". Using so-called quadrupole ion traps (QIT) it is further possible to generate multiple MS/MS spectra of selected fragments (MS^n) of one parent ion mass thereby providing a reasonable information content for structural elucidation of unknown compounds (Stafford, 2002, Tolstikov and Fiehn, 2002). Based on these features LCMS is at present a widely applied technique for the fast and sensitive characterization and quantification of metabolites and pharmaceutical compounds in complex biological fluids like plasma and tissue homogenates. However, most of the benefits of this instrumentation are currently related to the analysis of selected target metabolites in complex mixtures (Niessen, 1999). Consequently, metabolomic analysis using LCMS techniques requires further development and efforts with respect to the non-targeted metabolite analysis in complex mixtures (see Chapters 7 and 9). Deconvolution algorithms especially - comparable to that available for GCMS (Stein and Scott, 1994, Stein, 1999, Tong and Cheng, 1999) - have to be implemented to find peaks without prior knowledge of their abundance, mass spectral characteristics, or retention time. Furthermore, matrix effects and ion suppression have to be considered for the accurate quantification of metabolites in complex samples (Matuszewski *et al.*, 2003).

If these effects are not carefully validated - for instance by spiking targets or internal standards in different concentrations into complex matrices and testing their ESI efficiency - whole data sets are questionable. LCMS technologies provide a reasonable framework to combine various separation techniques. A great challenge remains as regards the analysis of polar compounds. Usually, normal phase or hydrophilic interaction chromatography is used (Tolstikov and Fiehn, 2002). Other alternatives are ion pair reagents, ion exchange chromatography, and novel separation phases combining hydrophobic and hydrophilic interaction such as hypercarb columns (Forgacs, 2002). These techniques are only applicable to a restricted set of polar compounds. Outside of this range, reproducibility and peak shapes are problematic. Since the number of putative metabolites in a complex sample is likely to exceed several thousands, even reversed phase chromatography suffers from restricted peak capacity and separation power. Recently, monolithic columns were introduced providing higher column length and peak capacities as compared to conventional particle-packed columns (Tanaka and Kobayashi, 2003). Combining the separation power of these columns with MS as a further dimension of separation is most promising for metabolomic and proteomic approaches (Tolstikov *et al.*, 2003, Wienkoop *et al.*, 2004). Alternatively, multidimensional chromatography exploiting orthogonal separation techniques may work for metabolomic approaches (Nobuo Tanaka, personal communication). High

resolution mass spectrometry such as FTICRMS (Hughey *et al.*, 2002) (detecting 11000 m/z in a single spectrum) and high resolution chromatography can be combined to increase the number of detectable metabolites in an unbiased way.

A valuable and complementary alternative to the traditional 2DE approach is the multidimensional LCMS analysis of a tryptic digest of a complex protein sample called shotgun proteomics. A major drawback of metabolomic technology yet to be overcome is the vast number of unknown compound structures. Here, LCMS techniques using MS^n, high accuracy mass spectrometers like FTICRMS, offline NMR as well as coupling of LC/NMR are highly required for structure elucidation.

Protein analysis is essentially based on two fundamentally different technologies: (i) protein separation using two-dimensional gel electrophoresis and subsequent MS analysis and (ii) shotgun proteomics on complex protein samples. The methodologies give overlapping but also complementary data on complex samples (Koller *et al.*, 2002, Schmidt *et al.*, 2004). Currently, 2DE has the highest protein resolution capacity of any separation technique. The subsequent identification process, however, is very laborious and depends strongly on protein staining and visualization techniques. Furthermore, the occurrence of many differentially modified protein species and protein isoforms complicate the analysis. A major drawback is the restricted loading capacity of the first dimension facing the enormous dynamic range of protein abundance.

Shotgun proteomics, a multidimensional LCMS analysis of tryptic digest of a complex protein sample is a valuable and complementary alternative to the traditional 2DE approach. A typical qualitative shotgun protein analysis in the range of 200 - 1000 proteins is proposed to be achievable in days (Yates, 2000, Washburn *et al.*, 2001, Aebersold and Mann, 2003, Strittmatter *et al.*, 2003, Wienkoop *et al.*, 2004, Weckwerth *et al.* 2004b). There are many critical issues for using this emerging technology. Database searches, for instance, are prone to generate hundreds of false positives and false negatives depending on the parameters used. Clear rules are missing and protein lists in the literature still provide empirical evaluation of the data. Comparisons among data sets are often limited by the parameters used: for shotgun approaches it will be of value to provide the raw-chromatograms or the MSMS spectra in text-format to allow other researchers to apply their own criteria for protein identification. False positive identifications and protein/peptide modifications (resulting in unreliable identification of high quality spectra) are liable to be the biggest hurdle.

In contrast to metabolomics, there are big differences between qualitative and quantitative protein analysis with respect to throughput. Although 2DE provides the most direct approach for quantification via staining of protein

spots, the process, and especially the reproducibility, is laborious and dependent on sample origin and biological variability. The reliability in the data analysis is always a matter of debate and many replicates are recommended. Only limited access to quantitative data has been demonstrated for shotgun proteomics using, for instance, metabolic or chemical stable isotope labeling techniques (Oda *et al.*, 1999, Goodlett *et al.*, 2001, Smolka *et al.*, 2001, Ong *et al.*, 2002). Quantitative studies are currently restricted to some hundreds of proteins and the time to evaluate the data is in the range of weeks to months to years. For instance, the evaluation of one dataset can take months depending on the software tools (Schmidt *et al.*, 2004). Furthermore, an experiment using differential stable isotope labeling is not a real multiplex analysis providing no statistical confidence of the data. Thus, many efforts are under way to enable the essential analysis of many replicates, considering technical and biological variability (Molloy *et al.*, 2003, Weckwerth, 2003, 2004b). In the world of 2DE, the situation is no better and high biological variation and restricted sample loading capacity (and consequently only high abundance protein detection) may confuse the analysis. More recent research proposes direct quantification from LCMS raw chromatograms without chemical or metabolic labeling, enabling fast access to multi replicate analysis (Chelius *et al.*, 2003, Strittmatter *et al.*, 2003, Wang *et al.*, 2003, Weckwerth *et al.*, 2004a). This seems to be a promising procedure circumventing all severe problems of quantitative chemical labeling (Smolka *et al.*, 2001) and filling the substantial need for replicate analysis. However, direct quantification in complex mixtures is still in the initial stages of development, and peak integration, proof of retention times, normalization to internal standards, fresh weight or TIC are done more or less manually. Direct quantification via peak integration involves all well known bottlenecks in the history of LCMS:

(i) Matrix effects due to ion suppression and enhancement

(ii) Signal to noise ratio, peak shape and retention time

(iii) Resolution capacity and reproducibility of the chromatography.

A major step forward in improved resolution chromatography for the analysis of complex samples is the invention of monolithic capillary columns because these columns provide dimensions not achievable with conventional packed columns owing to reduced backpressure (Premstaller *et al.*, 2001, Tanaka and Kobayashi, 2003, Tolstikov *et al.*, 2003, Wienkoop *et al.*, 2004). It is possible to use 100µm ID x 100cm length with moderate backpressure and appropriate flow rates resulting in very high peak resolution and loading capacity (Weckwerth, unpublished data). Another related way forward is the deconvolution of chromatograms to detect only statistically significant differences in samples (Duran *et al.*, 2003, Kenney and Shockcor, 2003, Tolstikov *et al.*, 2003). However, here one has to fight

against the typical noisy raw-files of GCMS or LCMS runs, as well as retention time shifts. After the detection of significant differences between samples, the structure of the compounds, whether they are peptides or metabolites remain to be identified. Last but not least, the protein coverage in shotgun proteomics can be used as a semi-quantitative measure but needs further proof and method validation (Florens *et al.*, 2002, Tabb *et al.*, 2002).

One major drawback of protein identification and quantification is the extreme dynamic range of protein concentration in tissue samples and no availability of protein amplification techniques analogous to transcript amplification via PCR. Some proteins in plant tissues like ATPase, photosystem I and II, RUBISCO small and large subunit, represent probably 50 - 80% or more of the total leaf tissue protein content. The same holds true for albumin in serum samples (Ahmed *et al.*, 2003). One can imagine that here the loading capacity of any protein separation technique is crucial to identify low abundance or even medium abundance proteins. A way around may be fast and reproducible pre-fractionation of high protein amounts and subsequent shotgun proteomics of the fractions (Wienkoop *et al.*, 2004). Besides the identification of a range of proteins constituting whole pathways, pre-fractionation enables a further confidence level for the identification process in shotgun protein sequencing. This is a very important feature facing the major problems of false positive and false negative identification rates. Other techniques involve the removal of highly abundant proteins using antibodies against, for example, RUBISCO or albumin. However, these techniques are limited in their general applicability.

All the limitations discussed above are likely to apply for metabolomics, too. However, owing to current technical limitations, protein identification and quantification cannot achieve a sample throughput comparable to that of metabolite profiling or metabolomics using GCMS and LCMS, thereby hampering any integrative approach. Thus, the availability of quantitative protein data, for instance a narrow step time series or the characterization of a phenotype for more than a dozen conditions, is missing. However, these data are ultimately needed to describe the protein *in vivo* dynamics of a living system on a statistically significant basis. In contrast, mRNA data are emerging for different kinds of organisms, with several experimental conditions and some even with time series. Often though, averaging over many different experiments, these databases are positive steps towards generating glimpses at the *in vivo* dynamics of biological model systems:

http://www.uni-frankfurt.de/fb15/botanik/mcb/AFGN/atgenex.htm
http://www.arabidopsis.org/info/expression/
http://www.yeastgenome.org/FEContents.shtml

3. METABOLIC NETWORKS

The increasing experimental capabilities described in the last sections have necessitated the simultaneous development of novel approaches to cope with this data algorithmically and conceptually. In this respect, metabolomics profits greatly from new computational methods, which were often already successfully applied in related fields, such as transcriptomics or other 'omic' approaches. Indeed, the most popular types of analysis are based on clustering, principal component analysis (PCA), or other unsupervised or supervised machine learning techniques, and are equally applicable to problems in metabolomics and transcriptomics (Kell *et al.*, 2001, Nicholson *et al.*, 2002, Taylor *et al.*, 2002, Goodacre *et al.*, 2004). Though currently often perceived as 'black box' methods, their power to significantly contribute to an analysis of complex metabolome data has already been demonstrated (Kell, 2002, Goodacre *et al.*, 2004). However, apart from rather pragmatically oriented questions, such as the search for biomarkers to indicate a disease status or a certain deficiency, understanding *global* metabolome data is still in its infancy. Also, the superficial universality of computational methods, irrespective of the particular types of data, often obliterates the unique features of metabolic systems.

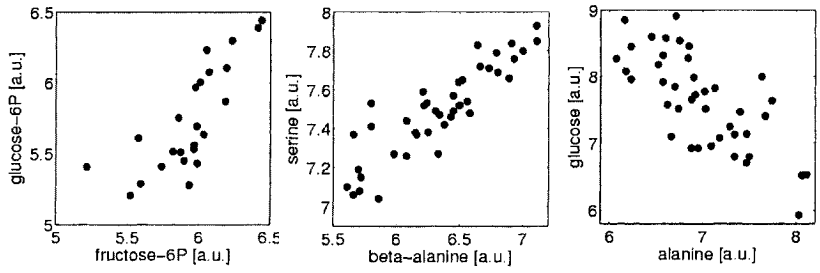


Figure 4. Metabolite levels exhibit a remarkable biological variability. Shown here are metabolite-metabolite scatterplots using samples from tuber tissue (wild type) obtained from an ensemble of identical genotypes under identical conditions with up to 43 measurement for each metabolite (all data are log-transformed and reported in arbitrary units).

Recently, we proposed a supplementary analysis to investigate the structure of metabolism from measurements of intracellular metabolite concentrations (Weckwerth *et al.*, 2001, 2004a, Weckwerth and Fiehn, 2002, Steuer *et al.*, 2003, Weckwerth, 2003).

As already discussed, we observe a remarkable biological variability in the metabolite levels, considerably exceeding the relative technical standard deviation. Importantly, as shown in Figure 4, this variation is not

independent. Rather, metabolites often tend to vary concertedly with other metabolites (Kose *et al.*, 2001, Weckwerth and Fiehn, 2002, Fiehn and Weckwerth, 2003, Steuer *et al.*, 2003, Weckwerth, 2003, Weckwerth *et al.*, 2004a). The resulting correlation between two metabolite concentrations within a given dataset can be quantified using the Pearson correlation coefficient

$$C_{ij} = \frac{\Gamma_{ij}}{\sqrt{\Gamma_{ii}\,\Gamma_{jj}}} \tag{1}$$

where Γ_{ij} denotes the covariance of two metabolite concentrations S_i and S_j .

$$\Gamma_{ij} = \langle S_i S_j \rangle - \langle S_i \rangle \langle S_j \rangle \tag{2}$$

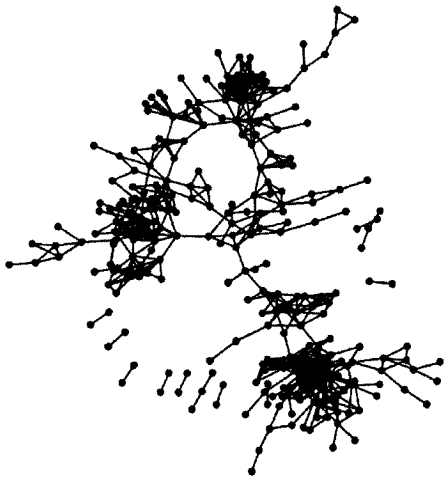


Figure 5. A metabolic correlation network obtained from a dataset of potato leaf samples for different thresholds C^T = 0:8 (refer text). Each dot corresponds to a metabolite, with the links indicating to which other metabolite it correlates stronger than a given threshold. Commonly the threshold is chosen such that the respective correlations are significant with respect to a given probability. Metabolites with no correlations larger than the threshold have been excluded from the plot.

To visualize the resulting pattern of correlations, the metabolites are integrated into a metabolomic correlation network: Each metabolite is assigned coordinates in a two-dimensional plane, such that the pairwise correlations ('similarities') are approximately reflected by the pairwise distances (Arkin and Ross, 1995, Arkin *et al.*, 1997, Steuer *et al.*, 2003, Weckwerth, 2003). Depending on whether the absolute value of their correlation exceeds a given threshold C^T, two metabolites are connected with a link. An example for a correlation network obtained from samples of potato leaf is depicted in Figure 5. Note that here the term 'network' should be understood in parenthesis. In contrast to other biological networks, we introduce the binary nature of the links deliberately and neglect marginal differences in the numerical values of the correlations. The threshold C^T is usually chosen in such a way as to ensure that the respective correlations are significant with respect to a given probability. Consequently, the correlation graph of Figure 5 represents the gross structure of the interconnectivity of metabolites with respect to their pair-wise correlations. As can be observed in Figure 5, this gross structure is remarkably complex and defies an intuitive analysis in terms of traditional biochemical knowledge. While some correlations conform to our intuitive expectations (e.g. F6P and G6P in Figure 4), most bear no obvious relation to the known structure of metabolic pathways (e.g. β-alanine and serine in Figure 4).

Nonetheless, the observed correlations, of course, are not arbitrary but are a direct consequence of the underlying biochemical system. Thus, as a prerequisite for further analysis, we need to achieve a more detailed understanding about how these correlations arise from the underlying metabolic system, what their relationship to biochemical pathways is and, whether we can eventually deduce novel insights about the global organization of metabolic systems from these data.

3.1 Models of metabolic co-regulation

We argue that the observed variability of metabolite concentrations must have biological causes, reflecting the intrinsic flexibility of metabolic networks (Steuer *et al.*, 2003, Weckwerth, 2003) That is, even in a population of identical genotypes under identical environmental conditions, (plant) metabolism is a highly dynamical system and subject to random fluctuations. For example, slight differences in light or nutrient uptake will induce variability in certain metabolic substrates, which in turn affects other metabolites, and ultimately creates an emergent pattern of correlations.

To illustrate this hypothesis, we can make use of a simple *in silico* experiment. Assume that a sequence of reactions, as shown in Figure 6, relies on the availability of certain metabolites (in this case the transport of

triosephosphates (TP) through a membrane). Even under approximately stationary experimental conditions, this supply will never be an exact constant, but will fluctuate due to numerous influences, which are not explicitly included in the model. Numerically, we thus simulate the external pool of triosephosphates TP_{ext} as a time-dependent random variable, using Langevin-type stochastic differential equations (Steuer *et al.*, 2003). The fluctuations in TP_{ext} will then propagate through the pathway and induce characteristic correlations between the remaining metabolites. Figure 7 shows results of numerical 'measurements' of the system. The metabolite concentrations are recorded from successive simulations using independent realizations of the fluctuations (or equivalently, recording the concentrations at successive points in time, so that the time between two 'measurements' is much longer than the correlation time of the system). As can be observed, the induced correlations between the metabolites bear no clear cut relationship to the pathway shown in Figure 6.

While some correlations again conform to our intuitive expectations, such as the strong correlation between G6P and F6P, corresponding to the fast isomerization reaction present in the model, most others defy such a straightforward explanation. For example, we observe a strong positive correlation between F6P and SP (sucrose-phosphate), but a negative correlation between UDP-glucose and SP. However, the observed correlations are not arbitrary. As shown recently (Steuer *et al.*, 2003), it is

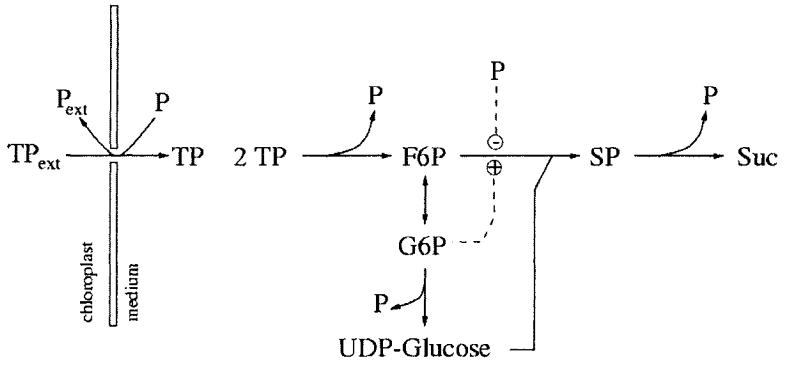


Figure 6. A simple example pathway: The reaction sequence resembles light dependent sucrose synthesis in plants starting from triosephosphate (TP) export from the chloroplast. The pathway is known to be under coarse control (Stitt et al., 1988). For convenience, we concentrate only on two control mechanisms of sucrose- phosphate synthase. This key-enzyme in light-dependent sucrose synthesis is activated via glucose-6-P and inorganic phosphate acts as a partial competitive inhibitor for fructose-6-P. The rate laws and parameters are given in Table 2.

Table 2. Reaction rates corresponding to Figure 6. Note that the purpose of this work, we do not necessarily aim at a realistic description of the system: All reactions are modeled as simple mass action kinetics with arbitrary parameters. $k_1 = 1$, $k_2 = 1$, $k_3 = 1$, $k_{+4} = 10$, $k_{-4} = k_{+4}/q$, $q = 2{:}3$, $k_5 = 0{:}1$. The functions: $f_1([P]) = (1 + [P]/K_p)^{-1}$ and $f_2([G6P]) = (1 + [G6P]=K_g)$ with $K_p = 1{:}0$ and $K_g = 1{:}0$. The total amount of phosphate is conserved: $P_{tot} = P + TP + F6P + G6P + SP$.

Reactions	Rate functions
$TP_{ext} + P \rightarrow TP + P_{ext}$	$v = k_{TP}\,[P][TP_{ext}]$
$TP + TP \rightarrow F6P$	$v = k_1[TP][TP]$
$F6P + UDP\text{-}gluc. \rightarrow SP$	$v = k_2\,[F6P]\,[UDP\text{-}gluc]\,f_1([P])\,f_2([G6P])$
$SP \rightarrow Suc + P$	$v = k_3[SP]$
$F6P \leftrightarrow G6P$	$v = k_{+4}[F6P] - k_{-4}[G6P]$
$G6P \rightarrow UDP\text{-}glucose + P$	$v = k_5\,[G6P]$

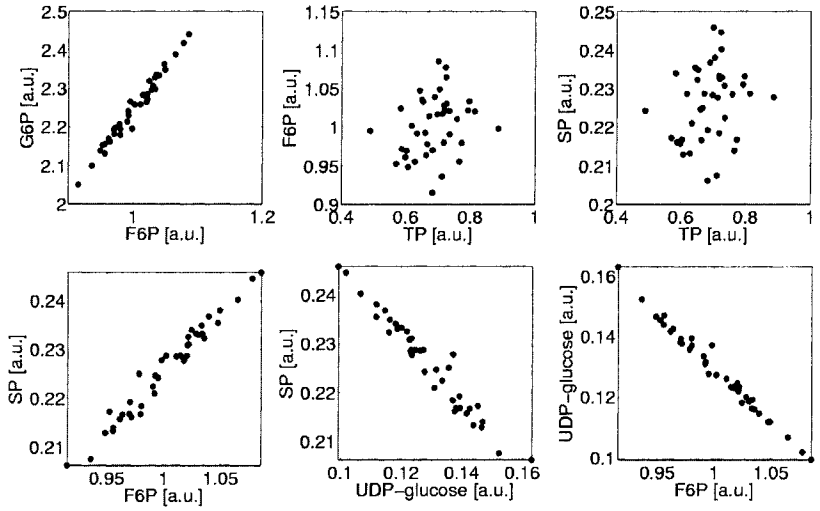


Figure 7. Examples of metabolite-metabolite scatterplots using *in silico* experiments. See text for details. Note that the observed correlations bear no straightforward relationship to the pathway shown in Figure 6.

possible to give an analytical description that provides a link between the observed correlation matrix and the Jacobian of the system (i.e. the linear approximation of the rate equations at the steady state). In particular, given an arbitrary Jacobian **J** and the fluctuation matrix **D**, the resulting covariance matrix Γ and hence the correlation matrix **C**, is given as the solution of a simple linear equation,

$$\mathbf{J\Gamma} + \mathbf{\Gamma J}^T = -2\mathbf{D} \tag{3}$$

where $\mathbf{J}^T$ denotes the transpose of Jacobian. Calculating the Jacobian for the rate equations given in Table 1.2, we can verify Eq. (3) for our simple example considered above. A solution of Eq. (3) together with Eq. (1), yields the correlation matrix **C**,

$$\mathbf{C} = \left(\begin{array}{cccccc} 1.00 & 0.79 & 0.37 & 0.26 & -0.29 & 0.35 \\ 0.79 & 1.00 & 0.27 & 0.11 & -0.16 & 0.28 \\ 0.37 & 0.27 & 1.00 & 0.99 & -0.99 & 0.99 \\ 0.26 & 0.11 & 0.99 & 1.00 & -1.00 & 0.97 \\ -0.29 & -0.16 & -0.99 & -1.00 & 1.00 & -0.98 \\ 0.35 & 0.28 & 0.99 & 0.97 & -0.98 & 1.00 \end{array}\right) \left|\begin{array}{l} \mathrm{TP}_{ext} \\ \mathrm{TP} \\ \mathrm{F6P} \\ \mathrm{G6P} \\ \mathrm{UDP\text{-}gluc.} \\ \mathrm{SP} \end{array}\right. \tag{4}$$

which is in good agreement with the numerical results. In particular, the theoretical solution confirms the unintuitive negative correlations displayed by UDP-glucose.

In general, Eq. (3) establishes a fundamental relationship between the observed covariance and the underlying reaction network. According to our hypothesis, the emergent pattern of correlations within a metabolic system can thus be interpreted as a specific 'fingerprint' of that system. In this way, measuring an ensemble of identical genotypes under identical experimental conditions exploits the intrinsic flexibility and variability in the concentrations to gain additional information about the current state of the system.

Importantly, the structure of Eq. (3) also emphasizes that the observed correlations represent a global property of the system, i.e. they do not depend on any single reaction, but are the combined result of (almost) all reactions in the system. Further, this underscores the fact that correlations observed in metabolome data are fundamentally different from their counterparts in transcriptomics. While for the latter, co-expressed genes are often clustered based on a 'guilt-by-association' principle (D'haeseleer *et al.*, 2000), a similar reasoning does not apply straightforwardly to correlations within metabolic networks. A similar conclusion can be obtained using a slightly different approach, based on metabolic control theory (MCA). Therein the local properties of a metabolic system are given as the (unscaled) elasticity coefficients ε (Heinrich and Schuster, 1996),

$$\epsilon = \left.\frac{\partial \nu}{\partial \mathbf{S}}\right|_{\mathbf{S}^0} \tag{5}$$

where **S** denotes the vector of substrate concentrations. In addition to its elasticities, the global or systemic properties of the system are described by the (unscaled) concentration control coefficients $\mathbf{C}^{\mathbf{S}}$, which characterize the response of a steady state concentration S_i to a change in the activity of a specific reaction v_k,

$$C^{S_i}_{\nu_k} = C^S_{ik} = \frac{\partial S_i / \partial p_k}{\partial \nu_k / \partial p_k} \qquad \text{or} \qquad \mathbf{C^S} = \frac{\partial \mathbf{S}}{\partial \mathbf{p}} \left(\frac{\partial \boldsymbol{\nu}}{\partial \mathbf{p}}\right)^{-1} \tag{6}$$

where the auxiliary parameter p_k acts specifically on the rate v_k (Heinrich and Schuster, 1996).

Thus, in addition to the dynamical stochastic fluctuations considered above, we can likewise assume that each sample, even if drawn from an ensemble of identical genotypes under identical experimental conditions, will still have slightly different parameters in its reaction rates. The concomitant change in two steady state concentrations upon such slight variations of a parameter p_k (acting specifically on a particular reaction rate v_k) is then given as the co-response coefficient of S_i and S_j (Hofmeyr *et al.*, 1993).

$$^{\nu_k}O^{S_i}_{S_j} = \frac{C^S_{ik}}{C^S_{jk}} \tag{7}$$

The co-response coefficient can be interpreted as the slope of the tangent to a plot of S_i against S_j (or $\ln X_i$ against $\ln X_j$, if scaled coefficients are used). For our simple example pathway, we get:

$$^{\mathrm{TP}}\mathbf{O} = \begin{pmatrix} {}^{\nu_{TP}}O^{TP}_{TP} & \cdots & {}^{\nu_{TP}}O^{TP}_{F6P} \\ \vdots & \ddots & \vdots \\ {}^{\nu_{TP}}O^{SP}_{TP} & \cdots & {}^{\nu_{TP}}O^{SP}_{SP} \end{pmatrix} \approx \begin{pmatrix} 1.0 & 0.3 & 0.2 & -0.9 & 1.5 \\ 3.0 & 1.0 & 0.4 & -2.8 & 4.5 \\ 6.7 & 2.3 & 1.0 & -6.3 & 10.0 \\ -1.1 & -0.4 & -0.2 & 1.0 & -1.6 \\ 0.7 & 0.2 & 0.1 & -0.6 & 1.0 \end{pmatrix}$$

Similar to the previous case, the response of a metabolic system in terms of its metabolite slopes is again a global or systemic property of the system.

3.2 Differential metabolic networks

Having established that observed correlations, or slopes, in metabolite scatterplots represent a 'fingerprint' of the system and its current state, we can explore the consequences for metabolomic data analysis. If we assume that the correlations are a global snapshot of the current state of the system, we must expect that plants measured under similar conditions, have likewise similar correlations. On the other hand, gross differences in the regulation within a metabolic system, should manifest itself in a distinct pattern of correlations; for example, as observed in a comparison of potato tuber versus leaf samples. Thus slight changes in the regulatory properties of a metabolic system should be detectable on the level of correlations (Weckwerth *et al.*, 2004a).

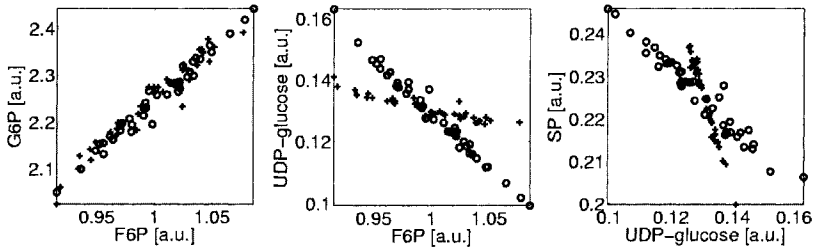


Figure 8. Changes is slope: Correlations in 'wild-type' (circles, o). Correlations without inhibition of v_2 by phosphate (crosses, +). ($f_1([P]) = 1$, $k_2 = 0{:}52$)

To illustrate this, we go back to the simple pathway of Figure 6, which demonstrates how metabolic control of an enzyme rate can change the slope of a metabolite-metabolite correlation. For example in spinach, SPS shows a strong light/dark modulation in activity. Under light conditions it is insensitive to Pi-inhibition (Stitt *et al.*, 1988, Winter and Huber, 2000). According to this, we have now neglected the inhibition exerted by free phosphate P (i.e. $f_1 = 1$ in Table 2). Figure 8 shows the resulting correlations, in comparison with the original case. As can be observed, the slopes have changed markedly though such effects must not necessarily be detectable by the average concentration. We exploited such a differential metabolic network analysis to distinguish a silent transgenic plant phenotype from its wild type (Weckwerth *et al.*, 2004a). Differences in network connectivity of specific metabolites and differences in slopes of metabolite pair correlations

pointed to altered flux partitioning corresponding to slight changes in enzyme activity.

4. OUTLOOK

The simple model described above prompts us to investigate several things in the future: The network topology with respect to correlations and slopes of these correlations has to be further studied. Metabolic connectivity networks might have power law properties (Weckwerth *et al.*, 2004a). This suggests that recent findings about metabolic flux distribution networks (Almaas *et al.*, 2004; de Menezes and Barabasi, 2004) can be related to metabolite connectivity networks. It is self-evident to extend this hypothesis based on the perception that metabolomic networks and the underlying biochemical or metabolic flux networks are ultimately connected as discussed in the sections above. With regard to the inherent causality of these networks it is very important to complement the already highly informative metabolite data with additional knowledge about gene expression, enzymatic activity, and regulation. The proposed model of light-dependent sucrose synthesis in plant leaf tissue brings enzyme activity modulation, for instance via phosphorylation (Stitt *et al.*, 1988; Winter and Huber, 2000), into focus (ter Kuile and Westerhoff, 2001). Consequently, measurement of enzyme activities or activation state (Weckwerth *et al.*, 2004a), protein levels (Weckwerth *et al.*, 2004b), kinase activities (Glinski *et al.*, 2003), and posttranslational modification of proteins will complement metabolomic datasets to nail down causal players of systems level regulation and *in vivo* dynamics.

ACKNOWLEDGMENTS

The authors would like to thank Megan McKenzie for careful reading the manuscript.

REFERENCES

Aebersold R and Mann M. Mass spectrometry-based proteomics. *Nature*, 422: 198–207 (2003).

Ahmed N, Barker G, Oliva K, Garfin D, Talmadge K, Georgiou H, Quinn M and Rice G. An approach to remove albumin for the proteomic analysis of low abundance biomarkers in human serum. *Proteomics*, 3: 1980–1987 (2003).

apRees T. Integration of pathways of synthesis and degradation of hexose phosphates. In Preiss, J. (ed.), *The Biochemistry of Plants*, volume 3, pages 1–29. Academic Press, New York (1980).

Arkin A and Ross J. Statistical construction of chemical-reaction mechanisms from measured time-series. *J. Phys. Chem.*, 99: 970–979 (1995).

Arkin A, Shen PD and Ross J. A test case of correlation metric construction of a reaction pathway from measurements. *Science*, 277: 1275–1279 (1997).

Castrillo JO and Oliver SG. Yeast as a touchstone in postgenomic research: Strategies for integrative analysis in functional genomics. *J. Biochem. Mol. Biol.*, 37: 93–106 (2004).

Chelius D, Zhang T, Wang GH and Shen RF. Global protein identification and quantification technology using two-dimensional liquid chromatography nanospray mass spectrometry. *Anal. Chem.*, 75: 6658–6665 (2003).

D'haeseleer P, Liang S and Somogyi R. Genetic network inference: From co-expression clustering to reverse engineering. *Bioinformatics*, 16: 707–726 (2000).

Dole M, Mack LL and Hines RL. Molecular beams of macroions. *J. Chem. Phys.*, 49: 2240–2249 (1968)

Duran AL, Yang J, Wang LJ and Sumner LW. Metabolomics spectral formatting, alignment and conversion tools (msfacts). *Bioinformatics*, 19: 2283–2293 (2003).

Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN and Willmitzer L. Metabolite profiling for plant functional genomics. *Nat. Biotechnol.*, 18: 1157–1161 (2000).

Fiehn O and Weckwerth W. Deciphering metabolic networks. *Eur. J. Biochem.*, 270: 579–588 (2003).

Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL, Witney AA, Wolters D, Wu YM, Gardner MJ, Holder AA, Sinden RE, Yates JR and Carucci, DJ. A proteomic view of the *Plasmodium falciparum* life cycle. *Nature*, 419: 520–526 (2002).

Forgacs E. Retention characteristics and practical applications of carbon sorbents. *J. Chromatogr. A*, 975: 229–243 (2002).

Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, and Kell DB. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.*, 22: 245–252 (2004).

Goodlett DR, Keller A, Watts JD, Newitt R, Yi EC, Purvine S, Eng JK, von Haller P, Aebersold R and Kolker E. Differential stable isotope labeling of peptides for quantitation and de novo sequence derivation. *Rapid Commun. Mass Spectrom.*, 15: 1214–1221 (2001).

Heinrich R and Schuster S. *The Regulation of Cellular Systems*. Chapman and Hall, New York (1996).

Hofmeyr JHS, Cornish-Bowden A, and Rohwer JM. Taking enzyme kinetics out of control: Putting control into regulation. *Eur. J. Biochem.*, 212: 833–837 (1993).

Hughey C, Rodgers R and Marshall A. Resolution of 11 000 compositionally distinct components in a single electrospray ionization Fourier transform ion cyclotron resonance mass spectrum of crude oil. *Anal. Chem.*, 74: 4145-9 (2002).

Ihmels J, Levy R and Barkai N. Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*. *Nat. Biotechnol.*, 22: 86–92 (2004).

Jennings KR. The changing impact of the collision-induced decomposition of ions on mass spectrometry. *Internat. J. Mass Spectrom.*, 200: 479–493 (2000).

Kacser H, Burns JA and Fell DA. The control of flux. *Biochem. Soc. Trans.*, 23: 341–366 (1995).

Kell DB. Metabolomics and machine learning: explanatory analysis of complex metabolome data using genetic programming to produce simple, robust rules. *Mol. Biol. Rep.*, 29: 237–241 (2002).

Kell DB, Darby RM and Draper J. Genomic computing. Explanatory analysis of plant expression profiling data using machine learning. *Plant Physiol.*, 126: 943–951 (2001).

Kenney B and Shockcor JP. Metabonomic studies. *Pharmagenomics*, Nov/Dec 56–63 (2003).

Koller A, Washburn MP, Lange BM, Andon NL, Deciu C, Haynes PA, Hays L, Schieltz D, Ulaszek R, Wei J, Wolters D and Yates JR. Proteomic survey of metabolic pathways in rice. *Proc. Natl. Acad. Sci. USA.*, 99: 11969–11974 (2002).

Kose F,Weckwerth W, Linke T and Fiehn O. Visualizing plant metabolomic correlation networks using clique-metabolite matrices. *Bioinformatics*, 17: 1198–1208 (2001).

Leonard C and Sacks R. Tunable-column selectivity and timeof-flight detection for high-speed gc/ms. *Anal. Chem.*, 71: 5177–5184 (1999).

Lorence A, Chevone BI, Mendes P and Nessler CL. Myoinositol oxygenase offers a possible entry point into plant ascorbate biosynthesis. *Plant Physiol.*, 134: 1200–1205 (2004).

Marcotte EM. The path not taken. *Nat. Biotechnol.*, 19: 626–627 (2001).

Matuszewski BK, Constanzer ML and Chavez-Eng CM. Strategies for the assessment of matrix effect in quantitative bioanalytical methods based on hplc-ms/ms. *Anal. Chem.*, 75: 3019–3030 (2003).

Molloy MP, Brzezinski EE, Hang JQ, McDowell MT and VanBogelen RA. Overcoming technical variation and biological variation in quantitative proteomics. *Proteomics*, 3: 1912–1919 (2003).

Nicholson JK, Connelly J, Lindon JC and Holmes E. Metabonomics: a platform for studying drug toxicity and gene function. *Nat. Rev. Drug Discov.*, 1: 153–161 (2002).

Nicholson JK, Lindon JC and Holmes E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, 29: 1181–1189 (1999).

Niessen WMA. State-of-the-art in liquid chromatography-mass spectrometry. *J. Chromatography A*, 856: 179–197 (1999).

Oda Y, Huang K, Cross FR, Cowburn D and Chait BT. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. USA*, 96: 6591–6596 (1999).

Oliver SG, Winson MK, Kell DB and Baganz F. Systematic functional analysis of the yeast genome. *Trends Biotechnol.*, 16: 373–378 (1998).

Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, and Mann M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics*, 1: 376–386 (2002).

Papin JA, Price ND, Wiback SJ, Fell DA and Palsson BO. Metabolic pathways in the post-genome era. *Trends Biochem. Sci.*, 28: 250–258 (2003).

Premstaller A, Oberacher H, Walcher W, Timperio AM, Zolla L, Chervet JP, Cavusoglu N, van Dorsselaer A and Huber CG. High-performance liquid chromatography-electrospray ionization mass spectrometry using monolithic capillary columns for proteomic studies. *Anal. Chem.*, 73: 2390–2396 (2001)

Rao CV, Wolf DM and Arkin AP. Control, exploitation and tolerance of intracellular noise. *Nature*, 420: 231–237 (2002).

Ravasz E and Barabasi AL. Hierarchical organization in complex networks. *Phys. Rev. E*, 67: 026112 (2003)

Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, Willmitzer L and Fernie AR. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell*, 13: 11–29 (2001).

Roessner U, Wagner C, Kopka J, Trethewey RN and Willmitzer L. Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant J.*, 23: 131–142 (2000).

Sauter H, Lauer M and Fritsch H. Metabolic profiling of plants - a new diagnostic-technique. *Abstr. Pap. Am. Chem. Soc.*, 195: 129 (1991)

Schmidt F, Donahoe S, Hagens K, Mattow J, Schaible UE, Kaufmann SHE, Aebersold R, and Jungblut PR. Complementary analysis of the mycobacterium tuberculosis proteome by twodimensional electrophoresis and isotope-coded affinity tag technology. *Mol. Cell Proteomics*, 3: 24–42 (2004).

Schuster S, Fell DA, and Dandekar T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.*, 18: 326–332 (2000).

Schuster S, Klamt S, Weckwerth W, Moldenhauer F and Pfeiffer T. Use of network analysis of metabolic systems in bioengineering. *Bioproc. Biosyst. Eng.*, 24: 363–372 (2002).

Sharom JR, Bellows DS and Tyers M. From large networks to small molecules. *Curr. Opin. Chem. Biol.*, 8: 81–90 (2004).

Smolka MB, Zhou HL, Purkayastha S and Aebersold R. Optimization of the isotope-coded affinity tag-labeling procedure for quantitative proteome analysis. *Anal. Biochem.*, 297: 25–31 (2001).

Stafford G. Ion trap mass spectrometry: A personal perspective. *J. Am. Soc. Mass Spectrom.*, 13: 589–596 (2002).

Stein SE. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *J. Am. Soc. Mass Spectrom.*, 10: 770–781 (1999).

Stein SE and Scott DR. Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.*, 5: 859–866 (1994).

Steuer R, Kurths J, Fiehn O and Weckwerth W. Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, 19: 1019–1026 (2003).

Stitt M, Wilke I, Feil R and Heldt HW. Coarse control of sucrose-phosphate synthase in leaves - alterations of the kinetic-properties in response to the rate of photosynthesis and the accumulation of sucrose. *Planta*, 174: 217–230 (1988).

Strittmatter EF, Ferguson PL, Tang KQ and Smith RD. Proteome analyses using accurate mass and elution time peptide tags with capillary lc time-of-flight mass spectrometry. *J. Am. Soc. Mass Spectrom.*, 14: 980–991 (2003).

Tabb DL, McDonald WH and Yates JR. Datselect and contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.*, 1: 21–26 (2002).

Tanaka N and Kobayashi H. Monolithic columns for liquid chromatography. *Anal. Bioanal. Chem.*, 376: 298–301 (2003).

Taylor J, King RD, Altmann T and Fiehn O. Application of metabolomics to plant genotype discrimination using statistics and machine learning. *Bioinformatics*, 18: S241–S248 (2002).

Tolstikov V, Lommen A, Nakanishi K, Tanaka N and Fiehn O. Monolithic silica-based capillary reversed-phase liquid chromatography/ electrospray mass spectrometry for plant metabolomics. *Anal. Chem.*, 75: 6737-40 (2003).

Tolstikov VV and Fiehn O. Analysis of highly polar compounds of plant origin: Combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Anal. Biochem.*, 301: 298–307 (2002).

Tong CS and Cheng KC. Mass spectral search method using the neural network approach. *Chemomet. Intell. Lab. Sys.*, 49: 135–150 (1999).

VerBerkmoes NC, Bundy JL, Hauser L, Asano KG, Razumovskaya J, Larimer F, Hettich RL and Stephenson Jr JL. Integrating "top-down" and "bottom-up" mass spectrometric

approaches for proteomic analysis of shewanella oneidensis. *J. Proteome Res.*, 1: 239–252 (2002).

Veriotti T and Sacks R. High-speed gc and gc/time-of-flight ms of lemon and lime oil samples. *Anal. Chem.*, 73: 4395–4402 (2001).

Wagner A. Can nonlinear epigenetic interactions obscure causal relations between genotype and phenotype? *Nonlinearity*, 9: 607–629 (1996).

Wagner A. Causality in complex systems. *Biology and Philosophy*, 14: 83–101 (1997).

Wang WX, Zhou HH, Lin H, Roy S, Shaler TA, Hill LR, Norton S, Kumar P, Anderle M and Becker CH. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.*, 75: 4818–4826 (2003).

Washburn MP, Wolters D and Yates JR. Large-scale analysis of the yeast proteome by multidimensional protein identification. *Nat. Biotechnol.*, 19: 242–247 (2001).

Watson JT, Schultz GA, Tecklenburg RE and Allison, J. Renaissance of gas-chromatography time-of-flight mass-spectrometry - meeting the challenge of capillary columns with a beam deflection instrument and time array detection. *J. Chromatography*, 518: 283–295 (1990).

Weckwerth W. Metabolomics in systems biology. *Ann. Rev. Plant Biol.*, 54: 669–689 (2003).

Weckwerth W and Fiehn O. Can we discover novel pathways using metabolomic analysis? *Curr. Opin. Biotechnol.*, 13: 156–160 (2002).

Weckwerth W, Loureiro M, Wenzel K and Fiehn O. Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc. Natl. Acad. Sci. USA*, 101: 7809–7814 (2004a).

Weckwerth W, Miyamoto K, Iinuma K, Krause M, Glinski M, Storm T, Bonse G, Kleinkauf H and Zocher R. Biosynthesis of pf1022a and related cyclooctadepsipeptides. *J. Biol. Chem.*, 275: 17909–17915 (2000).

Weckwerth W, Tolstikov V and Fiehn O. Metabolomic characterization of transgenic potato plants using gc/tof and lc/ms analysis reveals silent metabolic phenotypes. In *Proceedings of the 49th ASMS Conference on Mass spectrometry and Allied Topics*, volume 1-2. American Society of Mass Spectrometry, Chicago (2001).

Weckwerth W, Wenzel K and Fiehn O. Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. *Proteomics*, 4: 78–83 (2004b).

Wienkoop S, Glinski M, Tanaka N, Tolstikov V, Fiehn O and Weckwerth W. Linking protein fractionation with multidimensional monolithic RP peptide chromatography/mass spectrometry enhances protein identification from complex mixtures even in the presence of abundant proteins. *Rapid Commun. Mass Spectrom.*, 18: 643–650 (2004).

Winter H and Huber SC. Regulation of sucrose metabolism in higher plants: Localization and regulation of activity of key enzymes. *Crit. Rev. Biochem. Mol. Biol.*, 35: 253–289 (2000).

Yamashita M and Fenn JB. Electrospray ion-source – another variation on the free-jet theme. *J. Physical Chem.*, 88: 4451–4459 (1984).

Yates JR. Mass spectrometry - from genomics to proteomics. *Trends Genet.*, 16: 5–8 (2000).

Chapter 16

PARALLEL METABOLITE AND TRANSCRIPT PROFILING

Hypothesis generation for biotechnology

Alisdair R. Fernie, Ewa Urbanczyk-Wochniak and Lothar Willmitzer
Max-Planck-Institute für Pflanzenphysiologie, Am Mühlenberg 1, 14476 Golm, Germany

1. INTRODUCTION

Genome sequencing has driven a revolution in biology. The comprehensive sequence information it provides can be readily stored, accessed and analysed, and is readily transferable between laboratories. The widespread availability of such information is contributing to the development of powerful genetic and analytical resources that are speeding up the rate and precision of experimentation (Stitt and Fernie, 2003; Harrigan and Goodacre, 2003). Since information flow in biological systems follows the sequence DNA to RNA to protein, and gene function is generally described by the latter, the impact of this revolution is most immediate in research directly related to nucleic acids, where our ability to characterize and alter genotypes and allow genome-wide analysis of gene expression has been greatly facilitated. However, significant advances have also been made at the levels of proteins (Shevchenko *et al.*, 1996), metabolites (Raamsdonk *et al.*, 2000; Roessner *et al.*, 2001) and their fluxes (Boros *et al.*, 2003). More recently, combined studies of the transcriptome and a subset of the proteome have been carried out in *Saccharomyces cerevisiae* (Futcher *et al.*, 1999; Gygi *et al.*, 1999; Ideker *et al.*, 2001a;b) facilitating the use of systems biology approaches (Ideker *et al.*, 2001b; Kitano, 2002; Oltvai and Barabasi, 2002).

This chapter describes the parallel analysis of transcript and metabolite levels using studies on potato metabolism as a case study (Urbanczyk-Wochniak *et al.*, 2003). These studies were carried out using an established

GC-quadrupole-MS based metabolite profiling protocol alongside parallel analysis of gene expression using classical EST-based array technology. The principal aims of these experiments were to determine the relative power of these two phenotyping systems in the discrimination of biological systems and to assess whether the combined analysis of transcript and metabolite profiling represents a novel and meaningful approach to the identification of candidate genes for changing the metabolite composition of a biological system. Studies in which transcript and limited metabolite profiles have been integrated have led to improved yields of the medicinally important lovastatin and (+)-geodin from *Aspergillus* (Azkenazi *et al.* 2003). A similar approach has been taken to identify the control of the accumulation of a set of secondary metabolites following hormone application to tobacco cell suspension cultures (Goosens *et al.* 2003). Here we intend to review these studies as well as broader range studies carried out in the bacterial field in which transcriptomic and metabolomic data have been integrated with metabolic flux data (Krömer *et al.*, 2004).

2. TECHNOLOGY PLATFORMS

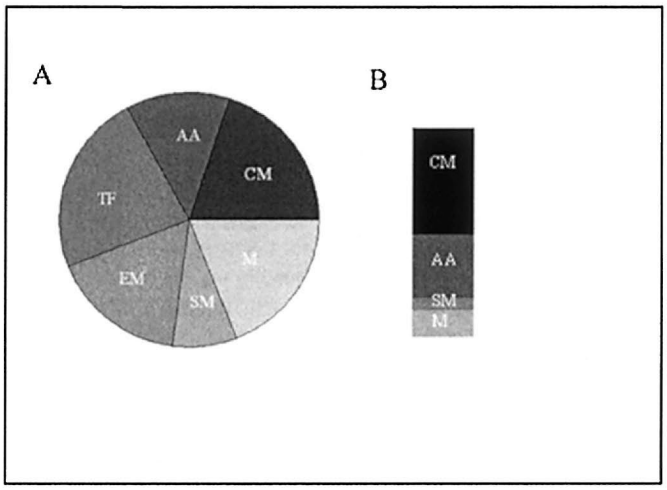


Figure 1. A) Functional categorization of genes belonging to custom Solanaceous macroarray;. B) Classification of the metabolites detected by GC-MS; AA – amino acids metabolism; CM – carbon metabolism ; TF – transcription factors; EM – energy metabolism; SM – secondary metabolism; M – miscellaneous

For our initial profiling experiment we decided to use a crop plant - the potato - rather than a model species (such as *Arabidopis*) to address our

aforementioned aims. The reasons for this were two-fold; firstly the potato tuber is a very homogenous organ that has several well-defined but nevertheless highly related developmental stages and allows the assessment of several well-characterized transgenic lines; secondly, our group is experienced in the analysis of this species on both molecular and biochemical levels (Fernie *et al.*, 2002). That said, the use of a crop species that has not been sequenced at the genome level imposed considerable restriction on the genome coverage possible for hybridization experiments.

As a first step we constructed a custom macroarray consisting of some 2200 Solanaceous ESTs, representing approximately 1000 genes; (Figure 1A) highly biased towards primary metabolism events (Urbanczyk-Wochniak *et al.*, 2003). This macroarray was used to profile the transcript levels of wild type potato tubers at various stages of development and of two distinct transgenic cases: one expressing a yeast invertase and the other a bacterial sucrose phosphorylase. Organic extracts were generated from the same samples and subjected to metabolite profiling by GC-MS, using a well established protocol (Roessner *et al.*, 2000) capable of the detection of over 70 metabolites comprising organic acids, amino acids, sugars, sugar alcohols and a handful of soluble secondary metabolites (Figure 1B). Both our transcript and metabolite profiling methods were therefore heavily biased toward primary metabolism. However, other studies that have integrated transcript and metabolite profiles have used similar instrumentation to determine the levels of even fewer metabolites. A subset of alkaloids and flavonoids and a subset of the transcriptome of tobacco cell suspension cultures were characterized by GC-MS and cDNA-AFLP profiling, following treatment with jasmonate (Goosens *et al.*, 2003), for example. Whilst the determination of the correlation between 21,000 arrayed elements of the *Aspergillus nidulan* genome and the commercially important secondary metabolites Iovastatin and (+)-geodin was performed using a combination of classical array technology, LC-MS methodologies and ultimately the use of NMR for structural confirmation were required for metabolite characterization. More recently an ambitious project aimed at a comprehensive understanding of a lysine producing strain of *Corynebacterium glutamicum* ATCC 13287 exploited transcriptome profiling of this bacterium, at different stages of batch culture via the use of DNA microarrays, in combination with intra- and extracellular measurement of metabolites and isotopomer modeling following supply of ^{13}C and subsequent GC-MS based analysis (Krömer *et al.*, 2004). Although the goals of these researchers were different to those outlined above it is likely that this data set will comprise a rich source of information that could be used in the identification of candidate genes. The successful implementation of a broad range of analytical techniques in the fairly limited number of

examples given here suggest that most properly validated techniques should be able to produce data of sufficient quality for such analyses. However, in most cases the real test of such approaches is some way off, namely, whether the candidate genes identified by such screens can indeed alter metabolite composition in biological systems. Whilst the platforms presented above allow correlations between many transcripts and many metabolites, it is clear that the adoption of whole genome profiling strategies and broad range metabolite profiling strategies using a combination of LC-MS and GC-MS (such as those discussed in Chapter 7) will prove more powerful. It is however, imperative that any such multi-entity analysis (transcript, protein, and metabolite) is applied to the same sample of any given biological material.

3. A REVIVAL OF THE CORRELATIVE APPROACH IN BIOLOGY

The advent and widespread adoption of molecular genetic approaches heralded a sea change in the way that biology was approached in the past three decades with prominence being given to "direct approaches". The previous experimental strategy of defining the importance of a gene or protein by association of a change in that entity under applied experimental conditions was supplanted by a direct functional analysis of the gene by analyzing the effect of its removal (from) or overproduction (in) an organism (see Simchen, 1978 and Stitt and Sonnewald, 1995 for examples of detailed reviews for this subject). The glut of information produced by post-genomic technologies has however seen biology swing full circle and once again embrace the correlative approach. Thus transcript, protein and metabolite profiles of various biological conditions are recorded and the information they contain is described to be diagnostic of these conditions. Whilst such "guilt by association" approaches clearly do not provide the investigator with any causal or mechanistic knowledge of the biological systems they do allow the identification of candidate genes or functions that can then be analysed in detail through complementary studies using genetic manipulations or more detailed study following reductionist principles (Kell and Oliver, 2004).

The first examples of mass correlation analysis naturally come from the most mature of the post-genomic technologies – transcript profiling. These studies have largely focused on the identification of co-regulated genes, operons and geometrical relationships in chromosomes (Allocco *et al.*, 2004; Yamanishi *et al.*, 2003; Sabatti *et al.*, 2002) and more recently have begun to

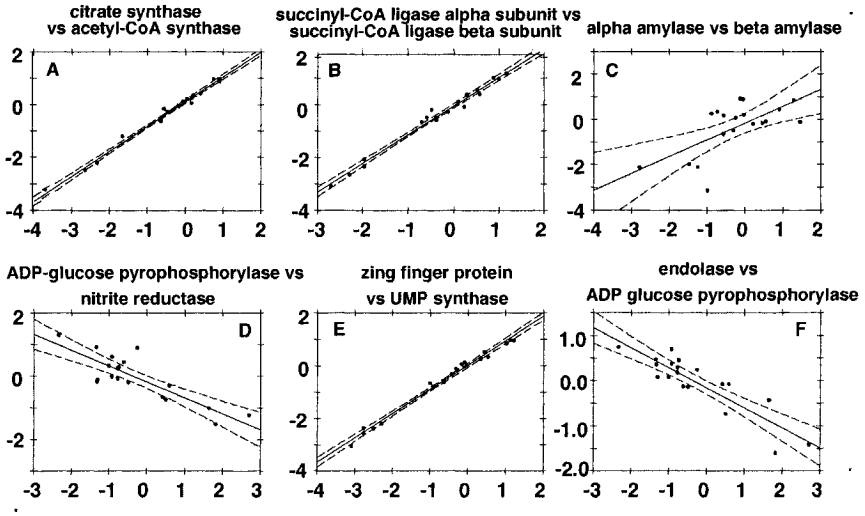


Figure 2. Correlation of transcript levels in the wild-type and the two transgenic lines of potato tuber during development (refer text for details); UMPsynthase – uridine monophosphate synthase

include the identification of structural motifs and the functional annotation of genes responsive to various classes of transcription factors. In our study on transcript levels in potato tuber during development many transcripts strongly correlated with one another (Figure 2). Whilst some of these could be expected such as correlations between TCA cycle proteins (Figures 2A and 2B), the weaker correlation between the different amylolytic activities (Figure 2C) and the negative correlation between ADP-glucose pyrophosphorylase and nitrite reductase (Figure 2D), others were not. Unexpected findings included the strong positive correlation between UMP synthase and a zinc finger protein (Figure 2E) and the negative correlation between enolase and ADPglucose pyrophosphorylase (Figure 2F). In addition to simple pairwise analyses of transcriptome data, substantial research has been carried out in the construction and analysis of gene networks (Davidson *et al.*, 2002). These currently suffer the problem that algorithms used in their generation are often not disclosed. This can create uncertainty in the interpretation of the results. Bioinformatic approaches to ameliorate this problem have recently been documented (Mendes *et al.*, 2003).

Metabolite-metabolite correlation analysis has been pursued in a similar vein to that described above for pair-wise transcript profiles. An initial systematic correlation analysis between all 70 metabolites profiled in our potato tuber systems revealed that, although the majority of metabolite pairs

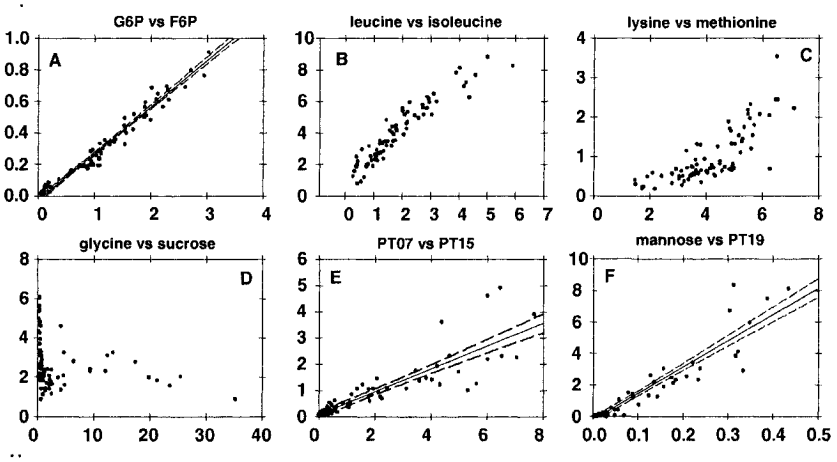


Figure 3. Correlation of metabolite levels in the wild-type and the two transgenic lines of potato tuber during development; G6P – glucose 6-phosphate, F6P – fructose 6-phosphate, PT07, PT15 and PT19 are metabolites of unknown chemical nature

showed little correlation, some were tightly co-regulated (see Figure 3A, 3E, 3F). Others were non-linearly co-regulated, perhaps indicating that the metabolites involved were in some way linked by an enzyme subject to strong regulation (Figures 3B, 3C; Roessner *et al.*, 2001). Yet, further novel insights were found in that some of the transgenic lines under investigation displayed different relationships between metabolites from those seen in the wild type (Figure 3D - although not directly distinguishable in the figure the transgenic lines are those scattered along the x-axis whilst the wild type are close to the origin on this axis). Moreover, the correlation of metabolites of unknown chemical nature may provide hints towards their biochemical synthesis (Figure 3E, F). Similarly co-response analysis, which is essentially an offshoot of metabolic control analysis, have been used both in pattern recognition of the metabolome (Raamsdonk *et al.*, 2001) and in recognition of co-variance of metabolites from the same organism under different conditions (see Kell, 2004 for a detailed review). Following this line of reasoning, Kose et al (2001) developed clique correlation analysis and more recently Steuer *et al.* (2003) related metabolite co-variance matrices to the Jacobian of the system (see Chapter 15). This new approach clearly represents a powerful tool in determining and understanding key regulatory points of the metabolic networks.

4. COMPARISON OF THE TECHNOLOGY PLATFORMS AVAILABLE

In addition to combining data obtained from transcript and metabolite profiling strategies we previously evaluated which of these platforms allowed the highest resolution for the discrimination of different biological systems. For fundamental reasons outlined in the theory of metabolic control analysis, changes in an individual enzyme (or presumably transcript) levels can have little effects on fluxes but can have major effects on metabolite levels (Kell, 2004; Cascante *et al.*, 2002). This would suggest the metabolite level to be the most appropriate for discriminating different biological situations. We wanted to test this experimentally. Thus, we analysed the potato tuber samples discussed above by both transcript and metabolite profiling and evaluated the resultant data via principal component analysis (Figures 4 and 5).

Some of the developmental situations can be readily discriminated from one another on the basis of their transcript profiles; for example, those harvested after 10-weeks of growth are distinct from the other samples. The exact reason for this discrimination however remains unclear. Examination of the clones that were upregulated at this time point suggests a higher metabolic activity at this stage; however this finding is not consistent with changes in the metabolites levels documented in identical samples. Despite the fact that these results were surprising they were consistent in each of the three replicates measured. The PCA also clearly illustrates the surprising result that the transgenic systems could not be discriminated from each other or from the corresponding wild type tuber samples. It is important to note here that a similar situation was also observed after PCA of the entire transcript data set (2200 ESTs) as for the ESTs deemed to give a reliable response (279 ESTs). Although it is apparent that the tuber samples harvested 10 weeks after transfer to the greenhouse were markedly different from those harvested at different developmental stages, the fact remains that transcriptional variation during development is greater than that after a relatively severe genetic perturbation of primary metabolism (Sonnewald *et al.*, 1997; Trethewey *et al.*, 2001).

Metabolic profiling by GC-MS was then carried out on the same samples to determine a wide range of primary metabolites including the nutritionally important lysine, methionine, tocopherol and ascorbate. When PCA was carried out on the data set obtained, a different situation was observed from that seen on analysis of the transcript data (Figures 4 and 5). In this case, the transgenics samples clustered independently of one another, and of the wild type; furthermore on the basis of their metabolic complement, samples of different developmental age could readily be distinguished from one another.

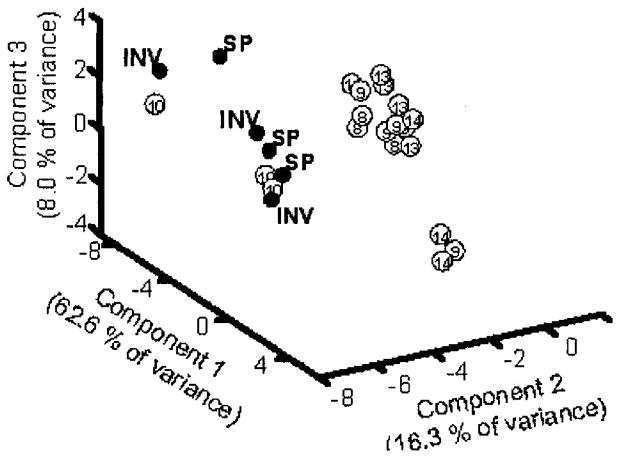


Figure 4. Principal component analysis of transcript levels of genetically and temporally distinct potato tuber systems. The percentage of variance explained by each component is shown in parentheses. The transgenic lines INV2-30 (INV) and SP 29 (SP) are represented by black circles, and wild-type harvest after 8, 9, 10, 13 and 14 weeks after transfer to the greenhouse are represented by open circles

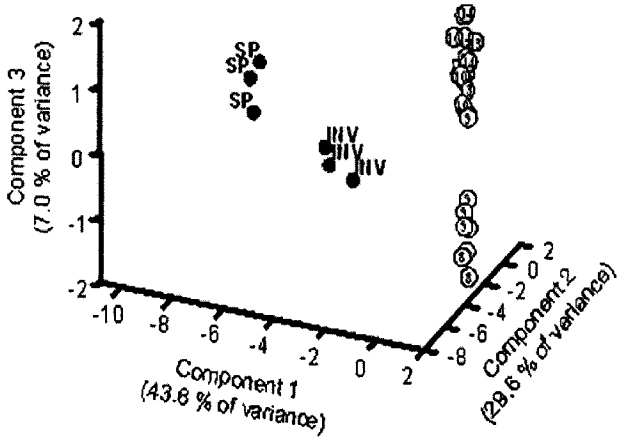


Figure 5. Principal component analysis of metabolite profiles of genetically and temporally distinct potato tuber systems. The percentage of variance explained by each component is shown in parentheses. The transgenic lines INV2-30 (INV) and SP 29 (SP) are represented by black circles, and wild-type harvest after 8, 9, 10, 13 and 14 weeks after transfer to the greenhouse are represented by open circles

As with the results presented above for transcript levels, we believe these data have important ramifications for the potential risks associated with transgenic organisms and theories of substantial equivalence (Kuiper *et al.*, 2001; Trewavas and Leaver, 1999). The conclusion of this work, which to our knowledge is the only direct comparison of the two profiling platforms, is that their discriminatory power is different, with the metabolic profiling allowing greater resolution of the different systems studied here. Whether this implies that changes at the transcript level are indeed less pronounced as compared to those at the metabolite level, or merely highlights limitations of the profiling methods used, remains an open question. It is possible that this result merely reflects low sensitivity of ESTs as probes and that profiling using full-length complementary DNAs (Seki *et al.*, 2001), or oligonucleotide-specific probes (Lo et al, 2003) would allow greater discrimination. Given this fact it would be interesting to see future comparisons as the technologies further evolve. However, whatever the reason for the differential discriminatory power, these results alongside previous experimental work comparing transcript and protein levels from identical samples (Futcher *et al.*, 1999; Gygi *et al.*, 1999; Ideker *et al.*, 2001), and the demonstration that statistical evaluation of the combined information yielded by metabolic and proteomic studies on *Arabidopsis* ecotypes revealed a high ability to discriminate distinct ecotypes from one another (Weckwerth, 2003), strongly suggest that the discrimination of biological systems should be performed at more than one level. As mentioned above this finding has resonance with respect to studies aimed at establishing substantial equivalence between transgenic and conventional crops.

5. INSIGHTS GAINED BY BIOINFORMATICS ON COMBINED DATA SETS

As stated above, our studies were initiated with two main objectives. Firstly, we wanted to compare the discriminatory power of both profiling approaches, and secondly, we were interested in using the combined evaluation of both analyses as a new approach in experimental systems biology (for a definition see Sweetlove *et al.*, 2003).

As a first step, we decided to run all data points through pairwise correlation analysis, like those described above, determining for each transcript whether it is correlated with any of the metabolites. Of the 26,616 pairs analysed, 363 positive and 208 negative correlations were detected, the total number of 571 correlations being well above that which might be expected by chance (266 at $P < 0.01$). Several, representative examples of

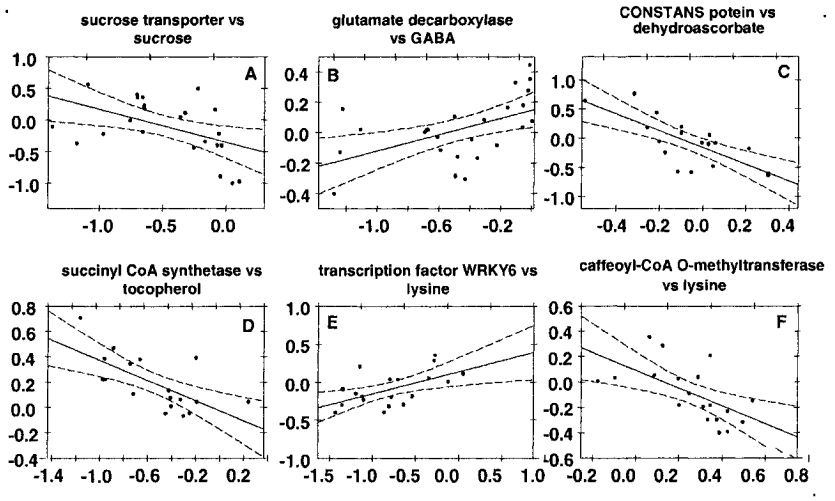


Figure 6. Correlations between metabolites and transcript levels in potato tuber during development. GABA - γ-aminobutyric acid.

transcript:-metabolite correlations are shown in Figure 6. As a primary evaluation we determined whether the data we obtained were in agreement with those made using different experimental strategies. This is clearly the case with a strong negative correlation between sucrose and sucrose transporter expression (Figure 6A) and a strong positive correlation between 4-aminobutyric acid and glutamate decarboxylase isoform I (Figure 6B). Since both these relationships have previously been reported in the literature (Vaughn *et al.*, 2002; Facchini *et al.*, 2000) these examples provide validitation of our approach. Secondly, many further correlations seem to have a functional basis that can be retrospectively explained. The positive correlations of both tryptophan and tyrosine with the β2chain of tryptophan synthase and ornithine carbamoyltransferase with serine and cysteine (Urbanczyk-Wochniak *et al.*, 2003) are two such examples. Thirdly, although several of the correlations, such as those described above, were predictable the majority of the correlations obtained following this approach were novel and not directly related to the biochemical pathway in which the gene products participate in. Whilst it is clear that many of these correlations are due to chance, several were observed between transcripts and metabolites of the same or related pathways a fact that may strengthen interpretation of the linkages. Examples of such instances include aminotransferase which correlates with both fructose-6-phosphate and glucose-6-phosphate (Urbanczyk-Wochniak *et al.*, 2003). Such comparisons may offer hints as to the function of the genes involved. It is also interesting

to note that several transcripts correlate with more than one metabolite such as highlighted above for the aminotransferase, other examples of this include glutamate decarboxylase isoform I which correlates both with spermidine and tyrosine (Urbanczyk-Wochniak *et al.*, 2003), whilst the same holds also true for 4-aminobutyric acid and tryptophan and various transcription factors. Finally it is exciting to see that nutritionally important metabolites such as ascorbate, tocopherol and lysine were tightly correlated to the expression levels of various genes or transcription factors: ascorbate being negatively correlated with a homologue of the clock gene CONSTANS (Figure 6C), tocopherol negatively correlated with succinyl CoA synthetase (Figure 6D) and lysine being positively regulated by transcription factor WRKY6 (Figure 6E), lysine being negatively regulated by caffeoyl-CoA O-methyltransferase (Figure 6F).

We believe these essentially unexpected correlations to be of great potential for biotechnological applications where the goal is the modification of metabolite compositions through genetic means. Whilst these data do not, in any way, prove causal or even mechanistic links between the different molecular entities, the approach of linking transcript and metabolite data via pair-wise correlation analysis presents a very powerful tool for the rapid identification of 'candidate genes', the function of which can be tested via further experimentation. Whilst initial results from these results look highly promising it will be some time before enough evidence is accumulated to be able to estimate the efficacy of this approach in plants. That said preliminary indications appear promising since the rapidity of genetic manipulation of microbial systems has already facilitated the identification of genetic factors underlying the production of commercially important pharmaceuticals (Azkenazi *et al.*, 2003).

6. SUMMARY AND FUTURE PROSPECTS

In analogy to gene and metabolite network analysis combined analyses across molecular entities can be carried out in attempt to define gene regulatory networks (Figure 7). This is a relatively novel and highly complicated type of analysis not least since metabolic regulation is often non-linear (Mendes and Kell, 1996, Barabasi and Oltvai, 2004)). That said analysis of co-regulation (regardless of the molecular entity under study) will likely be of great importance in further understanding the complex regulatory circuitry of metabolism.

The main conclusions of this discussion of the utility of combined metabolite and transcript profiling approaches is that it offers two major advantages over a single platform. Firstly, the results suggest that despite the

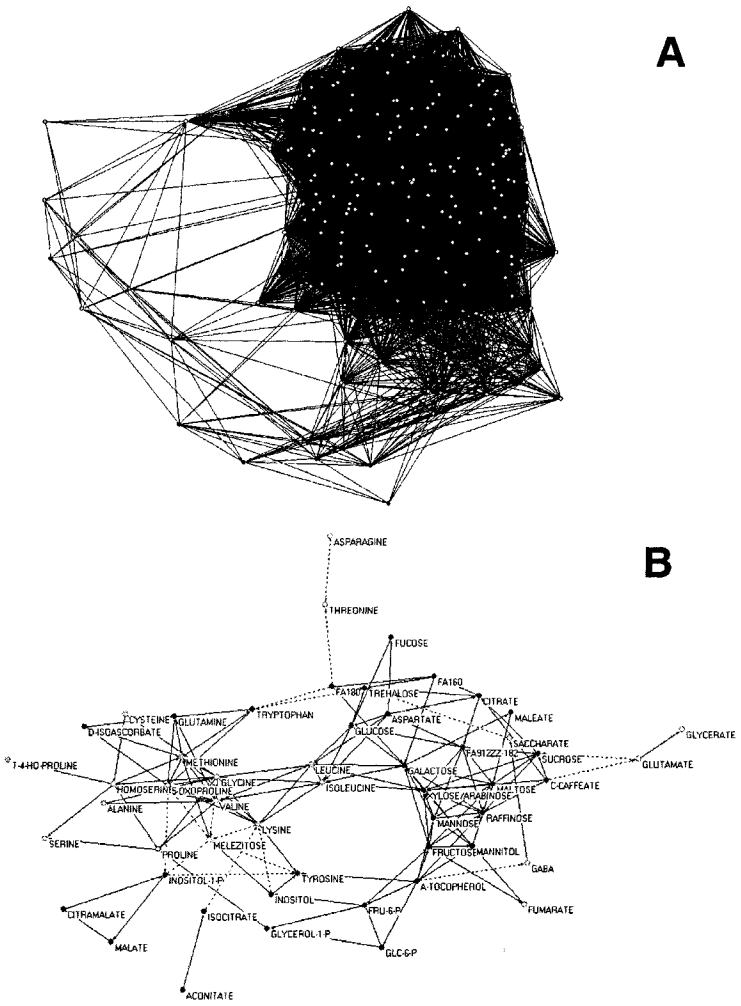


Figure 7. Metabolic correlation network for potato tuber. Pairwise correlations were calculated for every metabolite and transcript. If their correlation exceeded a given threshold ($p<0.01$), the two entities are connected with a link (Negative correlation being indicated by detached lines). A) complete transcript metabolite network of potato tuber development; B) Selected metabolite network.

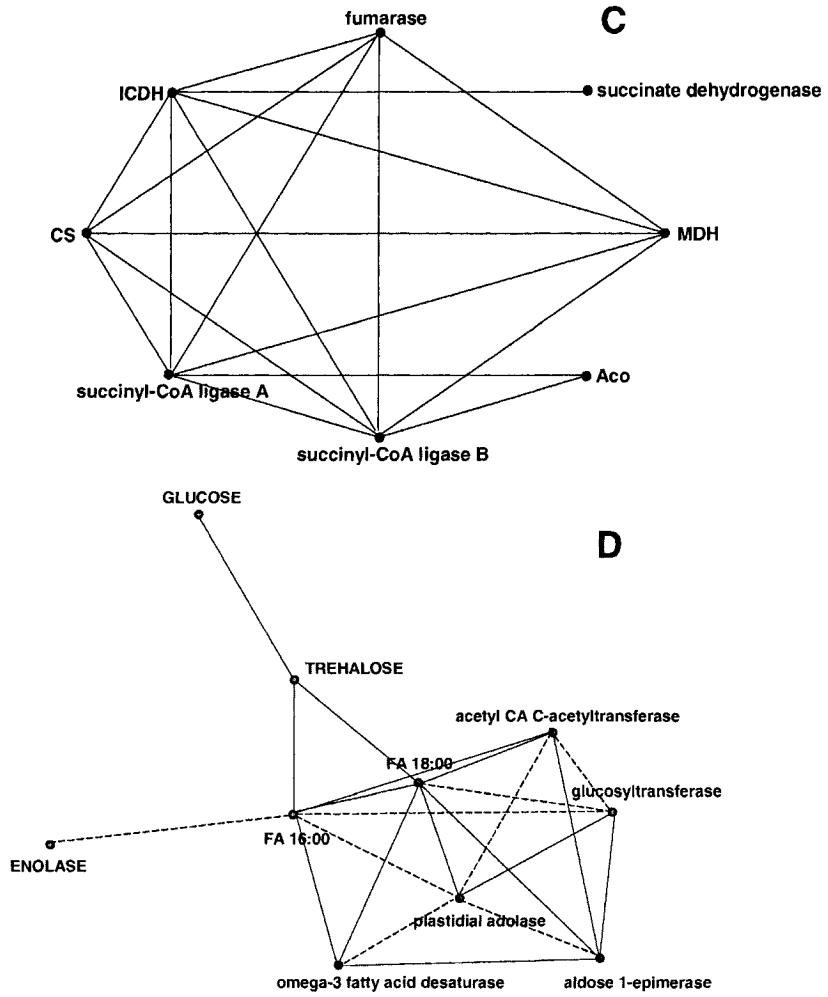


Figure7 (continued) C) Correlations between genes encoding enzymes of the TCA cycle; D) zoom in of A illustrating examples of correlations between metabolites and genes.

fact that metabolite profiling offered superior resolution of the distinct biological situations described here the discrimination of biological systems is best performed at more than one level. The second main conclusion of our work is that although the number of metabolite: transcript correlations was relatively small, as would be expected from previous studies comparing transcript and protein levels (Futcher *et al.*, 1999; Gygi *et al.*, 1999; Ideker *et al.*, 2001) and mathematical studies (ter Kuile and Westerhof, 2001), we

believe that they allow the generation of clearly testable hypotheses. Of particular interest in our data set was the correlation of genes with the essential amino acid lysine and with the vitamins ascorbate and tocopherol as these linkages define candidate genes for the manipulation of these nutritionally important compounds in plants. To our knowledge only the combination of the two platforms presented here is able to provide such novel hypothesis; therefore, the pairwise correlation between transcripts and metabolites is a potentially powerful new tool for hypothesis creation in systems biology.

ACKNOWLEDGEMENTS

The help of Dr. Victoria Nikiforova (same institute) in the network analyses and Drs. Joachim Kopka and Joachim Selbig (same institute) for collaboration in bioinformatic analysis described here is gratefully acknowledged.

REFERENCES

Allocco DJ, Kohane IS; Butte AJ. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics*, 5: 18-28 (2004).

Azkenazi M *et al.* Integrating transcriptional and metabolite profiles to direct the engineering of lovastatin-producing fungal strains. *Nat. Biotechnol.*, 21: 150-156 (2003).

Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat. Rev. Genetics.*, 5: 101-113 (2004).

Boros LG, Cascante M, Lee WNP. Stable isotope-based dynamic metabolic profiling in disease and health. In *Metabolic profiling: its role in biomarker discovery and gene function analysis* (Harrigan GG, Goodacre R, eds.). Kluwer Academic Publishers, Boston (2003).

Cascante M, Boros LG, Comin-Anduix B, de Atauri P, Centelles JJ, Lee WNP. Metabolic control analysis in drug discovery and disease. *Nat. Biotechnol.*, 20: 243-249 (2002).

Davidson EH *et al.* A provisional regulatory gene network for specification of endomesoderm in the sea urchin embryo. *Dev. Biol.*, 246: 162-190 (2002).

Facchini PJ, Huber-Allanach KL, Tari LW. Plant aromatic L-amino acid decarboxylases: evolution, biochemistry, regulation and metabolic engineering applications. *Phytochemistry,* 54: 121-138 (2000).

Fernie AR, Willmitzer L, Trethewey RN. Sucrose to starch: a transition in molecular plant physiology. *Trends Plant Sc.*, 7: 35-42 (2002).

Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN, Willmitzer L. Metabolite profiling for plant functional genomics. *Nat. Biotechnol.*, 18: 1157-1161 (2000).

Futcher B, Latter GI, Monardo P, McLaughlin CS, Garrells JI. A sampling of the yeast proteome. *Mol. Cell Biol.*, 19: 7357-7368 (1999).

Goossens A, Hakkinen ST, Laakso I, Seppanen-Laakso T, Biondi S, De Sutter V, Lammertyn F, Nuutila AM, Soderlund H, Zabeau M, Inze D, Oksman-Caldentey KM. A functional genomics approach toward the understanding of secondary metabolism in plant cells. *Proc. Natl. Acad. Sci. USA*, 100: 8595-8600 (2003).

Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell Biol.*, 19: 1720-1730 (1999).

Harrigan GG, Goodacre R *Metabolic profiling: its role in biomarker discovery and gene function analysis*, Kluwer Academic Publishers, Boston (2003).

Kitano H. Systems biology: A brief overview. *Science*, 295: 1662-1664 (2002).

Kell DB. Metabolomics and systems biology: making sense of the soup. *Curr Opin Microbiol.*, 7: 296-307 (2004).

Kell DB, Oliver SG. Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays*, 26: 99-105 (2004).

Kose F, Weckwerth W, Linke T, Fiehn O. Visualising plant metabolomic correlation networks using clique-metabolic matrices. *Bioinformatics*, 17: 1198-1208 (2001).

Krömer JO, Sorgenfrei O, Klopprogge K, Heinzle E, Wittmann C. In-depth profiling of lysine-producing Cornebacterium glutamicum by combined analysis of the transcriptome, metabolome and fluxome. *J. Bacteriol.*, 186: 1769-1784 (2004).

Kuiper HA, Kleter GA, Noteborn HPJM, Kok EJ. Assessment of the food safety issues related to genetically modified foods. *Plant J.*, 27: 503-528 (2001).

Lo HS, Wang ZN, Hu Y, Yang HH, Gere S, Buetow KH, Lee MP. Allelic variation in gene expression is common in the human genome. *Genome Research*, 13: 1855-1862 (2003).

Mendes P, Kell DB. Non-linear optimisation of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics*, 14: 869-883 (1998).

Mendes P, Sha W, Ye K. Artificial gene networks for objective comparison of analysis algoriths. *Bioinformatics*, 19: II122-II129 (2003).

Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, 292: 929-934 (2001a).

Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. *Ann. Rev. Genom. Hum. Genetics*, 2: 343-372 (2001b).

Oltvai ZN, Barabasi AL. Systems biology: Life's complexity pyramid. *Science*, 295: 1662-1664 (2002).

Raamsdonk LM, Teusink B, Broadhurst D, Zhang NS, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ, Westerhoff HV, van Dam K, Oliver SG. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.*, 19: 45-50 (2001).

Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, Willmitzer L, Fernie AR. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell*, 13: 11-29 (2001).

Sabati C, Rohlin L, Oh MK, Liao JC. Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acid Res.*, 30: 2886-2893 (2002).

Seki M *et al.* Arabidopsis encyclopedia using full-length cDNAs and its application. *Plant Physiol. Biochem.*, 39: 211-220 (2001).

Shevchenko A, Jensen ON, Podtelejnikov AV, Sagliocco F, Wilm M, Vorn O, Mortenson P, Shevchenko A, Boucherie H, Mann M. Linking genome and proteome by mass spectrometry: large scale identification of yeast proteins from two-dimensional gels. *Proc. Natl. Acad. Sci., USA*, 93: 14440-14445 (1996).

Simchen G. Cell-cycle mutants. *Ann. Rev. Genetics*, 12: 161-191 (1978).

Steuer R, Kurths J, Fiehn O, Weckwerth W. Observing and interpreting correlations in metabolomic networks. *Bioinformatics*, 19: 1019-1026 (2003).

Stitt M, Fernie AR. From measurement of metabolites to metabolomics: an "on the fly" perspective illustrated by recent studies of carbon-nitrogen interactions. *Curr. Opin. Biotechnol.*, 14: 136-144 (2003).

Stitt M, Sonnewald U. Regulation of metabolism in transgenic plants. *Ann. Rev. Plant Physiol. Plant Mol. Biol.*, 46: 341-381 (1995).

Sonnewald U, Hajirezaei MR, Kossmann J, Heyer AG, Trethewey RN, Willmitzer L. Increased potato tuber size results from apoplastic expression of yeast invertase. *Nat. Biotechnol.*, 15: 794-797 (1997).

Sweetlove LJ, Last RL, Fernie AR. Predictive metabolic engineering: a goal for systems biology. *Plant Physiol.*, 132: 420-425 (2003).

Trethewey RN, Fernie AR, Bachmann A, Fleisher-Notter H, Geigenberger P, Willmitzer L. Expression of a bacterial sucrose phosphorylase in potato tubers results in a glucose independent induction of glycolysis. *Plant Cell Environ.*, 24: 357-365 (2001).

Trewavas A, Leaver CJ. Conventional crops are the test of GM prejudice. *Nature*, 401: 640 (1999).

Urbanczyk-Wockniak E, Luedemann A, Kopka J, Selbig J, Roessner-Tunali U, Willmitzer L, Fernie AR. Parallel anaylsis of transcript and metabolite profiles: a new approach for systems biology. *EMBO reports*, 4: 989-993 (2003).

Vaughan MW, Harrington GN, Bush DR. Sucrose-mediated transcriptional regulation of sucrose symporter activity in the phloem. *Proc. Natl. Acad. Sci. USA*, 99: 10876-10880 (2002).

Weckwerth W. Metabolomics in systems biology. *Ann. Rev. Plant Biol.*, 54: 669-689 (2003).

Yamanishi Y, Vert JP, Nakaya A, Kanehisa M. Extraction of correlated gene clusters from multiple genomic data by generalized kernal canonical correlation analysis. *Bioinformatics*, 19: 323-330 (2003).

Chapter 17

FLUXOME PROFILING IN MICROBES

Nicola Zamboni and Uwe Sauer
Institute of Biotechnology, Swiss Federal Institute of Technology (ETH) Zürich, 8093 Zürich, Switzerland

1. INTRODUCTION

Similar to the readout from transcriptome and proteome analyses, comprehensive detection of intracellular metabolite levels, the metabolome, assesses metabolic network *composition*. In sharp contrast to such concentration measurements, quantification of fluxes between metabolites is a very different type of analysis that provides time-dependent information on metabolic network *operation*. Flux responses to environmental perturbations in the range of seconds or minutes are largely driven by kinetic regulation, which, in turn, is a function of metabolite concentrations and kinetic properties of the involved enzymes. On longer time-scales - in genetic variants or during growth in a given environment – flux responses result from both genetic and metabolic regulation. Quantification of molecular fluxes thus provides a direct link to the interactions between genes, proteins, and metabolites, where the actually observed flux distribution reflects the integration of all occurring regulation processes in the network (Bailey 1999; Hellerstein 2003).

Analogous to other 'omics approaches, the term fluxome was coined to define the metabolism-wide array of reaction velocities in an organism (Sauer *et al.,* 1999). Different from concentration measurements, however, molecular fluxes within a metabolic network are only indirectly accessible through extracellular metabolites that enter or exit the cell. These time-dependent concentration changes are then balanced within a stoichiometric model of the reaction network (Varma and Palsson, 1994). When applied to

realistic networks, such simple flux balancing has limited analytical potential because alternative pathways can catalyze a given conversion. This is particularly true for central metabolisms with multiple interconnected pathways. Biosynthetic reactions that yield cellular components, in contrast, develop linearly with very few redundancies. Thus, peripheral fluxes can be assessed accurately with detailed models of the macromolecular biomass composition that quantify withdrawal of precursors (Pramanik and Keasling, 1997; Dauner and Sauer, 2001). The major challenge in fluxome analysis is therefore largely restricted to quantification of fluxes through the 100-150 reactions of central metabolisms that flexibly catalyze the bulk of the carbon flow.

A breakthrough was achieved with ^{13}C-labeling experiments (Marx *et al.,* 1996; Sauer *et al.,* 1997), because pathway-specific conversion of substrates imprint characteristic ^{13}C- patterns in reaction intermediates or products thereof. This biochemical principle was extended from one or few pathways/reactions to the entire network (Wiechert, 2001; Sauer, 2004). In a typical experiment, microbes are grown in (quasi) steady-state using minimal media with a single ^{13}C-labeled substrate. After a few generations, biomass samples are collected and the labeling pattern of proteinogenic amino acids is analyzed by NMR or MS. Three basic approaches for ^{13}C-labeling pattern interpretation can be defined: integrated, analytical, and comparative (Sauer, 2004) (Figure 1). The integrative approach extends metabolite balancing with isotopomer balancing so that ^{13}C-data, extracellular material fluxes, and biomass composition are simultaneously interpreted within metabolic models of various complexity (Dauner *et al.,* 2001; Wiechert, 2001). To identify the flux distribution, the labeling state of all metabolic intermediates is balanced in an iterative fitting procedure. Since all available data are used, the integrative approach provides the greatest detail and has been used successfully with both NMR and MS data (Wiechert 2001; Sauer 2004).

Despite the attained analytical precision and the recognized value of the data, only 100-200 integrated flux analyses have been reported by the two groups in the field - primarily for two reasons. First, the necessity for high-quality physiological measurements typically requires tedious continuous cultures. Second, identifying the best-fit solution of unknown fluxes from the available data by iterative, numerical simulations is a mathematical/statistical challenge that requires multiple supervised runs, involving time and expertise. Since all available information is processed, imperfections such as measurement errors or incomplete networks propagate throughout the model and affect the entire flux solution, often leading to a statistical rejection of the model (van Winden *et al.,* 2001; Wiechert, 2001).

When this occurs, troubleshooting is tedious and user expertise is necessary to localize the problem.

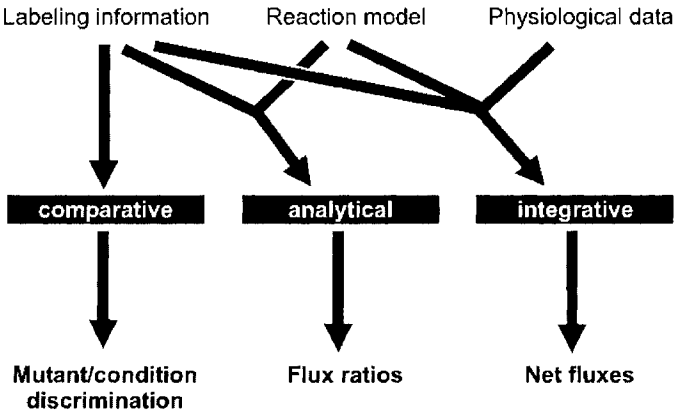


Figure 1. Roadmap for different types of fluxome analysis.

In contrast to integrative analysis, analytical or comparative interpretation of ^{13}C data is not based on balancing of metabolites or their labeling state. While it does not deliver absolute reaction velocities, it offers other important advantages (Szyperski, 1998; Sauer, 2004):

- physiological data are not necessary
- the computation is straightforward and rapid
- the analysis is local in nature, hence a particular measurement is largely independent of possible errors elsewhere in the network
- direct fluxome insight is gained because essential information is filtered from the large dataset of all labeling data.

Here we highlight key features that predispose analytical and comparative approaches for large-scale fluxome mapping in functional genomics and pharmaceutical research. The term *profiling* is used rather than flux analysis to indicate that these approaches do not attempt to quantify all fluxes.

2. ANALYTICAL FLUXOME PROFILING: METABOLIC FLUX RATIO ANALYSIS

Direct analytical interpretation of ^{13}C-labeling patterns with algebraic or probabilistic equations can quantify flux partitioning ratios of converging

pathways/reactions in microbes (Szyperski, 1995; Christensen *et al.,* 2001) or higher cells (Kelleher, 2001; Hellerstein, 2003; Sherry *et al.,* 2004). In contrast to integrated global balancing of all metabolite and isotopomer species, flux ratios are estimated locally from the labeling pattern of selected compounds. Incomplete networks, poor quality of some data, or fluxes that cannot be identified from the available data therefore affect only some, but not all quantified ratios (Szyperski, 1998; Sauer, 2004). A particularly useful approach to interrogate microbial metabolism is metabolic flux ratio analysis, first described for NMR data (Szyperski, 1995; Sauer *et al.,* 1997). The recent extension to sensitive and rapid MS analysis also has great potential for high-throughput studies in microscale cultures (Fischer and Sauer, 2003a; Fischer *et al.,* 2004). This state-of-the-art approach quantifies more than 10 independent ratios of key fluxes through converging pathways and reactions in bacterial or yeast metabolism by GC-MS analysis of proteinogenic amino acids from cells grown on ^{13}C-labeled glucose (Fischer and Sauer, 2003a; Blank and Sauer, 2004). The scientific potential of flux ratio analysis may be illustrated by recent discoveries: (i) a novel glucose catabolic pathway in *Escherichia coli* (Fischer and Sauer, 2003b); (ii) the reverse, anaplerotic function of a normally gluconeogenic enzyme in *Bacillus subtilis* (Zamboni *et al.,* 2004), (iii) experimental identification of metabolic network topology in poorly characterized bacterial species (Fuhrer *et al.*, unpublished); and (iv) an unexpected regulation of the Krebs cycle in *Saccharomyces cerevisiae* (Blank and Sauer, 2004).

Flux partitioning ratios *per se* do not provide absolute flux values, but may be used as constraints for their estimation. Combined with metabolite balances, they allow the quantification of absolute fluxes within a stoichiometric model. Such ^{13}C-constrained flux balancing was demonstrated with both NMR- (Sauer *et al.,* 1997) and MS-derived (Zamboni and Sauer, 2003; Fischer *et al.,* 2004) flux ratios. Although flux solutions obtained with comprehensive isotopomer balancing are of higher quality and greater detail (e. g. exchange fluxes in reversible reactions), ^{13}C-constrained flux balancing can yield statistically analogous results (Fischer *et al.,* 2004). Different from the integrative approach, this conceptually simpler flux method can be fully automated and requires negligible computation times. Since almost identical flux distributions were obtained with different methods in different batch cultivation systems (Fischer *et al.,* 2004), ^{13}C-constrained flux balancing from microtiter plate data appears to be a good compromise between accuracy and throughput for large-scale studies.

3. MODEL-INDEPENDENT COMPARATIVE FLUXOME PROFILING

The required mathematical frameworks remain a principal bottleneck because *a priori* network information is necessary to integrated flux and isotopomer balances or to derive probabilistic equations for analytical flux ratio analysis. Hence, experiments are generally done in synthetic media with one or two carbon sources to ensure model validity and to obtain precise readouts. While isotopomer balancing is feasible in rich media (Christiansen *et al.*, 2002), precise quantification of production and consumption of all carbon species render it rather tedious.

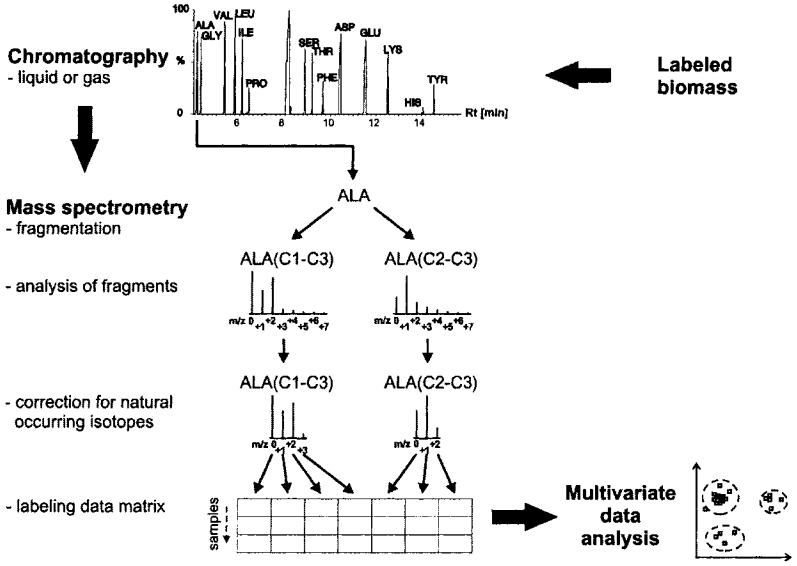


Figure 2. Schematic flow chart for multivariate statistical analysis of labeling data in comparative fluxome profiling. An example is given for the mass isotope distribution of two different alanine fragments that consist of a C3 and a C2 unit.

To overcome this fundamental limitation, we developed a novel concept to discriminate mutants or conditions by direct comparison of ^{13}C-patterns. In contrast to model-based approaches, pattern comparison does not deliver numerical values for fluxes or flux ratios, but aims to recognize discriminant information in the labeling patterns by statistical learning (Figure 2). Unsupervised methods reveal relevant features such as single outliers or conserved labeling pattern in redundant fragments that exhibit a high correlation. Such features are searched in the two-dimensional landscape of

all samples and mass distributions. This approach is henceforth referred to as *comparative fluxome profiling*.

Experiments may be done in any cultivation form, and labeled biomass or metabolites are best analyzed with MS. The raw mass distributions are then corrected for naturally occurring isotopes. For each sample, the mass distributions of all species, or fragments thereof, detected by MS are sequentially collected in the column of a table (Figure 2). Correction is not strictly necessary but reduces the data dimension, thus simplifying analysis and facilitating interpretation because isotopic effects that are not linked to the chosen label are filtered out. Unidentified peaks that may appear in chromatograms are ignored in model-based approaches, but may be included in the pattern comparison, in this case without correction for natural abundance. Statistical analysis of a comprehensive dataset then reveals unknown species that might exhibit relevant features, and may facilitate their identification by revealing pattern correlations within known metabolites.

3.1 Experimental proof-of-concept

A proof-of-principle for fluxome profiling by multivariate data analysis was obtained with GC-MS-analyzed proteinogenic amino acids from 12 *B. subtilis* mutants that were grown on ^{13}C- and ^{2}H-labeled substrates. Firstly, principal component analysis (PCA) (Jolliffe 2002) was applied, which projects the input variables in a space spanned by orthogonal principal components that are sequentially selected to maximize the variance of the projected data. As expected from the analogy between fluxome and metabolome profiling (Fiehn *et al.*, 2000; Allen *et al.*, 2003), PCA successfully discriminated mutant phenotypes. In contrast to metabolome profiling, however, the identified principal flux components were complex combinations of several input variables across the entire dataset. Hence, it was not possible to correlate the pattern to specific metabolic effects.

Hidden information in the labeling patterns was revealed when the corrected ^{13}C data were subjected to independent component analysis (ICA) (Hyvärinen *et al.*, 2001). Akin to PCA, a new multidimensional basis for the input variables space is defined by independent components. ICA identifies components that are statistically as independent as possible by selecting those with maximum non-Gaussian distribution. The resulting components are not only linearly independent, such as in PCA, but also possess minimal nonlinear correlations. When applied to labeling data from proteinogenic amino acids, ICA was able to identify components in the input variables that were dominated by either single or few related amino acids. The independent components could therefore be linked to specific shifts in the labeling pattern of metabolites. Specifically, ICA provided two types of information:

(i) it automatically identified signatures of independent metabolic responses that allowed the classification of samples, and (ii) it often grouped redundant signals in different amino acids within the same component, thus providing insights on the biochemical relation of species or fragments.

3.2 The comparative approach opens new dimensions in fluxome profiling: application to complex media and ^{2}H-tracers

Comparative interpretation of isotopic tracer information is data-driven and model-independent. Beside the obvious advantage for uncharacterized organisms, two unique features of comparative fluxome profiling pave the road to applications beyond microbes that grow in minimal media. Firstly, labeling experiments are feasible in complex media or in the presence of multiple labeled substrates. Our observations reveal that even when isotope patterns are recorded in the proteinogenic amino acids, sufficient information is available from *de novo* amino acid synthesis. This enables analysis of auxotrophic mutants from genomic libraries or organisms with complex nutrient requirements. Ultimately, the use of free metabolites would be desirable to increase the information content. The second important innovation is the applicability to any stable isotope, thus ^{18}O, ^{15}N, or ^{2}H may be used alone or in combination with ^{13}C. Profiling of hydrogen metabolism is particularly attractive due to the potential to monitor macromolecule turnover (McCabe and Previs, 2004), water release, or reactions that do not affect the carbon pattern, e. g. dehydrogenases (Siler *et al.*, 1999).

The responses in a dozen metabolic or regulatory *B. subtilis* knockout mutants during growth on fully deuterated [U-^{2}H]glucose, that we observed (*vide infra*), serve as an example of hydrogen fluxome profiling. To visualize mutant responses qualitatively, we normalized the mutant mass isotope distributions by subtracting the parental signals (Figure 3). Metabolic differences between strains are then reflected by the deviation from the null line. Figure 3A shows that the *sdhC* mutant with a disrupted Krebs cycle exhibits qualitatively similar responses in carbon and hydrogen metabolism. Aspartate and glutamate are the only obvious outliers due to the loss of ^{2}H during sample preparation. In other mutants ^{2}H-patterns differ from ^{13}C-pattern, e. g. in the glycolytic repressor mutant *cggR* (Figure 3B). Signatures of valine, leucine, and partially alanine in the *cggR* mutant revealed a double loss of ^{2}H in the pyruvate precursor. This agrees favorably with the derepression of the glycolytic enolase that promotes the reversible exchange of protons with water in the *cggR* mutant (Ludwig *et al.*, 2001). Thus, the

fingerprints can principally be mapped to their metabolic determinants and thereby reveal the underlying biochemical causality.

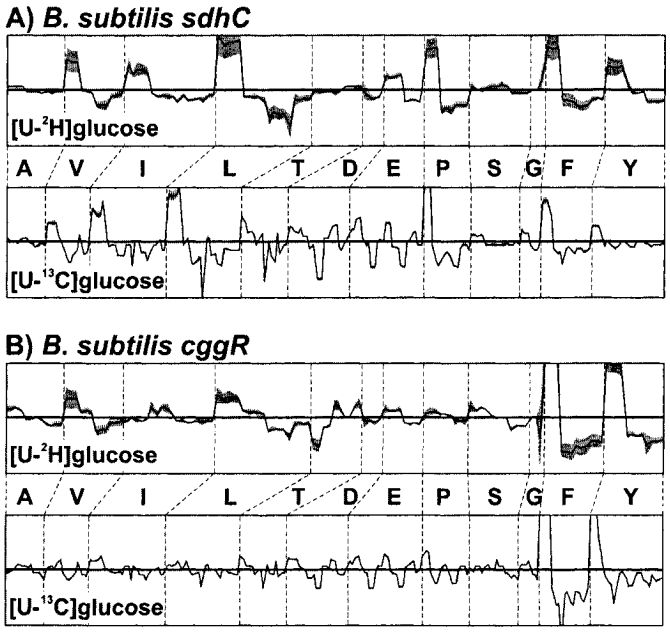


Figure 3. Comparison of wild-type-normalized labeling profiles in amino acids obtained from [U-^{2}H]glucose and [U-^{13}C]glucose experiments with two *B. subtilis* knockout mutants The line deviates above the null line when an amino acid (represented by their one-letter code) mass is more abundant in the mutant than in the parent, and vice versa. Within each amino acid, the available data points are in the order of their total mass, with the m0 at the left end. Shaded areas represent deviations between independent experiments.

3.3 Unsupervised versus supervised learning methods

For mutant/condition discrimination, comparative fluxome profiling by multivariate statistics from mass isotope distributions is feasible with unsupervised statistical learning methods such as PCA or ICA, but other component decomposition methods such as factor analysis (Jolliffe 2002) and independent factor analysis (Attias 1999)) may also be used. Additionally, we tested hierarchical cluster analysis and self-organizing maps (unpublished) as unsupervised classification methods. Although both recognized and grouped mutants with radical and distributed changes in their labeling pattern, the classification was error prone and failed to cluster mutants with less pronounced but statistically significant labeling effects.

In contrast to traditional, analytical or integrated flux analysis, comparative fluxome profiling by unsupervised statistical learning methods does not provide numerical values of flux ratios or net fluxes. In principle, this shortcoming may be overcome with supervised learning methods. Relevant correlations may be identified by training with datasets that contain both the input variables and the corresponding, expected outcome; either as a class (the inactivity of a particular pathway) or a scalar/vector (e.g. one or more split ratios). The prediction rate of the trained method must then be validated with test datasets for which the outcome is also known. In contrast to non-targeted analyses, such as PCA and ICA, supervised training promises higher resolution (Buckhaults *et al.,* 2003; Iizuka *et al.,* 2003) and quantitative estimates (Svetnik *et al.,* 2003).

Several classification algorithms were developed in statistics and machine learning to meet variegate requirements. Methods such as linear and quadratic discriminant analysis, support vector machines, *k*-nearest neighbor classifiers, bagging and bootstrapping trees (see for example Hastie *et al.,* 2001) appear to be compatible with labeling data. Among those, discriminant analysis aims to identify variables in *discriminant functions* that maximally discriminate between two or more groups, which are defined by the supervisor. For fluxome profiling, discriminant analysis is a prime candidate to cluster labeling patterns. In fact, both linear and quadratic discriminant analysis were already applied in metabolome profiling to discern mutants or physiological conditions (Raamsdonk *et al.,* 2001; Allen *et al.,* 2003), or to classify cancer cells from gene expression data (Nguyen and Rocke, 2002a). Notably, organism-wide scale problems generally exhibit more dimensions (variables) than samples, hence, these classification algorithms were applied on top of dimension reduction methods, usually PCA or partial least squares (Geladi and Kowalski, 1986; Höskuldsson, 1988; Nguyen and Rocke, 2002b). The potential of discriminant analysis for labeling data offers the opportunity to separate subpopulations based on their metabolic activity and to express causal connectivity between metabolites in the resulting discriminant functions. While fluxes cannot be derived from metabolite concentrations (nor from transcript levels), comparative fluxome profiling offers more direct access to complex flux traits, e.g. the activity of multiple enzyme pathways or activity alterations based on covalent modifications or allosteric regulators.

4. ANALYTICAL CHALLENGES

Although comparative fluxome profiling could, in principle, be done with NMR, only MS methods provide the sensitivity, low cost, and short analysis

times that are required for higher throughput. While it is already feasible with protein-bound amino acids (Sauer 2004), the full potential can only be exploited when labeling patterns are detected in the free metabolites. Technical challenges from high metabolite turnover rates and their low concentrations are then similar to those in metabolomics: (i) rapid sampling to quench ongoing metabolic activities (Schaefer *et al.,* 1999; Buziol *et al.,* 2002; Visser *et al.,* 2002), (ii) analytical sensitivity and robustness (Tolstikov and Fiehn, 2002; van Dam *et al.,* 2002; Soga *et al.,* 2003), and (iii) efficient extraction (Castrillo *et al.,* 2003; Maharjan and Ferenci, 2003). The latter is probably less critical for fluxome analysis because determination of labeling patterns does not rely on complete metabolite extraction. Measurement problems such as matrix-dependent ion suppression (Choi *et al.,* 2001) are also not a major issue, provided sufficient ions of interest are detected within convenient integration times. Hence, internal standards are not required to account for such effects.

Instead, two MS-related issues assume a greater importance. First, for each ion to be analyzed, *m/z* values between 10 and 15 must be detected to quantify the abundance of heavier mass isotopes from the tracer molecules. Fastidious overlaps between isotopomers of different ions are more likely to occur than in metabolite concentration analyses that focus on few values. Clean chromatographic fractionation of the analytes is necessary to minimize co-elution and mass superpositions, but increases measurement time. Second, ion fragmentation patterns are important. MS detects ion mass fractions that contain a given number of labeled atoms (m_0, m_{+1}, m_{+2} etc), but cannot directly identify the position of labeled atoms such as NMR (Szyperski, 1998). Positional information may be obtained, however, by comparing the mass distributions of parent and fragment ions or fragment pairs of the same parent ions (Figure 4). Two basic types of fragmentation can be distinguished, *in-source* and *post-source*.

In-source fragmentation occurs upstream of the (first) mass analyzer by the ionization and focusing steps where molecules are subjected to strong electric fields, high temperatures, or collisions with electrons or gas molecules (Cole, 1997). With small molecules, the phenomenon is typical for strong ionization methods, such as electron impact in GC-MS systems. For fluxome analysis with derivatized amino acids, GC-MS-based in-source fragmentation provided key data (Dauner and Sauer, 2000; Christensen *et al.,* 2002; Fischer and Sauer, 2003a). Because only 10-15 amino acids were analyzed, mass distributions could be quantified individually upon baseline separation by GC. The analysis of complex metabolite mixtures, in contrast, leads to co-elution of analytes (Fiehn *et al.,* 2000; Soga *et al.,* 2003; von Roepenack-Lahaye *et al.,* 2004) and in-source fragmentation complicates precursor ion identification and quantification because multiple signals are

generated for single compounds. To prevent premature fragmentation, most metabolome studies rely on mild ionization sources, typically electrospray ionization (ESI) (Fenn *et al.*, 1989).

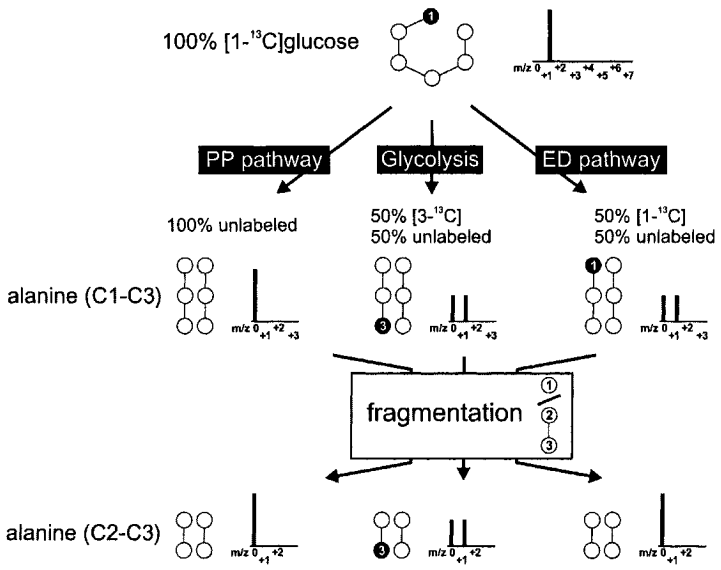


Figure 4. Illustration of positional labeling information that may be obtained from fragmented metabolites with MS. In *E. coli*, catabolism of [1-^{13}C]glucose may occur via three routes: glycolysis, the pentose phosphate (PP) or the Entner-Doudoroff (ED) pathways. Although they produce unique labeling pattern in alanine, MS analysis of intact alanine cannot discriminate between glycolysis and the ED pathway. However, their activity may be resolved by the mass distribution of, for example, the C2-C3 moiety of alanine. Naturally occurring stable isotopes were not considered for simplicity.

In contrast to in-source fragmentation that depends on the ionization source, induced post-source fragmentation can actually facilitate ion identification. Modern mass spectrometers allow for selective fragmentation of interesting ions and detect the resulting fragments. This so-called tandem MS (or MS/MS) analysis may be spatial or temporal with separate mass analyzers (e.g. in triple quadrupoles) or ion traps, respectively. The latter capture ions in a single chamber and perform all steps of parent ion selection, fragmentation, and product ion analysis sequentially. In tandem MS, the fragmentation is induced by collision with gas molecules (usually nitrogen or argon) and may be modulated by adjusting the collision energy. Since MS/MS can rapidly switch between full range and product-ion mode, data-driven acquisition methods can be used to obtain the additional mass distributions of fragments. Initially, a full range survey scan identifies the

eluted ions. Product ion analysis is then done for the identified parent ions. The MS continuously cycles between these two modes throughout the run. Again, a compromise must be sought between short cycle times imposed by the chromatography and acquisition times required for accurate mass distribution analysis. Several MS/MS instruments are available with very different characteristics of scanning speed, duty cycle, mass and dynamic ranges, resolution, sensitivity, and mass accuracy. For the identification and quantification of natural intermediates, hybrid MS/MS systems combining quadrupoles with accurate orthogonal time-of-flight tubes (Chernushevich *et al.,* 2001) or sensitive linear ion-traps (Hager and Le Blanc 2003; Xia *et al.,* 2003) are probably the best choice for fluxome profiling.

As an alternative to tandem MS, off-line MS analysis by matrix assisted laser desorption ionization (MALDI) of chromatographic fractions could potentially increase measurement time without affecting throughput. Several fraction collectors are commercially available that can directly spot samples on MALDI surfaces from liquid chromatography or capillary electrophoresis (Bodnar *et al.,* 2003). Moreover, robotic systems can be interfaced to nano-scale fluidic systems. Since MS detection is decoupled from chromatography, more time is available for MS/MS characterization of important, large, or rare compounds. While MALDI is extensively used for biopolymer analysis, it may become relevant for small metabolites because of its robustness, convenience, and speed (Cohen and Gusev 2002). Generally, MALDI appears to be more appropriate for fluxome than metabolome studies because the former does not rely on precise concentrations that are problematic for the irregular sample distribution in the matrix crystals. The main problem is the background signal produced by the matrix that severely compromises analysis of molecules below a mass of 300-400 Da. Nevertheless, MALDI-based approaches have been applied successfully to produced metabolites (Wittmann and Heinzle, 2001; Zabet-Moghaddam *et al.,* 2004). Notably, laser desorption/ionization from porous silicon (DIOS) is a very promising matrix-free alternative to MALDI for the analysis of small molecules (Go *et al.,* 2003). DIOS combines the advantages of MALDI with background signals and fragmentation patterns that are comparable to those of ESI. Hence, it is seemingly the best, off-line technique for high-throughput and metabolism-wide flux studies.

5. CONCLUSIONS

In contrast to transcriptome, proteome, and metabolome data that assess network *composition*, fluxome data assess the *operation* of networks that results from metabolite and protein interactions and the kinetic properties of

enzymes (Bailey, 1999; Hellerstein, 2003; Sauer, 2004). At the highest resolution, integrated flux analysis of ^{13}C-experiments with elaborate isotopomer models quantifies actual molecular fluxes or *in vivo* reaction rates (Dauner *et al.*, 2001; Wiechert, 2001), the functional determinants of cellular physiology. Here we discussed primarily the application potential for the conceptually novel approach of comparative fluxome profiling, which can discriminate mutants/conditions solely from raw mass isotope data by multivariate data analysis. While it does not provide direct flux information such as integrated or analytical flux analysis, particular profile changes may be related to the underlying metabolic causality, for example by using ICA or, perhaps more generally, by supervised learning methods. As a model-independent approach, comparative profiling is applicable to any organism, tracer molecule, or condition - provided a labeled molecule is metabolized and its pattern can be traced in metabolites.

The full potential of fluxome profiling could be exploited by the detection of patterns in the metabolites that are also assessed in metabolomics. There are three main reasons why metabolite-based labeling pattern are more informative than those from proteinogenic amino acids. First, they enable direct monitoring of flux imprints that cannot be inferred from amino acids that are not synthesized from a given metabolite. Second, transient phenomena can be followed because metabolite pools are more rapidly exchanged than protein is synthesized. Third, cells without *de novo* amino acid (or protein) synthesis may be analyzed, for example in rich media or in the absence of growth.

As a new methodological concept, metabolite-based comparative fluxome profiling holds promise for high-throughput applications in areas like functional genomics, chemogenomic profiling, toxicology, and metabolic disease profiling, both in microbes and multi-cellular organisms. By monitoring metabolic network operation, fluxome profiles provide a perspective that is fully complementary to the metabolome network composition.

REFERENCES

Allen J *et al.* High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nat. Biotechnol.*, 21: 692-696 (2003).

Attias H. Independent factor analysis. *Neural. Comput.*, 11: 803-851 (1999).

Bailey JE. Lessons from metabolic engineering for functional genomics and drug discovery. *Nat. Biotechnol.*, 17: 616-618 (1999).

Blank L, Sauer U. TCA cycle activity in *Saccharomyces cerevisiae* is a function of the environmentally determined growth and glucose uptake rates. *Microbiology*, 150: 1083-1093 (2004).

Bodnar WM *et al.* Exploiting the complementary nature of LC/MALDI/MS/MS and LC/ESI/MS/MS for increased proteome coverage. *J. Am. Soc. Mass Spectrom.*, 14: 971-979 (2003).

Buckhaults P *et al.* Identifying tumor origin using a gene expression-based classification map. *Cancer Res.*, 63: 4144-4149 (2003).

Buziol S *et al.* New bioreactor-coupled rapid stopped-flow sampling technique for measurements of metabolite dynamics on a subsecond time scale. *Biotechnol. Bioeng.*, 80: 632-636 (2002).

Castrillo JI *et al.* An optimized protocol for metabolome analysis in yeast using direct infusion electrospray mass spectrometry. *Phytochemistry*, 62: 929-937 (2003).

Chernushevich IV *et al.* An introduction to quadrupole-time-of-flight mass spectrometry. *J. Mass Spectrom.*, 36: 849-865 (2001).

Choi BK *et al.* Effect of liquid chromatography separation of complex matrices on liquid chromatography-tandem mass spectrometry signal suppression. *J. Chromatogr. A*, 907: 337-342 (2001).

Christensen B *et al.* Simple and robust method for estimation of the split between the oxidative pentose phosphate pathway and the Embden-Meyerhof-Parnas pathway in microorganisms. *Biotechnol. Bioeng.*, 74: 517-523 (2001).

Christensen B *et al.* Analysis of flux estimates based on ^{13}C-labeling experiments. *Eur. J. Biochem.*, 269: 2795-2800 (2002).

Christiansen T *et al.* Metabolic network analysis of *Bacillus clausii* on minimal and semirich medium using ^{13}C-labeled glucose. *Metab. Eng.*, 4: 159-169 (2002).

Cohen LH, Gusev AI. Small molecule analysis by MALDI mass spectrometry. *Anal. Bioanal. Chem.*, 373: 571-586 (2002).

Cole RB (ed). *Electrospray ionization mass spectrometry. Fundamentals, Instrumentation, and applications*. Wiley, New York (1997).

Dauner M *et al.* Metabolic flux analysis with a comprehensive isotopomer model in *Bacillus subtilis*. *Biotechnol. Bioeng.*, 76: 144-156 (2001).

Dauner M, Sauer U. GC-MS analysis of amino acids rapidly provides rich information for isotopomer balancing. *Biotechnol. Prog.*, 16: 642-649 (2000).

Dauner M, Sauer U. Stoichiometric growth model for riboflavin-producing *Bacillus subtilis*. *Biotechnol. Bioeng.*, 76: 132-143 (2001).

Fenn JB *et al.* Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246: 64-71 (1989).

Fiehn O *et al.* Metabolite profiling for plant functional genomics. *Nat. Biotechnol.*, 18: 1157-1161 (2000).

Fischer E, Sauer U. Metabolic flux profiling of *Escherichia coli* mutants in central carbon metabolism using GC-MS. *Eur. J. Biochem.*, 270: 880-891 (2003a).

Fischer E, Sauer U. A novel metabolic cycle catalyzes glucose oxidation and anaplerosis in hungry *Escherichia coli*. *J. Biol. Chem.*, 278: 46446-46451 (2003b).

Fischer E *et al.* High-throughput metabolic flux analysis based on gas chromatography-mass spectrometry derived ^{13}C constraints. *Anal. Biochem.*, 325: 308-316 (2004).

Geladi P, Kowalski BR. Partial least square regression: a tutorial. *Anal. Chim. Acta*, 185: 1-17 (1986).

Go EP *et al.* Desorption/ionization on silicon time-of-flight/time-of-flight mass spectrometry. *Anal. Chem.*, 75: 2504-2506 (2003).

Hager JW, Le Blanc JC. High-performance liquid chromatography-tandem mass spectrometry with a new quadrupole/linear ion trap instrument. *J. Chromatogr. A*, 1020: 3-9 (2003).

Hellerstein MK. In vivo measurement of fluxes through metabolic pathways: the missing link in functional genomics and pharmaceutical research. *Annu. Rev. Nutr.*, 23: 379-402 (2003).

Höskuldsson A. PLS regression methods. *J. Chemometr.*, 2: 211:228 (1988).

Hyvärinen A *et al. Independent component analysis.* John Wiley and Sons, Inc., New York (2001).

Iizuka N *et al.* Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *Lancet*, 361: 923-929 (2003).

Jolliffe IT. *Principal component analysis*, 2nd edn. Springer Verlag, New York (2002).

Kelleher JK. Flux estimation using isotopic tracers: common ground for metabolic physiology and metabolic engineering. *Metab. Eng.*, 3: 100-110 (2001).

Ludwig H *et al.,* Transcription of glycolytic genes and operons in *Bacillus subtilis*: evidence for the presence of multiple levels of control of the *gapA* operon. *Mol. Microbiol.*, 41: 409-422 (2001).

Maharjan RP, Ferenci T. Global metabolite analysis: the influence of extraction methodology on metabolome profiles of *Escherichia coli*. *Anal. Biochem.*, 313: 145-154 (2003).

Marx A *et al.,* Determination of the fluxes in the central metabolism of *Corynebacterium glutamicum* by nuclear magnetic resonance spectroscopy combined with metabolite balancing. *Biotechnol. Bioeng.*, 49: 111-129 (1996).

McCabe BJ, Previs SF. Using isotope tracers to study metabolism: application in mouse models. *Metab. Eng.*, 6: 25-35 (2004).

Nguyen DV, Rocke DM. Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*, 18: 1216-1226 (2002a).

Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18: 39-50 (2002b).

Pramanik J, Keasling JD. A stoichiometric model of *Escherichia coli* metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. *Biotechnol. Bioeng.*, 56: 398-421 (1997).

Raamsdonk LM *et al.* A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.*, 19: 45-50. (2001).

Sauer U. High-throughput phenomics: experimental methods for mapping fluxomes. *Curr. Opin. Biotechnol.*, 15: 58-63 (2004).

Sauer U *et al.* Metabolic fluxes in riboflavin-producing *Bacillus subtilis*. *Nat. Biotechnol.*, 15: 448-452 (1997).

Sauer U *et al.* Metabolic flux ratio analysis of genetic and environmental modulations of *Escherichia coli* central carbon metabolism. *J. Bacteriol.*, 181: 6679-6688 (1999).

Schaefer U *et al.* Automated sampling device for monitoring intracellular metabolite dynamics. *Anal. Biochem.*, 270: 88-96 (1999).

Sherry AD *et al.* Analytical solutions for ^{13}C isotopomer analysis of complex metabolic conditions: substrate oxidation, multiple pyruvate cycles, and gluconeogenesis. *Metab. Eng.*, 6: 12-24 (2004).

Siler SQ *et al.* De novo lipogenesis, lipid kinetics, and whole-body lipid balances in humans after acute alcohol consumption. *Am. J. Clin. Nutr.*, 70: 928-936 (1999).

Soga T *et al.* Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. *J. Proteome Res.*, 2: 488-494 (2003).

Svetnik V *et al.* Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.*, 43: 1947-1958 (2003).

Szyperski T. Biosynthetically directed fractional ^{13}C-labeling of proteinogenic amino acids. An efficient analytical tool to investigate intermediary metabolism. *Eur. J. Biochem.*, 232: 433-448 (1995).

Szyperski T. ^{13}C-NMR, MS and metabolic flux balancing in biotechnology research. *Q. Rev. Biophys.*, 31: 41-106 (1998).

Tolstikov VV, Fiehn O. Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Anal. Biochem.*, 301: 298-307 (2002).

van Dam JC *et al.* Analysis of glycolytic intermediates in *Saccharomyces cerevisiae* using anion exchange chromatography and electrospray ionization with tandem mass spectrometric detection. *Anal. Chim. Acta*, 460: 209-218 (2002).

van Winden W *et al.* Possible pitfalls of flux calculations based on ^{13}C-labeling. *Metab. Eng.*, 3: 151-162 (2001).

Varma A, Palsson BO. Metabolic flux balancing: Basic concepts, scientific, and practical use. *Bio/Technol.*, 12: 994-998 (1994).

Visser D *et al.* Rapid sampling for analysis of in vivo kinetics using the BioScope: a system for continuous-pulse experiments. *Biotechnol. Bioeng.*, 79: 674-681 (2002).

von Roepenack-Lahaye E *et al.* Profiling of *Arabidopsis* secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. *Plant. Physiol.*, 134: 548-559 (2004).

Wiechert W. ^{13}C metabolic flux analysis. *Metab. Eng.*, 3: 195-206 (2001).

Wittmann C, Heinzle E. MALDI-TOF MS for quantification of substrates and products in cultivations of *Corynebacterium glutamicum*. *Biotechnol. Bioeng.*, 72: 642-647 (2001).

Xia YQ *et al.* Use of a quadrupole linear ion trap mass spectrometer in metabolite identification and bioanalysis. *Rapid Commun. Mass Spectrom.*, 17: 1137-1145 (2003).

Zabet-Moghaddam M *et al.* Qualitative and quantitative analysis of low molecular weight compounds by ultraviolet matrix-assisted laser desorption/ionization mass spectrometry using ionic liquid matrices. *Rapid Commun. Mass Spectrom.*, 18: 141-148 (2004).

Zamboni N *et al.*, The phosphoenolpyruvate carboxykinase also catalyzes C_3 carboxylation at the interface of glycolysis and the TCA cycle of *Bacillus subtilis*. *Metab. Eng.*, 6:277-284 (2004).

Zamboni N, Sauer U. Knockout of the high-coupling cytochrome aa_3 oxidase reduces TCA cycle fluxes in *Bacillus subtilis*. *FEMS Microbiol. Lett.*, 226: 121-126 (2003).

Chapter 18

TARGETED DRUG DESIGN AND METABOLIC PATHWAY FLUX

László G. Boros and Wai-Nang Paul Lee
SIDMAP, LLC, 10021 Cheviot Drive, Los Angeles, CA 90064

1. INTRODUCTION

Tumor cells inherently possess various mechanisms to initiate and sustain any one of the following phenotypes: 1, proliferative; 2, differentiated; 3, transformed; 4, cycle arrested; 5; necrotic; and 6, apoptotic (Boros *et al.*, 2002a, b). In addition to multiple drug and apoptosis resistance, advanced and therapy-resistant tumors share a common phenotype characterized by rapid proliferation, poor differentiation and increased transformation. They also exhibit increased rates of metabolism using glucose as a primary substrate (Pitot and Jost, 1967; te Boekhorst *et al.*, 1995; Smith, 1998; Schwart *et al.*, 1986). As such, factors that govern a tumor cells' response (growth, differentiation, etc.) to exogenous and endogenous agents are deeply embedded in, and dependent on, the metabolic network supplying essential substrates for *de novo* macromolecule synthesis and energy production. Rates of cellular proliferation are closely associated with rates of *de novo* macromolecule synthesis, such as RNA, DNA, proteins and long-chain fatty acids (Eigenbrodt *et al.*, 1992). These complex molecules, which eventually become structural components of new and old progenies of tumor cells, are synthesized from small molecular weight substrates, such as glucose, short chain fatty acids and amino acids in an interconnected and complex metabolic network. All pathways in the network depend on one another via substrate sharing and channeling, and by regenerating shared co-factors that participate in oxidative degradation and reductive synthesis simultaneously. Such a close relationship is evident between direct glucose

oxidation in the pentose cycle and *de novo* fatty acid synthesis, where part of the reduced $NADP^+$ pool is regenerated allowing the irreversible glucose-6-phosphate dehydrogenase reaction to proceed for the synthesis of five carbon sugars (Kuhajda, 2000; Baron *et al.*; 2004). In turn, the reducing $NADP^+$ equivalent is used during reductive *de novo* synthesis of fatty acids, their chain elongation and de-saturation, allowing distant metabolic network processes to proceed in a well-controlled and synchronized fashion.

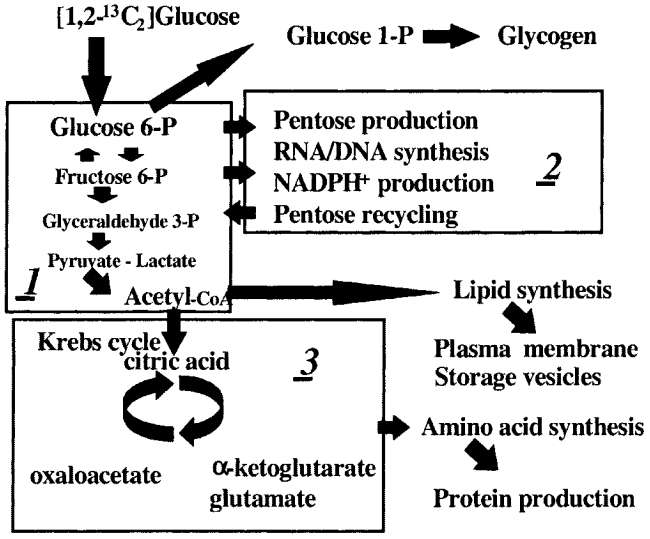


Figure 1. Interconnected metabolic pathways and their dynamic cross labeling by ^{13}C labeled glucose as the precursor. Glucose broadly utilized in mammalian cells readily labels major metabolite pools either as a direct substrate or through carbon exchange. The specificity for metabolic pathway substrate flow measurement is provided by the loss and rearrangements of the label from [1,2-$^{13}C_2$]glucose in various metabolites, intermediates and product pools. 1 glycolysis; 2 pentose cycle; 3 TCA cycle.

Stable isotope-labeled dynamic metabolic profiles (SIDMAP) can be particularly powerful in studies on the effect of endogenous and exogenous agents on intermediary metabolism in tumor cells. It can, for example, be applied to quantify induced changes in specific glucose metabolic reactions for nucleic acid synthesis, glucose oxidation and CO_2 production, amino acid synthesis, *de novo* lipid synthesis and TCA cycle anaplerotic flux, simultaneously, as shown in Figure 1 which illustrates the metabolic profiling potential of one particular labeled substrate [1,2-$^{13}C_2$] glucose, and thereby highlight the interconnectivity of these pathways. Well known

applications of SIDMAP include studies on the effect of novel anticancer agents such as Gleevec (Boros *et al.*, 2002b; 2003a,b) on glucose metabolism

Different cell phenotypes and their sensitivity to apoptosis show differences in their respective SIDMAPs of cross-regulated metabolic pathways in the network. Cells resistant to apoptosis can also be differentiated from cells more sensitive to programmed cell-death. The application of SIDMAP technology to uncover and interpret these metabolic differences is the primary focus of this review. The models discussed herein include therapy-resistant inflammatory breast cancer cells, apoptosis sensitive human fibroblasts and therapy sensitive pancreatic tumor cells. We herein argue that therapies targeting specific nodes or events in a metabolic network may overcome difficulties of drug design related to the enormous variability of the ever changing pool of genetic and proteomic targets in cancer (Cowan-Jacob *et al.*, 2004; Shah, *et al.*, 2004; Gullemard and Saragovi, 2004; Hu and Kavanagh, 2003). However, first one has to learn how to trace and read the map of the uniquely altered metabolic network of tumor cells in order to design new targeted therapies within and, as will be explained later, SIDMAP offers extraordinary potential here.

2. TARGETED THERAPIES OF CANCER USING GENETIC AND PROTEOMIC TARGETS

Targeted drugs are designed for a single or a very narrow range of genetic or protein targets. Although they have proven effective in the treatment of narrow cancer cell populations with a very favorable toxicity profile towards the host, they are limited by their high dependence on a single mechanistically defined target and by the rapid development of drug resistance. Drug resistance can arise from four major mechanisms including a decrease in target protein expression, mutations in target proteins, loss of target gene and/or construct due to clonal selection and increased drug transport from targeted cells. Drug resistance and a dependence on single target therapy have imposed significant delays in driving chemicals through the value chain of drug development and thereby imposed vast attrition related costs on the pharmaceutical industry. There are thus extensive research related cost accumulations and clinical disappointments with almost all targeted drugs designed so far. This is represented in Table 1, which summarizes major efforts in targeted drug design approaches against genetic and protein targets, as well as many of the major obstacles encountered.

Table 1. Targeted therapies of cancer and factors responsible for failures

	Pros	Cons	Latest developments	Ref
Monoclonal antibodies, and small molecule receptor antagonists	High specificity against cytokines and cell surface receptors	Host immune response, increased cytokine production	Low response rate especially in refractory disease	Dancey and Friedlin (2003)
Conjugated toxins or radioisotopes in leukemia	Targeted delivery, high efficacy drugs	Transporter and receptor dependent	Low response rate and recurrence	Nemecek and Matthews (2003); Sievers (2000)
Anti-sense oligonucleotides	Gene expression modifying RNA specific oligo-nucleotides	Delivery system is still not resolved, severe host response to viral vectors	Hemolytic anemia, renal failure and anasarca	Rudin *et al.* (2001)
Immunoliposome-encapsulated drugs	Targeted delivery of protected antibodies	Moderate stability, inability of the carrier to extravasate	Low efficacy, few clinical trials in progress and high failure rate	Matzku *et al.* (1990)
Small molecule inhibitors	Targets single oncogenic protein construct	Recurrence with blast disease in CML	Drug resistance is an emerging and severe problem	Hofmann *et al.* (2004)

The targeting of narrow oncogenic constructs with "magic bullets" is an approach that continues to elicit high expectations despite demonstrable limitations. Multiple resistance mechanisms arising from the ever changing genetic and proteomic maze of a tumor cell's regulatory network simply allow these cells to circumvent the desired effects of a candidate drug acting on a mechanistically and structurally defined target. This, of course, serves to make the "magic bullet" approach less effective and ever more expensive. New methodological and clinical approaches are clearly needed One proposed solution is to re-design drugs by imparting altered chemical features to hit a new target range (Shah *et al.*, 2004). This approach is limited however, given the logistics and costs involved in repeated pre-clinical and clinical trials for each new slightly modified drug, targeting slightly mutated proteins or increasing expression of genes. This will strain the research and development budgets of even large pharmaceutical companies. The cost of targeted therapy drug design is enormous compared to conventional treatments, yet returns on investment are limited due to

narrow demographic groups and limited number of patients who benefit from targeted therapies. It is also evident that health insurance companies will not be able to cover the ever growing costs of targeted therapies attempting to keep the ever changing genetic and protein profiles of tumor growth under long-term control (Danzon and Towse, 2002). Given only a few meaningful oncogenic permutations, as expected from gene and protein variations, targeted therapies may well be the most expensive journey medicine has yet taken with little in return regarding cancer disease outcome or improved population health.

A framework (or impetus) for addressing these issues can be provided by a look at the pharmaceutical industry. The exploratory stages of drug development represent the most expensive aspect of the value chain, including clinical trials. Early drug discovery is thus frequently identified (Boston Consulting Group, 2001) as the pharmaceutical area where re-organization and adoption of new enabling technologies is most likely to yield productivity gains. New approaches may additionally allow preclinical and clinical aspects to be addressed much earlier than current technologies allow. It is argued herein that metabolic pathway flux analysis represents a new enabling technology that can well serve cancer treatment and drug discovery

3. METABOLOMICS: THE STUDY OF THE TRANSFORMED METABOLIC NETWORK IN CANCER

Metabolomics can be considered a new enabling technology in medicine addressing and developing platforms, which will presumably afford better understanding of human biology and thereby allow more effective drug design against human diseases, including cancer (Schmidt, 2004). Metabolomics as a tool is designed around quantitative metabolite level measurement and ratios, which are mined using several pattern recognition techniques, including, but not restricted to, principal components analysis (Goodacre *et al.*, 2004). It has been defined as "comprehensive analysis of the metabolome under a given set of conditions" (Goodacre *et al.*, 2004). Derived from the greek "metabol" meaning change (metabolikos means changeable) metabolomics can be considered foremost as a science devoted to the analysis of metabolic changes in any biological system. There is a strong belief in the biomedical and agricultural communities that metabolomics will provide a strong complementary role to genetics and proteomics. This is exemplified, in the US, by the recent NIH Roadmap

Initiative. While there is still considerable reliance on functional genomics to elucidate further the role of genes and their protein products in human cancer, there is also an increasing necessity to define and understand how the genetic and protein networks conspire with the metabolism of particular tumor phenotypes (Griffin, 2004). In an excellent chapter in this volume (Chapter 2) Castrillo and Oliver also point out that evidence increasingly points to metabolites as much more than idle spectators of the "Central Molecular Dogma". They point out some recent findings such as evidence that endogenous metabolites excreted to the bloodstream (for example TCA cycle intermediates) have been found acting as signaling molecules for G-protein-coupled receptors, potentially linking intermediary metabolism and injury of tissues with blood pressure (He *et al.,* 2004; Hebert, 2004) and that metabolic pathways and metabolites (glycolysis and glucose) are associated with histone ubiquitination and gene silencing in yeast (Dong and Xu, 2004). In an increasingly cited test case study on Trypanosome glycolysis, ter Kuile and Westerhoff (2000) concluded that transcriptomics and proteomics analysis cannot suffice to adequately describe biological function.

Although it is known that metabolic networks vary in substrate utilization patterns and flux distribution according to cell function and phenotype, their control points have strictly been preserved throughout evolution and represent reliable drug targets considering the limited number of enzyme isoforms as well as the limited number of known major alternative metabolic routes. Studies on intermediary metabolism represent a venerable and traditional field of biochemical investigation but with continued and even new potential to assist drug development. Years after the glycolysis pathway was elucidated and its pioneering researchers were recognized with the Nobel Prize we are still exploring its role in phenotypic regulation. Metabolomics, including stable-isotope based methodologies, offers an approach to navigate the many interconnected pathways of a metabolic network, to trace and define targets and to determine and utilize existing control among distant metabolic pathways in the same network. This is now demonstrated in the following examples.

4. STABLE ISOTOPE LABELED METABOLIC NETWORK AND SENSITIVITY TO APOPTOSIS:

4.1 Apoptosis sensitive cells heavily depend on non-oxidative pentose cycle metabolism while lacking *de novo* fatty acid synthesis

The double tracer approach using stable isotope labeled glucose is particularly effective in revealing detailed substrate flow and distribution patterns in the complex metabolic network of human cells (see Figure 1). Applications in cancer have greatly facilitated an understanding of growth controlling mechanisms in transformed metabolic networks (reviewed in Boros *et al.*, 2002b; 2003a). A recent SIDMAP investigation of thiamine-reponsive megaloblastic anaemia (TRMA) was particularly illuminating.

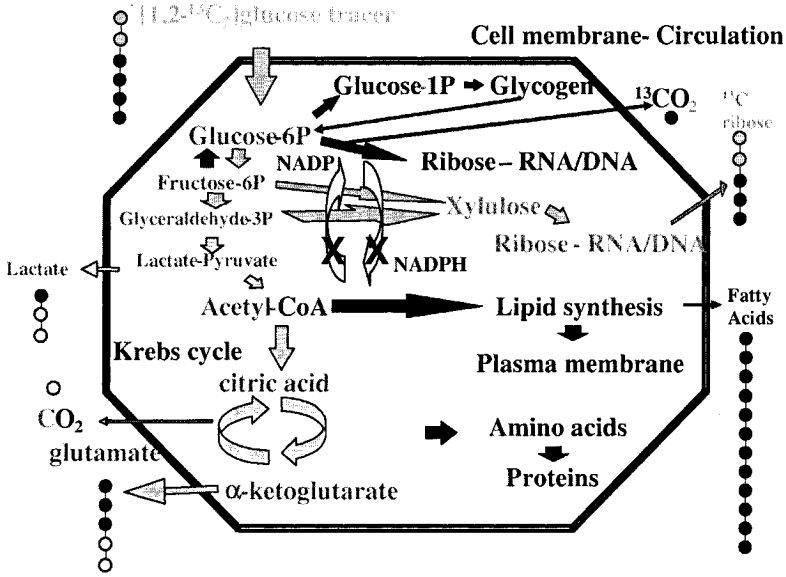


Figure 2. Stable isotope-based dynamic metabolic profile (SIDMAP) of apoptosis sensitive human cells. Grey arrows indicate routes of ^{13}C tracer glucose substrate carbons (grey filled circles) in the metabolic network. The heavy use of glucose carbons via the pentose cycle in human fibroblasts (TRMA cells) is primarily via the non-oxidative route, while the oxidative pathway is limited due to low NADP-NADPH cycling and fatty acid synthesis. Human fibroblasts with high affinity thiamine transport deficiency readily undergo spontaneous apoptosis (Boros *et al.*, 2003c) and MIA pancreatic adenocarcinoma cells show sensitivity and slow growth to non-oxidative pentose cycle inhibitors (Boros *et al.*, 2001a,b).

Apoptosis, after disruption of nucleic acid synthesis, is considered the final common pathway of hematopoiesis in TRMA (Green, 2003). SIDMAP was able to reveal that the underlying disruption of nucleic acid synthesis, which leads to premature apoptosis, resided in pentose cycle metabolism, specifically the transketolase enzyme which requires thiamine pyrophosphate as a cofactor (Boros *et al.*, 2003c). This thiamine co-factor becomes limited due to defective high affinity thiamine transport in thiamine responsive fibroblasts. By investigating normal and thiamine responsive fibroblasts in low and high-thiamine culture media Boros *et al.* (2003c) demonstrated that thiamine transport deficient human fibroblasts readily undergo apoptosis in culture with no rescue mechanism in place as they lack *de novo* fatty acid synthesis and therefore possess limited reserves of the oxidized form of $NADP^+$, which is the sole hydrogen acceptor during oxidative pentose synthesis from glucose in the cycle.

In a similar investigation, Boros *et al.* (1997) revealed that pancreatic adenocarcinoma cells show limited growth in response to pentose cycle inhibitors and possess a relatively low rate (20%) of *de novo* fatty acid synthesis and turnover during a 72 hour treatment period (Boros *et al.*, 2001a,b). Figure 2 demonstrates metabolic pathway substrate flow in apoptosis sensitive human cells using doubly labeled glucose as the tracer.

Pentose cycle inhibitors also induced cell cycle arrest in *in vivo* hosted Ehrlich's ascites carcinoma cells (Rais *et al.*, 1999), which demonstrated a high uptake of fatty acids and increased toxicity of sulfurated unsaturated fatty acids in culture due to limited *de novo* fatty acid synthesis and desaturation (Witek *et al.*, 1984)

4.2 Apoptosis resistant cells heavily depend on oxidative pentose cycle metabolism by maintaining high rate of *de novo* fatty acid synthesis and turnover

The SIDMAP metabolic profile of therapy and apoptosis resistant tumor cells is different from that of therapy sensitive cells shown on Figure 2. Figure 3 illustrates the metabolic profile of therapy resistant inflammatory breast cancer cells indicating intense tracer accumulation into fatty acids and oxidation of NADPH.

The main difference is the rate at which they synthesize medium and long chain saturated fatty acids up to the 16 carbon chain length, palmitate, and consequently elongate it to the 18 carbon length, stearate, and further into C:20-C:26 species. They also possess high fatty acid chain desaturase activity, further oxidizing NADPH and allowing the oxidative branch of the pentose cycle to operate under drug treatment. This is especially important

when nucleic acid synthesis inhibitors are targeting either branch of the pentose cycle; the operation of these alternative synthesis routes is essential for the survival of tumor cells and to endure apoptosis inducing drugs and signals.

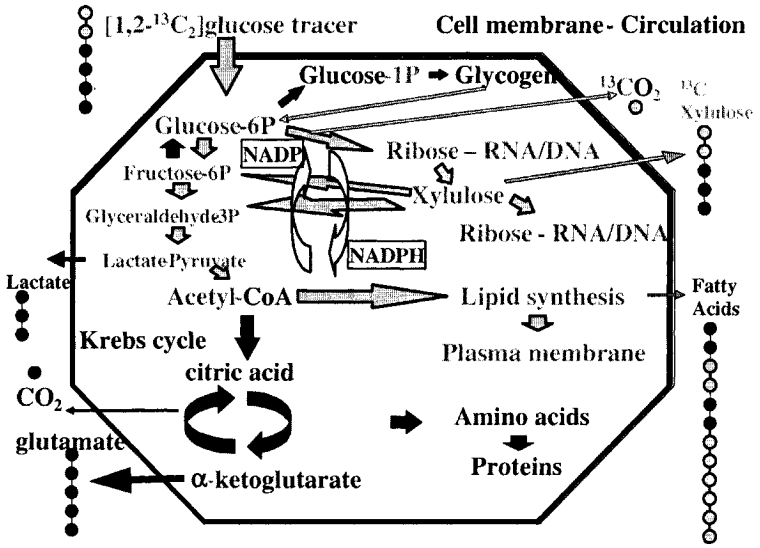


Figure 3. Stable isotope-based dynamic metabolic profile (SIDMAP) of apoptosis resistant tumor cells. Grey arrows indicate routes of ^{13}C tracer glucose substrate carbons (grey filled circles) in the metabolic network. NADPH-NADP cycling is active and is compensating in response to non-oxidative pentose cycle inhibitor treatment. Inflammatory breast cancer cells exhibiting this SIDMAP are extremely durable, treatment resistant and aggressive. Although growth retardation is achieved, inflammatory breast cancer cells cannot be forced into apoptosis even when the toxic glucose derivative 2-Deoxy-D-glucose derivative is given in high doses (5mM) (Boros, 2004).

5. FUTURE METABOLIC DRUG DESIGN SCENARIOS AND WHAT WE KNOW ALREADY

In the past two decades genetics and proteomics strategies have generated vast amounts of data to support the entry of new targeted therapies against unique genes and proteins into clinical practice for the treatment of cancer Significant drug resistance to new targeted therapies presents itself as a clinical challenge. Much laboratory research is now devoted to understanding and designing strategies to circumvent drug resistance. As

metabolic tracer data accumulates and the operation of the transformed metabolic network of tumor cells is further revealed by SIDMAP technologies, new opportunities have arisen via metabolic targeted therapies. The current challenges are several; firstly acquisition of more metabolomic and phenotypic data is needed and correlations between biological behavior and metabolic network characteristics established. Secondly, it is likely that metabolic targeted therapies against tumors will have to be aimed at multiple sites and control enzymes in the network due to the fact that metabolic networks are interconnected and alternative synthesis pathways are common. In other words, combination therapies may be required. Thirdly, individual SIDMAPs of tumor cells and host organs will facilitate the tailoring of metabolic targeted drugs to individual tumor growth characteristics in the host. Based on what is known so far, it can be predicted that limited *de novo* fatty acid synthesis of a given tumor will allow pentose cycle inhibitors to work effectively, while tumors possessing high rate of fatty acid turnover have to be targeted with a combined approach using fatty acid synthase, chain elongase and desaturase inhibitors, along with conventional drugs targeting pentose cycle synthesis, nucleic acid backbone sugar production, RNA synthesis, DNA replication and consequently cell proliferation. Tumor SIDMAPs can easily be determined both *in vitro* and *in vivo*, using non-invasive, non-radiating and natural sugar tracers for both diagnostic and metabolic targeting purposes.

6. CONCLUSIONS

As metabolic profiling allows new targets to be discovered, the promise of such targets is that they have significantly far less flexibility and variability than do genetic and signal protein targets to escape treatment. This is based on the fact that structures of metabolic enzymes and hierarchies of metabolic networks are well preserved throughout evolution and among species, and on the fact that tumor cells have to adhere to these hierarchies to survive. Regardless of the level of transformation and malignancy, tumor cells have to integrate and co-ordinate their metabolism with a complex host operating on a limited number of substrates and co-operative futile cycles. It is evident that mutations in growth signal proteins make them hidden from drugs that target them without losing function. Many growth signals exist contemporarily, they initiate downstream effects that can become constitutively active (Boren *et al.*, 2001) and they can maintain signaling by variations in gene expression. On the other hand, mutations in metabolic enzymes, albeit can let them escape newly designed metabolic reaction targeting drugs, also makes them non-functional and

defective to catalyze the metabolic reaction which the inhibitor tends to control. Over-expression of metabolic enzymes is rather a real threat to develop resistance against metabolic targeted therapies and this is where a combined approach of genomics, proteomics and metabolomics will change the future. The purpose of metabolomics in the new targeted era of drug design is to pinpoint targets in the fundamental component of cell function, the metabolome. These targets have very limited flexibility and variability to develop resistance by point mutations or structural/conformational changes, or any other mechanism that make genetic and protein targets weak and short-lived.

REFERENCES

Baron A, Migita T, Tang D, Loda M. Fatty acid synthase: a metabolic oncogene in prostate cancer? *J. Cell. Biochem.*, 91: 47-53 (2004).

Boren J, Cascante M, Marin S, Comin-Anduix B, Centelles JJ, Lim S, Bassilian S, Ahmed S, Lee WN, Boros LG. Gleevec (STI571) influences metabolic enzyme activities and glucose carbon flow toward nucleic acid and fatty acid synthesis in myeloid tumor cells. *J. Biol. Chem.*, 276: 37747-37753 (2001).

Boros LG, Puijaner J, Cascante M, Lee WN, Brandes JL, Bassilian S, Yusuf FI, Williams RD, Muscarella P, Melvin WS, Schirmer WJ. Oxythiamine and dehydroepiandrosterone inhibit the non-oxidative synthesis of ribose and tumor cell proliferation. *Cancer Res.*, 57: 4242-8 (1997).

Boros LG, Bassilian S, Lim S, Lee WN. Genistein inhibits non-oxidative ribose synthesis in MIA pancreatic adenocarcinoma cells: a new mechanism of controlling tumor growth. *Pancreas*, 22: 1-7 (2001a).

Boros LG, Lapis K, Szende B, Tomoskozi-Farkas R, Balogh A, Boren J, Marin S, Cascante M, Hidvegi M. Wheat germ extract decreases glucose uptake and RNA ribose formation but increases fatty acid synthesis in MIA pancreatic adenocarcinoma cells. *Pancreas*, 23: 141-147 (2001b).

Boros LG, Lee WN, Go VL. A metabolic hypothesis of cell growth and death in pancreatic cancer. *Pancreas*, 24: 26-33 (2002a).

Boros LG, Cascante M, Lee W-NP. Metabolic profiling of cell growth and death in cancer: applications in drug discovery. *Drug Discov. Today*, 7: 364-372 (2002b)..

Boros LG, Cascante M, Lee W-NP. Stable isotope-based dynamic metabolic profiling in disease and health. In: *Metabolic profiling: Its role in biomarker discovery and gene function analysis*. Eds. Harrigan GG, Goodacre R. Kluwer Academic Publishers, Boston (2003a).

Boros LG, Brackett DJ, Harrigan GG. Metabolic biomarkers and kinase drug targets in cancer using stable isotope-based dynamic metabolic profiling. *Curr. Cancer Drug Targets*, 3: 447-455 (2003b)..

Boros LG, Steinkamp MP, Fleming JC, Lee WN, Cascante M, Neufeld EJ. Defective RNA ribose synthesis in fibroblasts from patients with thiamine-responsive megaloblastic anemia (TRMA). *Blood*, 102, 3556-3561 (2003c).

Boros LG. Metabolic profile of inflammatory breast cancer: aiding diagnosis and treatment. George Washington University and the IBC Research Foundation co-sponsored "IBC Mini

Symposium" hosted by George Washington University: http://www.ibcresearch.org/ibcminisymposium/ (2004).

Boston Consulting Group. A revolution in R & D. How genomics and genetics are transforming the biopharmaceutical industry (2001).

Cowan-Jacob SW, Guez V, Fendrich G, Griffin JD, Fabbro D, Furet P, Liebetanz J, Mestan J, Manley PW. Imatinib (STI571) resistance in chronic myelogenous leukemia: molecular basis of the underlying mechanisms and potential strategies for treatment. *Mini. Rev. Med. Chem.*, 4: 285-299 (2004).

Dancey JE, Freidlin B. Targeting epidermal growth factor receptor--are we missing the mark? *Lancet*, 362: 62-64 (2003).

Danzon P, Towse A. The economics of gene therapy and of pharmacogenetics. *Value Health.* 5: 5-13 (2002).

Dong L and Xu CW. Carbohydrates induce mono-ubiquitination of H2B in yeast. *J. Biol. Chem.*, 279: 1577-1580 (2004).

Eigenbrodt E, Reinacher M, Scheefers-Borchel U, Scheefers H, Friis R. Double role for pyruvate kinase type M2 in the expansion of phosphometabolite pools found in tumor cells. *Crit. Rev. Oncog.*, 3: 91-115 (1992).

Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB. Metabolomics by numbers; acquiring and understanding global metabolite data. *Trends Biotechnol.*, 22: 245-252 (2004).

Green R. Mystery of thiamine-responsive megaloblastic anemia unlocked. *Blood*, 102: 3464-3465 (2003).

Griffin JL. Metabolic profiles to define the genome: can we hear the phenotypes? *Philos. Trans R Soc. Lond B Biol. Sci.*, 359: 857-871 (2004).

Guillemard V, Saragovi HU. Novel approaches for targeted cancer therapy. *Curr. Cancer Drug Targets.*, 4: 313-326 (2004).

He W, Miao FJ, Lin DC, Schwandner RT, Wang Z, Gao J, Chen JL, Tian H and Ling L. Citric acid cycle intermediates as ligands for orphan G-protein-coupled receptors. *Nature*, 429: 188-193 (2004).

Hebert SC. Physiology: orphan detectors of metabolism. *Nature*, 429: 143-145 (2004).

Hofmann WK, Komor M, Hoelzer D, Ottmann OG. Mechanisms of resistance to STI571 (Imatinib) in Philadelphia-chromosome positive acute lymphoblastic leukemia. *Leuk. Lymphoma*, 45: 655-660 (2004).

Hu W, Kavanagh JJ. Anticancer therapy targeting the apoptotic pathway. *Lancet Oncol.*, 4: 721-729 (2003).

Kuhajda FP. Fatty-acid synthase and human cancer: new perspectives on its role in tumor biology. *Nutrition*, 16: 202-208 (2000).

Matzku S, Krempel H, Weckenmann HP, Schirrmacher V, Sinn H, Stricker H. Tumor targeting with antibody-coupled liposomes: failure to achieve accumulation in xenografts and spontaneous liver metastases. *Cancer Immunol. Immunother.*, 31: 285-291 (1990).

Nemecek ER, Matthews DC. Use of radiolabeled antibodies in the treatment of childhood acute leukemia. *Pediatr. Transplant.*, 3: 89-94 (2003).

Pitot HC, Jost JP. Control of biochemical expression in morphologically related cells *in vivo* and *in vitro*. *Natl. Cancer Inst. Monogr.*, 26: 145-166 (1967).

Rais B, Comin B, Puigjaner J, Brandes JL, Creppy E, Saboureau D, Ennamany R, Lee WN, Boros LG, Cascante M. Oxythiamine and dehydroepiandrosterone induce a G1 pase cycle arrest in Ehrlich's tumor cells through inhibition of the pentose cycle. *FEBS Lett.*, 456: 113-118 (1999).

Rudin CM, Holmlund J, Fleming GF, Mani S, Stadler WM, Schumm P, Monia BP, Johnston JF, Geary R, Yu RZ, Kwoh TJ, Dorr FA, Ratain MJ. Phase I Trial of ISIS 5132, an

antisense oligonucleotide inhibitor of c-raf-1, administered by 24-hour weekly infusion to patients with advanced cancer. *Clin. Cancer. Res.*, 7: 1214-1220 (2001).

Schmidt C. Metabolomics takes its place as latest up-and-coming "omic" science. *J. Natl. Cancer Inst.*, 96: 732-734 (2004).

Schwartz AG, Pashko L, Whitcomb JM. Inhibition of tumor development by dehydroepiandrosterone and related steroids. *Toxicol. Pathol.*, 14: 357-362 (1986).

Shah NP, Tran C, Lee FY, Chen P, Norris D, Sawyers CL. Overriding imatinib resistance with a novel ABL kinase inhibitor. *Science*, 305: 399-401 (2004).

Sievers EL. Targeted therapy of acute myeloid leukemia with monoclonal antibodies and immunoconjugates. *Cancer Chemother. Pharmacol.*, 46: S18-22 (2000).

Smith TA. FDG uptake, tumor characteristics and response to therapy: a review. *Nucl. Med. Commun.*, 19: 97-105 (1998).

te Boekhorst PA, Lowenberg B, van Kapel J, Nooter K, Sonneveld P. Multidrug resistant cells with high proliferative capacity determine response to therapy in acute myeloid leukemia. *Leukemia*, 9: 1025-1031 (1995).

ter Kuile BH and Westerhoff HV. Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett.*, 500: 169-171 (2001).

Witek R, Kubis A, Krupa S. The cytotoxic action *in vitro* of catalytically sulphurated unsaturated fatty acids on Ehrlich's ascites cancer and normal peritonela exudates (leucocytes). *Pharmazie*, 39: 482-483 (1984).

Chapter 19

METABONOMICS IN THE PHARMACEUTICAL INDUSTRY

Current practice and future prospects

Eva M. Lenz, Rebecca Williams and Ian D. Wilson
Dept. of Drug Metabolism and Pharmacokinetics, Mereside, Alderley Park, Macclesfield, Cheshire SK10 4TG, UK.

1. INTRODUCTION

The development of ever more powerful analytical methodologies, combined with an increasing awareness that a more complete view of changing metabolic profiles is needed to understand living systems, has led to the development of a range of global metabolite fingerprinting strategies. These have been variously termed by their advocates as metabonomics and metabolomics. There is currently some confusion between the terms metabonomics and metabolomics and they are often used interchangeably. In an attempt to clarify this situation Nicholson and co-workers defined metabonomics as *"the quantitative measurement of the dynamic multiparametric response of living systems to pathophysiological stimuli or genetic modification"* (Nicholson *et al.*, 1999), whilst proposing that metabolomics provides similar information on metabolite profiles in cell-based systems rather than the whole organism. From the perspective of the pharmaceutical industry and the US Food and Drug Administration, metabonomics is the term routinely employed and, in general, metabonomics will be used here. Irrespective of the terminology the aim is to provide a "global" view of the metabolic status of an organism by determining metabolic fingerprints, rather than target analysis of particular compounds. To date, in the pharmaceutical industry, the main application of metabonomics has been in the area of toxicology (Nicholson *et al.*, 2002), although examples are now appearing of its use for studying disease/disease

models. Amongst other things this interest in toxicological applications has led to the formation of a group of major pharmaceutical companies and academics in the Consortium on Metabonomic Toxicology ("COMET"). The aim of the consortium was to build a large database of ^{1}H NMR-based metabonomic data for a range of ca. 100 model toxins in the rat. On the basis of the organ-specific toxicities studied, and the characteristic urinary metabolic fingerprints that result, it is expected that metabonomics can be used to detect similar toxicities produced by novel compounds at an early stage in drug discovery and development. Some of the results of the work of this consortium are now appearing in the literature (Lindon *et al.*, 2003).

2. ANALYTICAL PLATFORMS FOR METABONOMICS

The ideal characteristics of analytical techniques for metabonomic studies in pharmaceutical research are that such methods should provide as comprehensive a metabolic fingerprint as possible in a reasonably short analysis time so as to enable moderate to high throughput. Equally important, the technique should provide sufficient structural data as to enable the investigator to identify the marker or markers detected. Currently the two major analytical methods used for obtaining metabonomic data are based on either high-resolution proton nuclear magnetic resonance spectroscopy (^{1}H NMR) or, more recently, high performance liquid chromatography coupled to mass spectrometry (HPLC-MS).

The sample types that can be analysed by these techniques encompass all of those that might be required for toxicological analysis, including urine, bile, blood plasma, intact tissues and tissue extracts. When combined with chemometric techniques such as principal component analysis (PCA) etc., particular metabolites, or groups of metabolites that provide specific markers for a particular condition (e.g. toxicity, disease, physiological variation etc.) can be identified. In many ways urine provides an ideal method for the non-invasive study of the effects of such conditions on endogenous metabolic pathways. Samples can be taken over the duration of the study and provide a time course of effects that can be used to pinpoint onset and severity of toxicity, and determine the best times for other more invasive investigations. In addition, unless small rodents such as mice are involved, there are usually few restrictions on the size of the sample that is obtained. Blood plasma provides a more direct "window" on the organism under study, but clearly requires more invasive procedures. There are also well defined limits on the amounts of sample, and number of sampling times that can be taken in any given study. Tissue samples obtained from target organs clearly require

surgical intervention, which in animal studies are usually only obtained on autopsy. In the case of humans the removal of e.g. tumours, or diseased organs as part of therapy can afford the possibility of the direct study of these tissues.

However, when considering sampling, great care must be taken in all metabonomic studies to ensure both integrity and validity of the sample. Many factors can result in changes to sample composition, and for good results to be obtained these must be controlled. Perhaps the most obvious is that biofluid samples provide ideal growth media for bacteria and unless steps are taken to preserve e.g. urine being collected from animals in metabolism cages, then the metabolic profile observed may be more indicative of fermentation than a response to an experimental treatment. More subtle factors such as the time of day of collection, and the gender, age, strain and diet of the animals (or humans), as well as exercise and physical activity can have very significant effects on global metabolite profiles (e.g., see Bollard *et al.*, 2001; Gavaghan *et al.*, 2000, 2002; Holmes *et al.*, 1994).

2.1 Nuclear Magnetic Resonance (NMR) spectroscopy:

2.1.1 Liquid samples

NMR spectroscopy provides, in many ways, an ideal methodology for the non-targeted analysis of liquid samples for low molecular mass organic compounds. In particular, there is no need to pre-select the analytical conditions as biofluid samples such as plasma and urine can be directly introduced into the instrument without the need for any form of sample pre-treatment (other than the addition of a small amount of D_2O to act as a field frequency lock for the spectrometer). The spectra themselves have very high information content providing the possibility of rapid identification of analytes. In addition, because NMR spectroscopy is non-destructive, the sample can be used for other analyses (e.g. HPLC-MS, GC-MS). The technique also allows equilibria to be observed, which are usually destroyed when e.g. chromatographic methods are employed. An oft-quoted criticism of NMR spectroscopy as a bioanalytical tool for metabonomics is that it is relatively insensitive. However, this has to be set against the advantage that, unlike many other techniques (e.g. MS and UV spectroscopy), it is equally sensitive for all protons. Instrumental sensitivity is also continually improving as a result of increasing field strength, better probe design and innovations such as cryogenically cooled electronics and currently lies in the ng range. The use of NMR spectroscopy for the analysis of liquid samples

has been reviewed by Nicholson and co-workers (Nicholson and Wilson, 1989; Lindon *et al.*, 2004).

2.1.2 Solid Samples

Whilst best known for the analysis of liquid samples, NMR spectroscopy can also be used for the investigation of solid and semi-solid samples, such as tissue, via the technique of magic angle spinning (MAS). High resolution MAS has been employed in a number of metabonomic investigations including investigations of e.g., intact kidney tissues (Moka *et al.*, 1997) tumours (Cheng *et al.*, 1996; Tomlins *et al.*, 1998) and liver (Waters *et al.*, 2000) etc. The technique is complementary to the analysis of tissue extracts but has the advantage over the latter in that intracellular compartmentation is preserved. In a recent study the effect of the hepatotoxin paracetamol (acetaminophen) was investigated using both conventional solution ^{1}H NMR spectroscopy and high resolution ^{1}H MAS spectroscopy (Coen *et al.*, 2003, 2004). The MAS spectroscopic studies clearly showed large changes in the content of the liver tissue with a rapid decline in glucose and glycogen and an increase in lipid content (albeit with changing lipid composition with time). When combined with transcriptomic (Coen *et al.*, 2004) and proteomic (Tonge *et al.*, 2002) investigations this revealed an overall pattern consistent with a global energy failure in the liver. In addition to organ tissue, the technique can also be used on isolated organelles as illustrated by a recent study on metabolic compartmentation involving cardiac mitochondria (Bollard *et al.*, 2003).

2.2 Mass Spectrometry (MS) and High Performance Liquid Chromatography (HPLC)-MS

Unlike the situation noted above for NMR it is difficult to advocate the use of MS directly on biological fluids as the problems of e.g. ion suppression currently seem almost insurmountable. Indeed, our own (unpublished) observations have shown that the quality of the data obtained by direct infusion approaches for urine into the ion source *vs.* HPLC-MS clearly show the superiority of the latter method. The use of mass spectrometry in this area therefore relies on the use of the hyphenated techniques such as HPLC-MS and GC-MS. Although GC-MS has been more widely investigated for metabolome analyses in microbial and plant systems (e.g. Fiehn *et al.*, 2000 and Chapter 7 in this book), few investigations concentrate on metabonomic applications in the pharmaceutical industry. This does not reflect a lack of potential for GC-MS in this area, but simply a lack of application and there is every reason to suppose that examples will

appear in the future. To date relatively few HPLC-MS studies have been published though this can be expected to be a rapid area for growth. However, the current applications cover reasonably diverse topics including the study of toxicity in rats (Idborg-Bjorkman *et al.*, 2003; Lafaye *et al.*, 2003; Lenz *et al.*, 2004a, b; Plumb *et al.*, 2002) and metabotyping (metabolic phenotyping) of strain, gender and diurnal variation in mice (Plumb *et al.*, 2003). In our own studies we have used gradient reversed-phase HPLC-orthogonal acceleration (oa)-TOF-MS(MS) for the examination of urine obtained from rats exposed to a number of nephrotoxins (Lenz *et al.*, 2004a, b). In these studies a simple linear gradient has usually been applied with the samples analysed using both positive and negative electrospray ionisation (in separate analytical runs). The particular advantage of using a time of flight instrument is that accurate mass data can be obtained enabling atomic compositions to be deduced which, when combined with information on fragmentation can greatly help with the identification of unknowns. As well as conventional HPLC other formats are possible including the use of capillary HPLC columns and recently introduced "UPLC" (Ultra performance LC)-MS (Wilson *et al.*, in press).

3. APPLICATIONS OF METABONOMICS

3.1 The study of toxicity

As indicated above, an area where metabonomic research is already well established within the pharmaceutical industry is in the study of toxicity. The bulk of these studies have been conducted using NMR spectroscopic methods, but more recently HPLC-MS-based analysis has begun to be performed. The combination of the two techniques is particularly powerful as the different sensitivities and specificities of the enable a more complete metabolic profile to be generated. In one of these studies the effects of the administration of a single dose of the model nephrotoxin mercuric chloride (2.0 mg/kg, subcutaneous) to male Wistar-derived rats on the urinary metabolite profiles of a range of endogenous metabolites has been investigated (Lenz *et al.*, 2004a). Urine was collected for 9 days with analysis by HPLC-oa-TOF/MS and ^{1}H NMR spectroscopy both of which revealed marked changes in the pattern of endogenous metabolites as a result of $HgCl_2$-induced nephrotoxicity toxicity. The greatest disturbances in the urinary metabolite profiles was detected at 3 days post dose after which the metabolite profile gradually returned to a more normal composition. The urinary markers of toxicity detected using ^{1}H NMR spectroscopy included increases in lactate, alanine, acetate, succinate, trimethylamine (TMA), and

glucose together with reductions in the amounts of citrate and α-ketoglutarate. In contrast the HPLC-MS-detected markers (in positive ESI) included decreased kynurenic acid, xanthurenic acid, pantothenic acid and 7-methylguanine concentrations, whilst an ion at *m/z* 188, possibly 3-amino-2-naphthoic acid, was observed to increase. In addition, unidentified ions at *m/z* 297 and 267 also decreased after dosing. Negative ESI revealed a number of sulphated compounds such as phenol sulphate and benzene diol sulphate, both of which appeared to decrease in concentration in response to dosing, together with an unidentified glucuronide (*m/z* 326). One conclusion from this study was that both NMR and HPLC-MS (positive and negative ESI) give similar time courses for the onset of toxicity and recovery. However, the markers seen were quite different for each technique clearly suggesting a role for both types of analysis. Similar conclusions about the complementary nature of NMR and HPLC-MS were confirmed in an investigation of the nephrotoxicity of the immuno-suppressant, cyclosporin A (Lenz *et al.*, 2004b). In this study, 9 daily doses of 45 mg/kg/day for 9 days were given, with toxicity only becoming apparent after 7 days of administration. In this instance HPLC-MS analysis was complicated by the presence of ions derived from cyclosporin, its metabolites and the dosing vehicle. These had to be eliminated from the HPLC-MS data prior to analysis by PCA. There was excellent concordance between the observed time course of toxicity, whichever technique was used. However, as with the mercuric chloride example given above, the markers were different depending upon whether NMR or HPLC-MS was examined and in general we would therefore recommend that, wherever possible, both techniques should be used to analyse samples. The use of ^{1}H NMR spectroscopy for the study of toxicity is now well established within the pharmaceutical industry, and the complementary nature of HPLC-MS suggests that this technique will also become a routine tool for this type of investigation. Similarly a future role for GC-MS in this type of investigation seems highly likely.

3.2 The investigation of disease models

As well as providing organ specific biomarkers of toxicity metabonomics has similar potential in the investigation of animal disease models. Such studies may provide a means to understand the model better, and how it relates to the human disease process, and may also be able to provide novel biomarkers that can be used to monitor efficacy. An important part of using metabonomics in such disease models is, of course, to determine the differences between normal animals and the model system. As part of such background studies we have examined the urinary metabolite profiles of nude mice (extensively used in cancer models) and compared them with

normal black and white and nude mice. More recently, we have also begun to investigate the urinary and plasma metabolite profiles of "Zucker" rats (used as a model of diabetes). These basic investigations have reinforced the fact that metabolic profiles depend on strain, age, gender and factors such as diet, diurnal variation and gut microfloral populations. Such studies clearly show that a lack of care in experimental design will produce enough variability in sample composition to confound metabonomic (and other omic) analysis by masking treatment-induced changes.

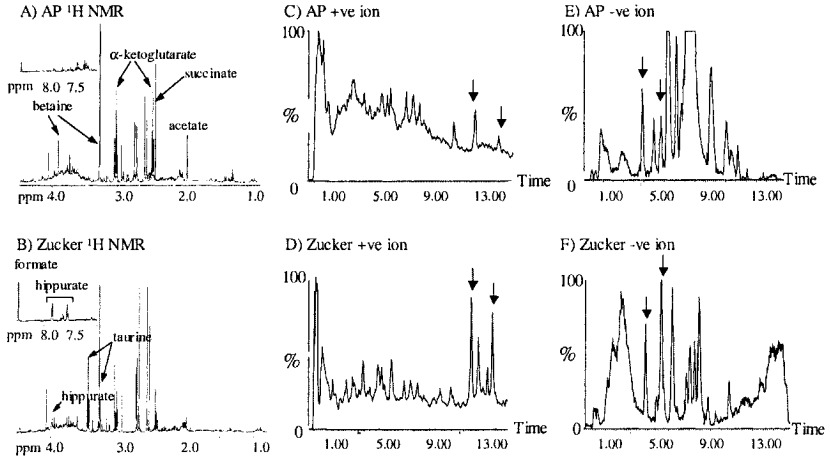


Figure 1. ^{1}H NMR spectra (0.8-4.5 and 7.0-8.5ppm) and positive and negative ion total ion chromatograms (TICs) (HPLC-MS analysis) obtained from urine samples collected from a 3 month old male AP rat (A, C, E) and a male Zucker rats (B, D, F) respectively.

In the case of the Zucker rat metabonomic analysis of urine from Zucker obese (fa/fa) rats was performed using ^{1}H-NMR and HPLC-MS to generate metabolite fingerprints. Diurnal and gender-based differences, as well as comparison with a Wistar-derived strain, were investigated using PCA and discriminant partial least squares analysis to analyse the spectroscopic data. Strain differences were evident in the ^{1}H NMR spectra as increased taurine, hippurate and formate and decreased betaine, α-ketoglutarate, succinate and acetate in Zucker compared to Wistar-derived rats. Similarly, HPLC-MS identified increased amounts of hippurate and unidentified ions at *m/z* 255.0640 and 285.0770 in positive, and 245.0122 and 261.0065 in negative, ESI. Both techniques revealed diurnal variation in the urine of Zucker rats due to elevated taurine, creatinine, allantoin and α-ketoglutarate by ^{1}H NMR

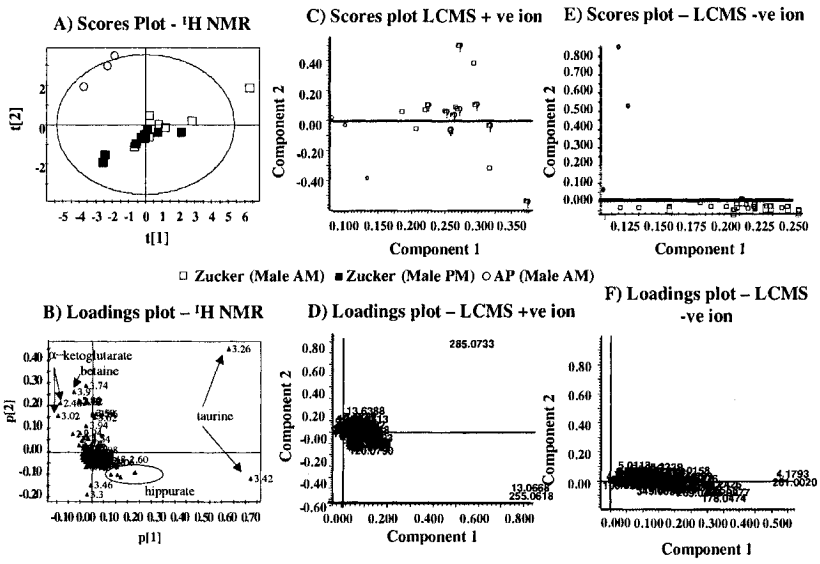


Figure 2. PCA scores (A,C,E) and loadings (B,D,F) plots (component 1 versus component 2) obtained from ^{1}H NMR spectra (A,B) and positive (C,D) and negative (E,F) ion HPLC-MS data from urine samples collected from 3 month old male AP and Zucker rats (each point represents a single sample). Loadings shown are the mid-segment value (ppm) for ^{1}H NMR data and the retention time and m/z for HPLC-MS data.

spectroscopy, whilst HPLC-MS showed that and ions at *m/z* 285.0753, 291.0536 and 297.1492 (positive ESI) and 461.1939 (negative ESI) were higher in the evening samples. Gender was also a discriminating factor with hippurate, succinate, α-ketoglutarate and dimethylglycine elevated in the ^{1}H NMR spectra of the urine of female Zucker rats compared to the males. Gender discrimination could also be obtained using HPLC-MS with ions at *m/z* 431.1047, 325.0655, 271.0635 and 447.0946 (positive ESI) and *m/z* 815.5495 and 459.0985 (negative ESI) by HPLC-MS. Typical spectra obtained by ^{1}H-NMR spectroscopy the urine of 12 week old male AP and Zucker rats are shown in Figure 1a and b whilst the corresponding HPLC total ion current chromatogram (TIC) mass chromatograms are shown in Figure 1c and d (+ve ESI) and e and f (-ve ESI). As indicated above, one of the "markers" identified by both NMR spectroscopy and HPLC-MS was hippuric acid. Interestingly, hippurate is largely derived *via* gut microfloral metabolism of dietary compounds (Phipps *et al.* 1997; 1998; Williams *et al.*, 2002). In this instance, therefore, the difference in hippurate concentrations between normal Wistar-derived and Zucker rats may be due simply to different populations of gut microflora rather than an underlying, directly

disease-related, difference in biochemistry. The scores and loadings plots shown in Figure 2 a-f show the results of PCA of the ^{1}H-NMR and HPLC-MS data obtained from the male AP and Zucker rats. In all scores plots (Figure 2a, c and e) there is a clear separation of the urine from the two strains with the AP samples clustering away from the Zucker ones.

3.3 Metabonomics in the clinic:

3.3.1 Studies in man

Greater methodological problems are associated with performing studies in humans rather than animals. The most obvious of these is the much greater inherent variability in human populations, compared with studies which are undertaken on inbred strains of animals. The animals are housed in uniform, and carefully controlled laboratory conditions, and are all of the same age, gender, weight range and diet. In comparison humans, in addition to not being housed in uniform environmental conditions, are subject to great variation in virtually all of the things that are carefully controlled in animal studies. Diet in particular can have a very large effect on the urinary metabonome. Recently, we have shown that dietary components (ethanol and ethyl glucoside) associated with the probable use of rice wine in cooking (or Saki consumption) appear (Teague *et al.,* 2004) in the urine. Similarly, we have found that presumed differences in diet, probably associated with a higher consumption of fish, can be used to separate British and Swedish populations (Lenz *et al.,* 2004). Changes in the dietary habits of volunteers can also have dramatic effects on metabolic profiles. This is illustrated by the urinary profile of a female subject from whom two samples were obtained some months apart. When the first sample was obtained the subject was following the "Atkins" diet (characterised by high meat consumption) but by the time of the second sample this had been discontinued (Figure 3). Instead, other life style markers were evident, such as alcohol consumption and a change to a fish diet. All of this can greatly increase the "metabolic noise" associated with human studies but, to some extent, once recognised these variables can be controlled. We have recently demonstrated that subjects in a Clinical Pharmacology Unit, with diet controlled on two study days a fortnight apart, although showing great inter-individual variation could be used as their own controls (Lenz *et al.,* 2003) thereby enabling the use of the technique in clinical trials.

^{1}H NMR spectroscopy has been used to examine the renal toxicity of ifosphamide in cancer patients undergoing therapy over a number of treatment cycles, with maximum renal toxicity seen by the fourth treatment

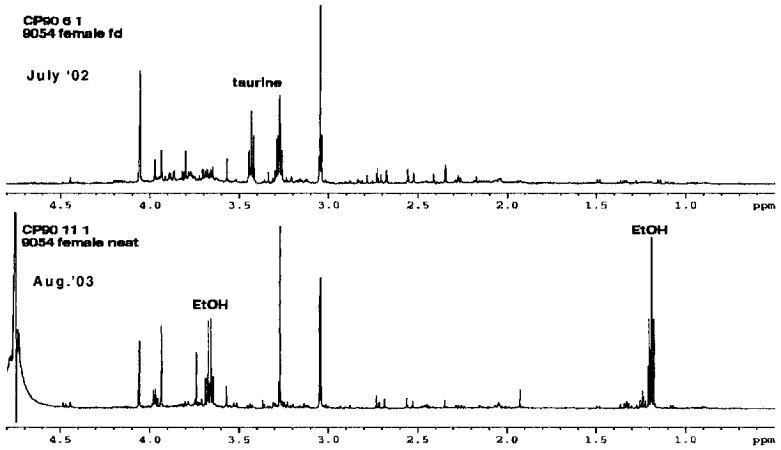


Figure 3. ^{1}H NMR urine spectra of a British female volunteer showing high concentrations of taurine due to the Atkins diet (upper spectrum) and that obtained some 12 months later. The prominent triplet at 1.19 ppm is for ethanol (the smaller triplet adjacent to it is probably ethyl glucoside).

cycle (Foxall *et al.*, 1997), illustrating the potential of metabonomics to monitor drug toxicity in the clinic. As well as detecting toxicity the potential for clinical use of metabonomics in diagnosis has recently been graphically illustrated by a study of cardiovascular disease in man in which ^{1}H NMR spectroscopy of blood plasma was able to accurately assess the level of coronary atherosclerosis (Brindle *et al.*, 2002), which is currently only possible with invasive techniques such as angiography.

It seems therefore, that just as in toxicological and animal model studies metabonomics, in combination with the other omic technologies such as genomics, transcriptomics and proteomics may well have an important part to play in the discovery and development of new medicines.

REFERENCES

Bollard ME *et al.* Investigations into biochemical changes due to diurnal variation and estrus cycle in female rats using high-resolution ^{1}H NMR spectroscopy of urine and pattern recognition. *Anal. Biochem.*, 295: 194-202 (2001).

Bollard ME *et al.* A study of metabolic compartmentation in the rat heart and cardiac mitochondria using high-resolution magic angle spinning ^{1}H NMR spectroscopy. *FEBS Lett.*, 533: 73-78 (2003).

Brindle JT *et al.* Rapid and non-invasive diagnosis of the presence and severity of coronary heart disease using ^{1}H-NMR-based "metabonomics". *Nat. Medicine*, 8: 1439-1444 (2002).

Cheng LL *et al.* Enhanced resolution of proton NMR spectra of malignant lymph nodes using magic angle spinning. *Magn. Reson. Med.*, 36: 653-658 (1996).

Coen M *et al.* An integrated metabonomic investigation of acetaminophen toxicity in the mouse using NMR spectroscopy. *Chem. Res. Toxicol.*, 16: 295-303 (2003).

Coen M *et al.* Integrated application of transcriptomics and metabonomics yields new insight into the toxicity due to paracetamol in the mouse. *J. Pharm. Biomed. Anal.*, 35: 93-105 (2004).

Fiehn O *et al.* Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal. Chem.*, 72: 3573-3580 (2000).

Foxall PD *et al.* Urinary proton magnetic resonance studies of early ifosfamide-induced nephrotoxicity and encephalopathy. *Clin. Cancer Res.*, 3: 1507-1518 (1997).

Gavaghan CL *et al.* An NMR-based metabonomic approach to investigate the biochemical consequences of genetic strain differences; application to the C57Bl10J and Alpk:Apcf CD mouse. *FEBS Letters*, 484: 169-174 (2000).

Gavaghan CL, Wilson ID and Nicholson JK. Physiological variation in metabolic phenotyping and functional genomic studies: Use of orthogonal signal correction and PLS-DA, *FEBS Letters*, 530: 191-196 (2002).

Holmes E *et al.* Automatic data reduction and pattern recognition methods for analysis of ^{1}H nuclear magnetic resonance spectra of human urine from normal and pathological states. *Anal. Biochem.*, 220: 284-296 (1994).

Idborg-Bjorkman H *et al.* Screening of biomarkers in rat urine using LC/electrospray ionization-MS and two-way data analysis. *Anal. Chem.*, 75: 4784-4792 (2003).

Lafaye A *et al.* Metabolite profiling in rat urine by liquid chromatography/electrospray ion trap mass spectrometry. Application to the study of heavy metal toxicity, *Rapid Commun. Mass Spectrom.*, 17: 2541-2549 (2003).

Lenz EM *et al.* A metabonomic investigation of the biochemical effects of mercuric chloride in the rat using ^{1}H NMR and HPLC-TOF/MS: Time dependant changes in the urinary profile of endogenous metabolites as a result of nephrotoxicity. *The Analyst*, 129: 535-541 (2004).

Lenz EM *et al.* B. Cyclosporin A-induced changes in endogenous metabolites in rat urine: A metabonomic investigation using high field ^{1}H NMR spectroscopy, HPLC-TOF/MS and chemometrics. *J. Pharm. Biomed. Anal.*, 35: 599-608 (2004).

Lenz EM, Bright J, Wilson ID, Morgan SR and Nash AFP. A ^{1}H NMR-based metabonomic study of urine and plasma samples obtained from healthy human subjects. *J. Pharm. Biomed. Anal.*, 33: 1103-1115 (2003).

Lenz EM *et al.* Metabonomics, Dietary influences and cultural differences: A ^{1}H NMR-based study of urine samples obtained from healthy British and Swedish subjects. *J. Pharm. Biomed. Anal.*, 36: 841-849 (2004).

Lindon JC *et al.* Contemporary issues in toxicology:The role of metabonomics in toxicology and its evaluation by the COMET project. *Toxicol. Appl. Pharmacol.*, 187: 137-146 (2003).

Lindon JC, Holmes E and Nicholson JK. Toxicological applications of magnetuic resonance, *Prog. NMR Spectrosc.*, 45: 109-143 (2004).

Moka D *et al.* Magic angle spinning proton NMR spectroscopic analysis of intact kidney tissue samples. *Anal. Commun.*, 34:107-109 (1997).

Nicholson JK, Lindon JC and Holmes E. "Metabonomics": Understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR data. *Xenobiotica*, 29: 1181-1189 (1999).

Nicholson K and Wilson ID. High resolution proton magnetic resonance spectroscopy of biological fluids, *Prog. NMR Spectrosc.*, 21: 449-501 (1989)

Nicholson JK *et al.* Metabonomics: a platform for studying drug toxicity and gene function. *Nature Rev. Drug Disc.*, 1: 253-258 (2002).

Phipps AN *et al.* Use of proton NMR for determining changes in metabolite excretion profiles induced by dietary changes in the rat, *Pharmaceutical Sciences*, 3: 143-146 (1997).

Phipps AN *et al.* Effect of diet on the urinary excretion of hippuric acid and other dietary-derived aromatics in rat. A complex interaction between diet, gut microflora and substrate specificity. *Xenobiotica*, 28: 527-537 (1998)

Plumb RS *et al.* Metabonomics: the use of electrospray mass spectrometry coupled to reversed-phase liquid chromatography shows potential for the screening of rat urine in drug development. *Rapid Commun. Mass Spectrom.*, 16: 1991-1996 (2002).

Plumb R *et al.* Metabonomic analysis of mouse urine by liquid-chromatography-time of flight mass spectrometry (LC-TOFMS): detection of strain, diurnal and gender differences. *The Analyst*, 128: 819-823 (2003).

Teague C *et al.* Ethyl glucoside in human urine following dietary exposure: detection by ^{1}H NMR spectroscopy as a result of metabonomic screening of humans. *The Analyst*, 129: 259-264 (2004).

Tomlins A *et al.* High resolution magic angle spinning ^{1}H nuclear magnetic resonance analysis of intact prostatic hyperplasic and tumour tissues. *Anal. Commun.*, 35: 113-115. (1998).

Tonge R *et al.* Genomics and proteomics analysis of acetaminophen toxicity in mouse liver, *Toxicol. Sci.*, 65: 135-150 (2002).

Waters NJ *et al.* High resolution magic angle spinning NMR spectroscopy of intact liver and kidney: optimisation of sample preparation procedures and biochemical stability of tissue during spectral acquisition. *Anal. Biochem.*, 262: 16-23 (2000).

Williams RE *et al.* Effect of intestinal microflora on the urinary metabolic profile of rats: a ^{1}H-nuclear magnetic resonance spectroscopy study. *Xenobiotica*, 32: 783-794 (2002)

Wilson ID *et al.* HPLC-MS-based methods for the study of metabonomics, *J.Chrom., B., (in press)*

Chapter 20

HOW LIPIDOMIC APPROACHES WILL BENEFIT THE PHARMACEUTICAL INDUSTRY

Alvin Berger
Head of Biochemical Profiling, Icoria Inc. (formerly Paradigm Genetics, Inc)., 108 Alexander Dr., Research Triangle Park, NC, 27709

1. WHAT IS LIPIDOMICS

Metabolomics is the latest of the 'omic' based sciences that is beginning to garner both academic, government and industrial interest worldwide (Adams, 2003; Fiehn, 2002; Phelps *et al.*, 2002; Sumner and Liu, 2002; Varnau and Singhania, 2002; Watkins and German, 2002; Weckwerth and Fiehn, 2002). Metabolomics is the science of identifying all of the low molecular weight metabolites in a biological sample such as cell, tissue or biofluid. Metabolomics provides for an accurate estimate of phenotypic changes to an organism. In human studies, metabolomics is most commonly used to compare normal and diseased biological samples, and for comparing placebo and drug-treated samples. The aim from such studies is typically to identify disease or drug efficacy and safety biomarkers. Lipids represent key signaling molecules which control, or are (bio)-markers of, physiological and disease processes. They are also key structural components of cellular membranes. Lipidomics is thus a subset of metabolomics that aims to detect and quantify all lipid species within a biological sample (Ejsing *et al.*, 2004; Fisher-Wilson, 2003; Forrester *et al.*, 2004; Han and Gross, 2003; Ivanova *et al.*, 2004; Watkins, 2004; Welti and Wang, 2003; Yang, 2003). In a lipidomic approach, collected data is typically analyzed using modern statistical approaches such as clustering and related approaches (Mutch *et al.*, in press); presented in an intuitive format such as a heat map; and the data stored so that global patterns of change in measured lipids can be

recalled to facilitate our understanding of signaling cascades responsible for observed patterns of change. For example, in mouse liver, lipidomic investigation of fish oil feeding, containing docosahexaenoic acid (DHA), and administration of the PPARα agonist Wy-14,643, lead to similar changes in lipids since both DHA and Wy-14,643 signal through PPARα (Berger and Roberts, 2004).

Lipidomics, may be combined with metabolomics, proteomics (Hanash, 2003; Wulfkuhle *et al.*, 2003; Ziboh *et al.*, 2002), phospho-proteomics (Mann and Jensen, 2003), and transcriptomics, in a systems biology approach. This can provide a particularly powerful suite of tools to examine phenotypic changes. In such a systems biology approach, it is possible to correlate lipidome changes to changes in the global transcriptome, and physiological sequel (Berger *et al.*, 2002a; Berger *et al.*, 2002b). Experiments that monitor changes to mRNA transcripts (in response to treatment or disease state) which are expected to affect lipid metabolism should, however, not be termed lipidomics, but rather transcriptomics.

For many years, lipid researchers commonly quantified levels of 25 or more fatty acids in individual phospholipids from tissues and biofluids using TLC-GC FID approaches, and in some cases mass spectral approaches. The data was typically not stored in an omic approach as described above. Today, it is fashionable to refer to limited lipid characterizations as targeted lipidomics (see Section 3 for further explanation).

2. LIPID CLASSIFICATIONS

Lipids may be classified according to chemical structure, function, and polarity. In structural classification schemes, lipids are classified according to their common back bone. For example, sphingolipids, phospholipids, and neutral lipids contain sphingoid, glycerol-3-phosphate, and glycerol, back bones respectively. Functional classifications include: barrier lipids (such as in the skin), signaling lipids (such as in caveolae), and storage lipids (such as in adipose tissue). Polarity classifications include: Highly polar (glycolipids), polar (phospholipids), and non-polar (neutral lipids such as fatty acids and triacylglycerol). Polarity is also related to solubility, thus another classification could be: water soluble; acetone soluble (glycolipids), methanol soluble (phospholipids), and hexane soluble (neutral lipids, etc.), although lipids will not partition uniquely into one solvent class.

Lipids, such as eicosanoid classes, may also be separated by chirality. This is important, since during the disease state, chiral enantiomers of key lipids may be formed with altered functions. For example, in psoriasis, a novel 12R lipoxygenase (LOX) product forms (12R hydroxyeicosatetraenoic

acid; HETE), rather than the usual 12S HETE product (Boeglin *et al.*, 1998; Bowcock *et al.*, 2001). In the diseased state, and when applying drugs, it is important to determine not only quantitative changes to specific bioactive lipids, but also lipid chirality with chiral columns.

Lipids may also be classified according to their sub cellular fractionation. For example, sphingolipids, gangliosides, and cholesterol concentrate in specific domains known as rafts and caveolae, which may affect caveolae functioning and signalling to the nucleus (Foster *et al.*, 2003; Parton, 2003). Caveolae perturbations have been linked to disease states including cancer (Bender *et al.*, 2002) and Alzheimer's disease (Dufour *et al.*, 2003). Caveolae represent a possible vesicular trafficking pathway through cell barriers including endothelium and epithelium, and may permit for targeted drug therapies (Schnitzer, 2001). Examining changes to whole cell lipids in the diseased state or following drug treatment, is not sufficient, since the most dramatic changes to lipids may be localized to lipid rafts.

In cells, lipids are asymmetrically distributed on the two leaflets of the plasma membrane and can thus be classified by whether they are predominately on the inner or outer leaflet. Maintenance of transbilayer lipid asymmetry is essential for normal membrane function, and disruption of this asymmetry is associated with cell activation or pathologic conditions. As an example, phosphatidylserine (PS) is known to be externalized during apoptosis and platelet activation (Daleke, 2003), and unregulated loss of PS asymmetry has been linked to heart disease, stroke, and diabetes. Following drug treatment and when comparing normal to the diseased state, one should assess changes to lipid asymmetric distribution by treating cells with phospholipase A_2 (Gascard *et al.*, 1991) and using other published techniques. If there are changes to the normal membrane assymetry, new drug targets could be developed to restore the normal asymmetric distribution.

3. CONVENTIONAL APPROACHES *VS* LIPIDOMICS

Before the advent of lipidomics, an experimenter might evaluate a precise hypothesis: Does drug X affect PGE_2 levels? In targeted lipidomics, the question posed would be: does drug X affect prostaglandin levels? Since about 1920, lipid researchers have conducted what would now be called targeted lipidomics. That is, researchers asked whether a particular diet, drug, or disease state influenced a class of lipids. Prior to recent advances in LC, mass spectroscopy (MS) and hyphenated approaches, most lipid researchers used gas chromatography (GC) to examine changes to non-polar

lipid classes, most commonly 20-40 fatty acids, analyzed as their methyl esters.

In a more open-ended, but still targeted lipidomic approach, one would pose the following question: does drug X affect eicosanoid levels? There are easily more than 1000 eicosanoids (defined here as hydroxylated and epoxylated derivatives of 18-22 carbon fatty acids and fatty acid derivatives, such as primary amides, ethanolamides, etc.). Hence such an approach would not have been possible with past technologies.

In broad lipidomics, the hypothesis would be: does drug X affect levels of all measurable lipids? To even pose such a question 10 years ago would be considered ludicrous. This goal will likely be realized within the next 2-5 years with the advent of various lipidomic academic and commercial large scale, well-funded initiatives, and the advent of more advanced MS techniques such as FTMS.

In a broad lipidomic-combined systems biology approach, the hypothesis would be: does drug X affect levels of all measurable lipids? What are the corresponding changes to the transcriptome? At what rate do the lipids change in the pathways affected?

Various systems biology academic programs and commercial companies have recently emerged, bent on addressing such questions.

4. HOW MANY LIPIDS ARE THERE?

Table 1. Calculation of selected theoretical neutral lipids (FA, FACoA, TAG, DAG, and MAG)

Types of lipid	Number of lipids in class
40 common FA	40
1,2,3 TAG	40*40*40=64,000
1-,2-, and 3-MAG	40+40+40=120
40 FA acyl CoA	40
1,2 + 2,3+ 1,3 DAG	(40*40)*3=4800
SUM	69,000

FA, fatty acid; TAG, triacylglycerol, MAG, monoacylglycerol, DAG, diacylglycerol

New classes of lipids are still being discovered, which makes it difficult to calculate the true number of lipids. Such new molecules include endocannabinoids, N-acyl linked- amino acids and dopamine, and various oxygenated and epoxylated bioactive derivatives of AA and DHA (Amer *et al.*, 2003; Berger *et al.*, 2002a; Burstein *et al.*, 2002; Capdevila *et al.*, 2003; Chu *et al.*, 2003; Cowart *et al.*, 2002; Hong *et al.*, 2002; Huang *et al.*, 2002; Serhan, 2002; Walker and Huang, 2002). A calculation of the true number of

lipids is beyond our present scope, but examples of such calculations appear in Table 1.

5. HOW TO MEASURE THE LIPIDOME?

Lipids were traditionally separated by TLC and 2-dimensional TLC, then methylated and analyzed by GC with a capillary column to assess acyl changes. GC/MS was used to confirm identity. HPLC was used to quantify more polar lipids such as hydroxylated lipids (e.g., HETES). Today, a large number of laboratories use MS, LC/MS, and MS^n approaches to analyze lipids of all classes. Sensitivity has been increased by derivitization, ionization, and the detector type. As shown in Table 2, techniques such as FTMS have enabled researchers to separate up to 11,000 peaks in a single run (Hughey *et al.*, 2002). Clearly, mass spectroscopy advances are driving the metabolomic and lipidomic fields.

Table 2. Summary of methods used to perform lipidomics

Separation Method	Max Peak capacity	Theoretical plates/Resolution
HP-TLC	25	1,000
Gradient LC	200	60,000
HPLC	1,000	1.5 million
CE	1,000	1.5 million
GC	1,000	1.5 million
ESI-FT-MS	130,000	2.5 billion

Modified from (He, 2002).

Abbreviations: HP-TLC, high performance TLC; HPLC (high performance) liquid chromatography; CE, capillary electrophoresis; GC, gas chromatography; ESI-FTMS, electrospray ionisation Fourier transform mass spectroscopy.

6. DIVIDING THE LIPIDOME INTO MODULES

Various groups around the world have the expertise to characterize fully a particular class of lipids. To date, there is not a single group, or single technique, that has been developed to characterize all lipids. For this reason, it is reasonable to conquer the lipidome by dividing it into classes separated by polarity, or by function as shown in the illustrative Table 3.

Table 20-3. Lipidome modules

Lipid Module	Molecules Studied	Function
P450ome	P450-derived lipids (EETs, ω and ω-1 HETES)	Dilation, constriction of arteries, mitogenic properties
Retinome	Retinoids	Signalling molecules
Calcitriome	Vitamin D compounds	Signalling molecules, calcium homeostasis
Prostaglandome	Prostaglanins	Inflammatory responses (includes thromboxanes)
Leukotrienome	Leukotrienes	Inflammatory responses
HETEome	Hydroxylated fatty acids	Inflammatory responses
Antioxidantome	Tocopherols, tocotrienes, lycopenes, etc.	Antioxidant role
PAFome	Platelet activating factor and derivatives	Inflammatory responses
Inositome	Lipids with inositide backbone	Signalling molecules, Ca homeostasis
Sphingome	Lipids with sphingoid backbone	Structural and signalling roles
NAEome	N-acyl ethanolamines, primary amides, N-acylated lipids (e.g., prostamides	N-acyl ethanolamine neuronal signalling via CB1 receptors
Bile acidome	Bile acid species	Cholesterol homeostasis
PLome	Phospholipid species	Structural and signalling roles
Lysolipidome	Lysophospholipid species	Lysophospho/sphingolipids have mitogenic roles and precursor roles
Steroidome	All steroids, including cholesterol	Structural, hormonal, and adaptive roles
NLome	Remaining neutral lipids (waxes, TAG, DAG, MAG)	Storage and barrier roles

7. COMBINED LIPID CLASS MODULES

Another approach for studying lipids in an omic fashion is to study all the lipids that affect a particular process or reside in a particular organ or even intra-cellular location. Examples are shown in Table 4.

Table 4. Combined lipid class modules

Lipid Module	Molecules Studied	Function
Skin Lipidome	Sphingolipids, neutral lipids, eicosanoids, NAE	Roles in inflammation and structural roles
Milk Lipidome	Neutral lipids, phospholipids, glycolipids, gangliosides	Structural and signaling lipids important for health and development
Microbial Lipidome	All classes	Roles in growth, and temperature adaptation. Lipids in pathogenic microbes could become drug targets. Knowledge of the lipids in healthy pro-biotic bacteria could facilitate selection processes
Psychiatric Lipidome	All classes found, particular focus is on gangliosides, anandamide-like molecules, cholesterol, and LC-PUFA	Lipids have structural and signalling roles in the brain, and have been linked to etiology of Alzheimer's, multiple sclerosis, schizophrenia, Parkinson's, adrenoleukodystrophy, stroke, macular degeneration, retinitis pigmentosa, and dyslexia
Macrophageome	See NIH LipidMapp consortium for details	NIH LipidMapp consortium

Abbreviations: NIH, National Institute of Health USA; LC-PUFA, long chain polyunsaturated fatty acid.

8. HOW LIPIDOMIC APPROACHES WILL BENEFIT THE PHARMACEUTICAL INDUSTRY

The science of metabolomics, and its important lipidomic component, is advertised to:

- Discover new drug and nutritional targets through discovery of new metabolites and mechanisms of action; and alternatively identify toxic effects of drugs.
- Examining changes to a wide range of metabolites could uncover new drugable targets. At the same time, hidden toxicities of drugs could be found. Hidden drug toxicity is often missed with current pharmaceutical practices.

- Promote early detection of disease by providing for early novel biomarkers.
- Biomarkers must be reliable and ideally obtainable from circulating fluids rather than biopsied tissue.
- Promote theranostic (diagnostic) kits for early detection of disease.
- It follows that a reliable early biomarker may be readily measured, and thus adapted to kit form.
- Validate and de-validate animal and cell models.
- Cheap, reliable models are extremely important to the drug, cosmetic, and nutrition industry.
- Identify common mechanisms of action.
- By clustering lipid changes statistically, lipidomics can find common mechanisms of actions amongst drug candidates and natural extracts, which may accelerate drug selection and approval processes. This approach could be used to validate that generic drugs behave like their patented counterparts.
- Promote pharmacogenomics (customized medicine).
- Lipidomics can be used as a tool to monitor whether drugs affect lipid metabolism differently in individuals to develop customized medicines.
- Complement genomics (Transcriptomic-lipidomic connectivity).
- Lipidomics is a perfect complement to massively parallel transcriptomic-, knock out-, and knock in approaches. Lipidomics will also be a valuable tool for studying: silent and overt mutations; introduction of transgenes; single-nucleotide polymorphisms (SNPs); evolutionary genomics; and lipid differences amongst cloned animals. Currently, cloned animals suffer from a number of abnormalities, despite having presumed identical genomes.

In Table 5 below, the recent literature has been probed to provide specific examples of how lipidomic approaches are rapidly advancing the goals setforth for metabolomics described above. The references have been selected based on the novelty of the approach, and the importance of the finding, with a focus on recent literature using mass spectral approaches.

Table 5. Summary of representative current lipidomic research

Organization/ Company (Researcher)	Description	Reference
New mechanisms of action of drugs		
Harvard Med. Sc., Boston (CN Serhan)	Hydroxylated docosatrienes discovered while investigating mechanisms of action of aspirin	(Serhan *et al.*, 2002)
Drug toxicity		
Univ. London (JK Nicholson)	FA metabolites are biomarkers for drug toxicity, with NMR	(Coen *et al.*, 2003; Coen *et al.*, 2004; Mortishire-Smith *et al.*, 2004)
Normal vs disease state: understanding changes in Lipids[1]		
Univ. South Carolina (YA Hannun)	Quantitative analysis of endogenous ceramides in lipid extracts, with LC-APCI-MS	(Pettus *et al.*, 2004)
Diabetes and glucose metabolism		
Duke Univ. Durham, NC (C Newgard)	Role of fatty acids in islet studies, by NMR	(Boucher *et al.*, 2004)
Obesity and food intake		
Burke Med. Res. Inst., White Plains, NY (B Kristal)	Profiling of metabolites following food intake manipulation	(Kaddurah-Daouk *et al.*, 2004)
Cancer and tumorigenicity		
Georgia Inst. Tech. (AH Merrill)	Sphingolipids and gangliosides as biomarkers, and mechanisms of action, in gliomas, with ESI-MS-MS	(Sullards *et al.*, 2003)
Atherosclerosis, heart disease, vasoconstriction		
Univ. Cambridge, UK (D Grainger)	Coronary heart disease, with NMR	(Brindle *et al.*, 2002) www.magicad.org.uk
Cholesterol metabolism, lipid transport, lipid-protein binding (animal model validation)		
Numerous	Quantification of individual bile acid species to understand how drugs and diet affect cholesterol absorption, recirculation, conversion to bile acids, and metabolism	(Griffiths, 2003; Mims and Hercules, 2003; Perwaiz *et al.*, 2002)
Nestle Res. Ctr., Lausanne, Switzerland (A Berger)	Absorption and synthesis of dietary cholesterol, studied with on-line GC/combustion and GC/pyrolysis/IRMS MS	(Godin *et al.*, 2004; Gremaud *et al.*, 2001; Pouteau *et al.*, 2003a; Pouteau *et al.*, 2003b; Richelle *et al.*, 2004)
Mental disease, neurotransmission		
Wash. Univ. Sc. Med., St. Louis, MO (X Han)	Determined lipidomes of dorsal root ganglions of wild-type and apoE knockout mice (ApoE may	(Cheng and Han, 2004)

Organization/ Company (Researcher)	Description	Reference
	be linked to Alzheimer's pathogenesis), with MS. ApoE regulated sulfatide metabolism similar to CNS, but also influenced mass content and composition of TAG (potential energy reservoir in the peripheral nervous system)	
Alcoholism		
NIH, Bethesda, MD (HY Kim)	MS was used to monitor effects of EtOH on PS molecular species, and to delineate that EtOH decreased PS synthesis, but not PS degradation, using functional assays; could be used to monitor individual- and species differences in EtOH toxicity	(Wen and Kim, 2004)
Drug-induced oxidation, oxidative stress related disease		
Univ. Colorado (RC Murphy	Free radical oxidation of AA and cholesterol, as occurs during oxidative stress (e.g., pulmonary hypertension, ozone exposure, cerebral edema), may affect disease etiology, with MS	(Bowers *et al.*, 2004; Di Gennaro *et al.*, 2004; Pulfer and Murphy, 2004; Zarini and Murphy, 2003)
Skin disease		
Vanderbilt Univ., Nashville, TN (HA Brown)	Computational algorithms to interpret ESI-MS phospholipid spectra, to understand barrier and cell signaling roles	(Forrester *et al.*, 2004; Ivanova *et al.*, 2004)
Fatty acid β-oxidation disorders		
Womens and Childrens Hosp, North Adelaide, Australia (DW Johnson)	Peroxisomal biogenesis defect patients showed higher levels of hexadecanedioyl (C16DC) carnitine than cystic fibrosis patients and normals. Other carnitine species could aid in distinguishing inherited diseases from generalized dicarboxylic aciduria	(Johnson, 2004)
Inflammatory and immune diseases		
Harvard Med. Sc., Boston,	Novel DAG species perturb	(Bannenberg *et al.*, 2004;

Organization/ Company (Researcher)	Description	Reference
MA (CN Serhan)	signaling in diseases of inflammation, with LC-MS; mechanism of action of plant pathogens; role of lipids in osteoporosis	Gronert *et al.*, 2004; Hong *et al.*, 2003; Serhan, 2004; Serhan and Chiang, 2004)
Ocular disease		
Univ. New Orleans (RB Cole)	Found increased neutral lipids in eye tear samples from persons with dry eye syndrome, with ESI-MS	(Ham *et al.*, 2004)
Lung disease		
Univ. Leipzig, Germany (D Sommerer)	Composition of lung surfactant in different species, with TLC-MALDI-TOF-MS; markers of lung disease?	(Sommerer *et al.*, 2004)
Disease caused by parasitic, bacterial, and viral invasion (mechanism of action, development of better drugs)		
VA Bioinformatics Inst., Blackburg, VA (N Deighton)	Utilization of lipids by malaria parasite, during infestation of RBC	(Deighton *et al.*, 2004)
Nestle Res. Ctr., Lausanne, Switzerland (A Berger)	Understanding biological activity of new milk sources (buffalo milk) by quantification of gangliosides and binding of gangliosides to Cholera toxin and other bacteria	(Colarow *et al.*, 2003)
Transcriptomic-lipidomic connectivity		
Nestle Res. Ctr., Lausanne, Switzerland (A Berger)	Effects of AA and fish oils in mice, with microarrays and GC-FID to study quantitative changes to fatty acids in PL classes	(Berger *et al.*, 2002a)
Understanding role of subcellular organelles		
Wash. Univ. Sc. Med., St. Louis, MO (RW Gross)	Lipid signaling in specific cell locations, with ESI-MS	(Han and Gross, 2003; Pike *et al.*, 2002)
Biomarkers of environmental pollution		
Florida State Univ. (AG Marshall)	Lipid profiling identified 11,000 compounds in crude oil, with FT-MS. Technology could be used to detect pollutants in crude oil and crude oil traces following oil spills	(Hughey *et al.*, 2002)
Lipid consortiums and programs with broad interest in lipidomics		
Max Plank, Germany (CS	General interest in advancing	(Ejsing *et al.*, 2004; Weckwerth

Organization/ Company (Researcher)	Description	Reference
Ejsing, O Fiehn)	instrumentation for metabolomics	*et al.*, 2004)
London Metropolitan Univ. (M Crawford)	Forming lipidomic consortium	http://eoi.cordis.lu/dsp_details.cfm?ID=28372
Kansas State Univ. (R Welti, X Wang)	ESI-MS Service for lipidomic and metabolomic work	(Welti and Wang, 2003)
NIH Lipid Map Consortium, Bethesda, MD	Quantify all lipids in macrophage	http://www.lipidmaps.org/
Imperial College Consortium on Metabonomic Toxicology (COMET), UK	Metabolomics, to uncover hidden toxicity of drugs	(Lindon *et al.*, 2003)
Alliance for Cellular Signaling (AfCS)	Metabolomics, to dissect signal transduction cascades	http://www.signaling-gateway.org/
Austrian Genomics of Lipid-associated Disorders" consortium (W Hofmann)	Expression profiling of mouse models to understand lipid associated disorders (a lipidomic component would benefit this effort)	http://www.bmt.tugraz.at/research/Gen.htm
Univ. Szeged, Hungary (B Penke)	Neuroprotective effects of FA on Alzheimer's disease, with MS	http://www.lipidiet.org/szeged/szeged.htm

[1]Beyond Genomics, Biomol Chenomx, Bristol Meyers Squibb, Eli Lilly, Hoffman-La Roche, Metabolon, Metanomics, METabolic Explorer, Novo Nordisk, Paradigm Genetics, Pfizer, Pharmacia, Phenomenome, Surromed, and TNO are additional companies with metabolomic programs and a general interest in lipids and using metabolomics to uncover drug toxicity.

Abbreviations: NMR, nuclear magnetic resonance; SPE, solid phase extraction; APCI, atmospheric pressure chemical ionisation; TCA, tricarboxylic acid; CL, cardiolipin; ER, endoplasmic reticulum; PC, phosphatidylcholine; PL, phospholipid; FID, flame ionisation detector; FABP, fatty acid binding protein; FA, fatty acid; AA, arachidonic acid; EtOH, ethanol; PS, phosphatidylserine; LDL, low density lipoprotein; PG, prostaglandin; LPC, lysophosphatidylcholine; Cer, ceramide; SPN, sphingomyelin.

9. CONCLUSION

The science of lipidomics is advancing more rapidly than could have been predicted when the term was first coined in the late 1990s by Cytochroma, Inc. (Canada). This is evident by the number of important lipidomic based discoveries described in Table 5. This acceleration in progress is due to two factors: 1) the formation of University and Private Consortiums devoted to advancing lipidomic science; and 2) an explosive growth in mass spectral based approaches to studying lipidomics. Both consortia and equipment manufacturers are eager to share their technology

with the 'masses' (providing they have financial means), and to make the technologies more and more user friendly for newcomers to join the lipidomic band wagon, which will further expedite lipidomic progress.

REFERENCES

Adams A. Metabolomics: small-molecule 'omics. *The Scientist*, 17: 38 (2003).

Amer RK, Pace-Asciak CR, Mills LR. A lipoxygenase product, hepoxilin A(3), enhances nerve growth factor-dependent neurite regeneration post-axotomy in rat superior cervical ganglion neurons *in vitro*. *Neuroscience*, 116: 935-946 (2003).

Bannenberg GL *et al.* Exogenous pathogen and plant 15-lipoxygenase initiate endogenous lipoxin A4 biosynthesis. *J. Exp. Med.*, 199: 515-523 (2004).

Bender F *et al.* Caveolae and caveolae-like membrane domains in cellular signaling and disease: identification of downstream targets for the tumor suppressor protein caveolin-1. *Biol. Res.*, 35: 151-167 (2002).

Berger A *et al.* Dietary effects of arachidonate-rich fungal oil and fish oil on murine hepatic and hippocampal gene expression. *Lipids Health Dis.*, 1: 2 (2002a).

Berger A *et al.* Unraveling lipid metabolism with microarrays: effects of arachidonate and docosahexaenoate acid on murine hepatic and hippocampal gene expression. *Genome Biol* 3: PREPRINT0004, May (2002b)

Berger A, Roberts MA. Dietary effects of arachidonate-rich fungal oil and fish oil on murine hepatic gene expression. In Berger A and Roberts MA (eds), *Unraveling lipid metabolism with microarrays and other "omic" approaches*, Marcel Dekker, NY, 2004.

Boeglin WE, Kim RB, Brash AR. A 12R-lipoxygenase in human skin: mechanistic evidence, molecular cloning, and expression. *Proc. Natl. Acad. Sci. USA*, 95: 6744-6749 (1998).

Boucher A *et al.* Biochemical mechanism of lipid-induced impairment of glucose-stimulated insulin secretion and reversal with a malate analogue. *J. Biol. Chem.*, 279: 27263-27271 (2004).

Bowcock AM *et al.* Insights into psoriasis and other inflammatory diseases from large-scale gene expression studies. *Hum. Mol. Genet.*, 10: 1793-1805 (2001).

Bowers R *et al.* Oxidative stress in severe pulmonary hypertension. *Am. J. Respir. Crit. Care Med.*, 169: 764-769 (2004).

Brindle JT *et al.* Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using ^{1}H-NMR-based metabonomics. *Nat. Med.*, 8: 1439-1445 (2002).

Burstein SH *et al.* Regulation of anandamide tissue levels by N-arachidonylglycine. *Biochem. Pharmacol.*, 64: 1147-1150 (2002).

Capdevila JH, Nakagawa K, Holla V. The CYP P450 arachidonate monooxygenases: enzymatic relays for the control of kidney function and blood pressure. *Adv. Exp. Med. Biol.*, 525: 39-46 (2003).

Cheng H, Han X. The effects of ApoE on the lipidome of mouse peripheral nervous system: A two-dimensional electrospray ionization mass spectrometric study. *Abstract 193, ASMS 2004 Meeting, Nashville, TN, May 23-27*, (2004).

Chu CJ *et al.* *N*-oleoyldopamine, a novel endogenous capsaicin-like lipid that produces hyperalgesia. *J. Biol. Chem.*, 278: 13633-13639 (2003).

Coen M *et al.* An integrated metabonomic investigation of acetaminophen toxicity in the mouse using NMR spectroscopy. *Chem. Res. Toxicol.*, 16: 295-303 (2003).

Coen M *et al.* Integrated application of transcriptomics and metabonomics yields new insight into the toxicity due to paracetamol in the mouse. *J. Pharm. Biomed. Anal.*, 35: 93-105 (2004).

Colarow L *et al.* Characterization and biological activity of gangliosides in buffalo milk. *Biochim. Biophys. Acta*, 1631: 94-106 (2003).

Cowart LA *et al.* The CYP4A isoforms hydroxylate epoxyeicosatrienoic acids to form high affinity peroxisome proliferator-activated receptor ligands. *J. Biol. Chem.*, 277: 35105-35112 (2002).

Daleke DL. Regulation of transbilayer plasma membrane phospholipid asymmetry. *J. Lipid Res.*, 44: 233-242 (2003).

Deighton N *et al.* A metabolomics study of *Plasmodium falciparum* infection of red blood cells in the absence and presence of antimalarials. Abstract 257, *ASMS 2004 Meeting*, May 23-27, (2004).

Di Gennaro A *et al.* Cysteinyl-leukotrienes receptor activation in brain inflammatory reactions and cerebral edema formation: a role for transcellular biosynthesis of cysteinyl-leukotrienes. *FASEB J.*, 18: 842-844 (2004).

Dufour F *et al.* Abnormal cholesterol processing in Alzheimer's disease patients fibroblasts. *Neurobiol. Lipids*, 1: 34 - 45 (2003).

Ejsing CS *et al.* Shotgun Lipidomics: high throughput profiling of the molecular composition of phospholipids. *Oral presentation, ASMS 2004 Meeting, Nashville, TN, May 23-27,* (2004).

Fiehn O. Metabolomics--the link between genotypes and phenotypes. *Plant Mol. Biol.*, 48: 155-171 (2002).

Fisher-Wilson J. Long-suffering lipids gain respect: Technical advances and enhanced understanding of lipid biology fuel a trend toward lipidomics. *The Scientist*, 17: 5 (2003).

Forrester JS *et al.* Computational lipidomics: a multiplexed analysis of dynamic changes in membrane lipid composition during signal transduction. *Mol. Pharmacol.*, 65: 813-821 (2004).

Foster LJ, De Hoog CL, Mann M. Unbiased quantitative proteomics of lipid rafts reveals high specificity for signaling factors. *Proc. Natl. Acad. Sci. USA*, 100: 5813-5818 (2003).

Gascard P *et al.* Asymmetric distribution of phosphoinositides and phosphatidic acid in the human erythrocyte membrane. *Biochim. Biophys. Acta*, 1069: 27-36 (1991).

Godin JP *et al.* [^{2}H/H] Isotope ratio analyses of [$^2H_5$]cholesterol using high-temperature conversion elemental analyser isotope-ratio mass spectrometry: determination of cholesterol absorption in normocholesterolemic volunteers. *Rapid Commun. Mass Spectrom.*, 18: 325-330 (2004).

Gremaud G *et al.* Simultaneous assessment of cholesterol absorption and synthesis in humans using on-line gas chromatography/ combustion and gas chromatography/pyrolysis/isotope-ratio mass spectrometry. *Rapid Commun. Mass Spectrom.*, 15: 1207-1213 (2001).

Griffiths WJ. Tandem mass spectrometry in the study of fatty acids, bile acids, and steroids. *Mass Spectrom. Rev.*, 22: 81-152 (2003).

Gronert K *et al.* A molecular defect in intracellular lipid signaling in human neutrophils in localized aggressive periodontal tissue damage. *J. Immunol.*, 172: 1856-1861 (2004).

Ham BM *et al.* Identification, quantification and comparison of major nonpolar lipids in normal and dry eye tears by ES-MS/MS. *Oral presentation, ASMS 2004 Meeting, Nashville, TN, May 23-27,* (2004).

Han X, Gross RW. Global analyses of cellular lipidomes directly from crude extracts of biological samples by ESI mass spectrometry: a bridge to lipidomics. *J. Lipid Res.*, 44: 1071-1079 (2003).

Hanash S. Disease proteomics. *Nature*, 422: 226-232 (2003).

He F. Measuring metabolic responses of hepatocytes to drug treatment using FTMS. *Presented at Cambridge Healthtech Institute's 2nd Annual Metabolic Profiling: Pathways in Discovery, Durham, North Carolina, Dec 2-3, 2002,* (2002).

Hong MY *et al.* Fish oil increases mitochondrial phospholipid unsaturation, upregulating reactive oxygen species and apoptosis in rat colonocytes. *Carcinogenesis*, 23: 1919-1925 (2002).

Hong S *et al.* Novel docosatrienes and 17S-resolvins generated from docosahexaenoic acid in murine brain, human blood, and glial cells. Autacoids in anti-inflammation. *J. Biol. Chem.*, 278: 14677-14687 (2003).

Huang SM *et al.* An endogenous capsaicin-like substance with high potency at recombinant and native vanilloid VR1 receptors. *Proc. Natl. Acad. Sci USA*, 99: 8400-8405 (2002).

Hughey CA, Rodgers RP, Marshall AG. Resolution of 11,000 compositionally distinct components in a single electrospray ionization Fourier transform ion cyclotron resonance mass spectrum of crude oil. *Anal. Chem.*, 74: 4145-4149 (2002).

Ivanova PT *et al.* LIPID Arrays: New tools in the understanding of membrane dynamics and lipid signaling. *Mol. Interv.*, 4: 86-96 (2004).

Johnson DW. Deuterium labeled dicarboxylic acylcarnitines for the differentiation of fatty acid oxidation disorders by tandem mass spectrometry. *Abstract 44, ASMS 2004 Meeting, Nashville, TN, May 23-27,* (2004).

Kaddurah-Daouk R *et al.* Bioanalytical advances for metabolomics and metabolic profiling. *PharmaGenomics*, Jan: 46-52 (2004).

Lindon JC *et al.* Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project. *Toxicol. Appl. Pharmacol.*, 187: 137-146 (2003).

Mann M, Jensen ON. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.*, 21: 255-261 (2003).

Mims D, Hercules D. Quantification of bile acids directly from urine by MALDI-TOF-MS. *Anal. Bioanal. Chem.*, 375: 609-616 (2003).

Mortishire-Smith RJ *et al.* Use of metabonomics to identify impaired fatty acid metabolism as the mechanism of a drug-induced toxicity. *Chem. Res. Toxicol.*, 17: 165-173 (2004).

Mutch DM *et al.* An integrative metabolism approach identifies stearoyl-CoA desaturase as a target for an arachidonate-enriched diet. *FASEB J. (In press).*

Parton RG. Caveolae--from ultrastructure to molecular mechanisms. *Nat. Rev. Mol. Cell. Biol.*, 4: 162-167 (2003).

Perwaiz S *et al.* Rapid and improved method for the determination of bile acids in human feces using MS. *Lipids*, 37: 1093-1100 (2002).

Pettus BJ *et al.* Quantitative measurement of different ceramide species from crude cellular extracts by normal-phase high-performance liquid chromatography coupled to atmospheric

pressure ionization mass spectrometry. *Rapid Commun. Mass Spectrom.*, 18: 577-583 (2004).

Phelps TJ, Palumbo AV, Beliaev AS. Metabolomics and microarrays for improved understanding of phenotypic characteristics controlled by both genomics and environmental constraints. *Curr. Opin. Biotechnol.*, 13: 20-24 (2002).

Pike LJ *et al.* Lipid rafts are enriched in arachidonic acid and plasmenylethanolamine and their composition is independent of caveolin-1 expression: a quantitative electrospray ionization/mass spectrometric analysis. *Biochemistry (Mosc)*, 41: 2075-2088 (2002).

Pouteau E *et al.* Determination of cholesterol absorption in humans: from radiolabel to stable isotope studies. *Isotopes Environ. Health Stud.*, 39: 247-257 (2003a).

Pouteau EB *et al.* Non-esterified plant sterols solubilized in low fat milks inhibit cholesterol absorption--a stable isotope double-blind crossover study. *Eur. J. Nutr.*, 42: 154-164 (2003b).

Pulfer MK, Murphy RC. Formation of biologically active oxysterols during ozonolysis of cholesterol present in lung surfactant. *J. Biol. Chem.*, 279: 26331-26338 (2004).

Richelle M *et al.* Both free and esterified plant sterols decrease cholsterol absorption and the bioavailability of β-carotene and α-tocopherol, in normocholesterolemic humans. *Am. J. Clin. Nutr.*, 80: 171-177 (2004).

Schnitzer JE. Caveolae: from basic trafficking mechanisms to targeting transcytosis for tissue-specific drug and gene delivery *in vivo*. *Adv. Drug Deliv. Rev.*, 49: 265-280 (2001).

Serhan CN. Clues for new therapeutics in osteoporosis. *N. Engl. J. Med.*, 350: 1902-1903 (2004).

Serhan CN. Lipoxins and aspirin-triggered 15-epi-lipoxin biosynthesis: an update and role in anti-inflammation and pro-resolution. *Prostaglandins Other Lipid Mediat.*, 68-69: 433-455 (2002).

Serhan CN, Chiang N. Novel endogenous small molecules as the checkpoint controllers in inflammation and resolution: entree for resoleomics. *Rheum. Dis. Clin. North Am.*, 30: 69-95 (2004).

Serhan CN *et al.* Resolvins: a family of bioactive products of omega-3 fatty acid transformation circuits initiated by aspirin treatment that counter proinflammation signals. *J. Exp. Med.*, 196: 1025-1037 (2002).

Sommerer D *et al.* Analysis of the phospholipid composition of bronchoalveolar lavage (BAL) fluid from man and minipig by MALDI-TOF mass spectrometry in combination with TLC. *J. Pharm. Biomed. Anal.*, 35: 199-206 (2004).

Sullards MC *et al.* Metabolomic profiling of sphingolipids in human glioma cell lines by liquid chromatography tandem mass spectrometry. *Cell. Mol. Biol,*. 49: 789-797 (2003).

Sumner SCJ, Liu G. Metabolomics holds key to intelligent discovery efforts. *Genetic Engineering News*, 22: 32-33 (2002).

Varnau M, Singhania A. Dynamic metabolomics industry emerges. *Genetic Engineering News*, 22: 15-17; 93 (2002).

Walker JM, Huang SM. Endocannabinoids in pain modulation. *Prostaglandins Leukot Essent Fatty Acids*, 66: 235-242 (2002).

Watkins SM. Lipomic profiling in drug discovery, development and clinical trial evaluation. *Curr. Opin. Drug Discov. Devel.*, 7: 112-117 (2004).

Watkins SM, German JB. Metabolomics and biochemical profiling in drug discovery and development. *Curr. Opin. Mol. Ther.*, 4: 224-228 (2002).

Weckwerth W, Fiehn O. Can we discover novel pathways using metabolomic analysis? *Curr. Opin. Biotechnol.*, 13: 156-160 (2002).

Weckwerth W, Wenzel K, Fiehn O. Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. *Proteomics*, 4: 78-83 (2004).

Welti R, Wang X. Lipidomics. *Inform.*, 14: 607-608 (2003).

Wen Z, Kim H-Y. Alterations in hippocampal phospholipid profile by prenatal exposure to ethanol. *J. Neurochem.*, 89: 1368-1377 (2004).

Wulfkuhle JD, Liotta LA, Petricoin EF. Proteomic applications for the early detection of cancer. *Nat. Rev. Cancer*, 3: 267-275 (2003).

Yang W. Lipomics: mastering metabolites. *Biocentury, The Bernstein report on BioBusiness* A13 (2003).

Zarini S, Murphy RC. Biosynthesis of 5-oxo-6,8,11,14-eicosatetraenoic acid from 5-hydroperoxyeicosatetraenoic acid in the murine macrophage. *J. Biol. Chem.*, 278: 11190-11196 (2003).

Ziboh VA *et al.* Biological significance of essential fatty acids/prostanoids/lipoxygenase-derived monohydroxy fatty acids in the skin. *Arch. Pharm. Res.*, 25: 747-758 (2002).

Chapter 21

METABOLITES AND FUNGAL VIRULENCE

An integrated perspective on pathogenic fungal physiology

Edward M. Driggers[1] and Axel A. Brakhage[2]

[1]Microbia, Inc., 320 Bent St., Cambridge, MA 02141, edriggers@microbia.com ; [2]Institute of Microbiology, University of Hannover, Schneiderberg 50, D-30167, Hannover, Germany.

1. ANTIFUNGAL DRUGS AND THE EVOLVING DEMOGRAPHICS OF INFECTION

The history and current status of anti-fungal drug development has been recently reviewed in the literature (Odds *et al.*, 2003). However, it is worth highlighting some specifics of that history. The current status of anti-fungal drug development suggests a practical role for metabolite profiling in this field, and forms a useful framework for considering fungal virulence, especially as it relates to secondary metabolism and the profiling of secondary metabolites.

A healthy person typically encounters pathogenic fungi only as localized or epidermal infections with minor impact. However, the growing number of immunosuppressed patients during the last three decades has greatly increased the frequency of far more serious systemic mycoses with very high rates of mortality (up to 80% for invasive aspergillosis). Treatment of these infections has been primarily through the use of either amphotericin B (a polyene), or with one of the azole-based compounds, such as fluconazole (Figure 1); despite limitations, both are now well-established therapeutic approaches.

The azoles attack the fungal cell wall by inhibiting the biosynthesis of ergosterol, an abundant sterol in the fungal cell membrane. The compounds have specificity for Erg11P, a P-450 enzyme which serves to demethylate the pathway intermediate lanosterol at the 14α position; nitrogen in the imidazole or triazole ring serves as a disruptive ligand for the protoporphyrin

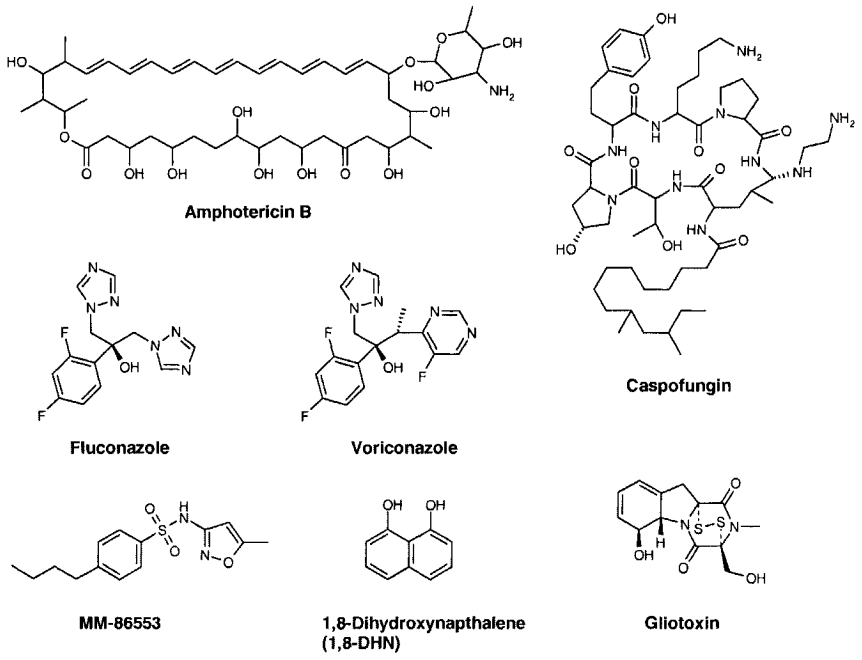


Figure 1. Antifungal agents and noteworthy fungal secondary metabolites

iron cofactor in the enzyme. The mechanism of action for amphotericin B is through the same pathway; however the polyene serves to bind the ergosterol molecule itself. Both classes of compounds thus have their effect by perturbing the function of the fungal cell membrane.

Amphotericin B exhibits toxic effects on mammalian cells, as well as significant nephrotoxicity in the clinic. Its toxicity appears to be mediated through binding of cholesterol, a mammalian sterol for which it also has affinity. The toxicity of amphotericin B and other antifungal compounds has been a significant factor in the drive to develop improved therapeutics for systemic mycoses.

Amphotericin B exhibits broad spectrum of action against fungal species, including very difficult to treat invasive aspergillosis. The absolute and relative frequency of aspergillosis mortality has been on the rise for a number of years (Steinbach *et al.*, 2003; Odds *et al.*, 2003) due to multiple factors, including aggressive tumor treatments, increased organ transplants, and the success of more recent AIDS therapies that have reduced the frequency of systemic *Candida spp.* infections. Consequently, clinicians continue to use amphotericin B to treat systemic fungal infections. Recently,

new antifungal compounds have become available, including second generation azoles (e.g. voriconazole) and echinocandins (e.g. caspofungin) (Figure 1), which function by inhibiting synthesis of β-1,3-glucan polysaccharides in the cell wall. These new compounds show promise against aspergillosis, however limitations persist, including drug-drug interactions (azoles), intravenus-only formulations (echinocandins), and incomplete efficacy (azoles and echinocandins) – even with the newest treatments, mortality rates from systemic fungal infections remain unacceptably high. Therefore researchers continue to seek new therapeutic strategies and targets to either kill or suppress the virulence of pathogenic fungi. One such approach aims to inhibit key, evolutionarily-conserved (and therefore broad spectrum) pathways that are essential for the development of pathogenic morphology in infective fungi. This strategy has lead to the identification of a class of molecules known as "Anti-Invasins" (eg. MM-86553, Figure 1) (Summers *et al.*, 2003; Mayorga *et al.*, 2003).

Interestingly, from the perspective of metabolite profiling, *Aspergillus fumigatus* is a fungus with a rich and complex secondary metabolism. Over eighty distinct secondary metabolites are associated with the species (Buckingham, 2001), as compared with three for *C. albicans* (note also that *A. terreus*, a more recently emerging cause of aspergillosis (Baddley *et al.*, 2003; Mosquera *et al.*, 2002), is known to produce over one hundred secondary metabolites under various conditions). A number of these secondary metabolites have been identified as virulence factors (in a variety of other pathogenic fungi as well; see below). As with many secondary metabolites, the production of these complex structures is linked closely with developmental steps such as conidiation (Calvo *et al.*, 2002; Demain and Fang, 2000), which are also central to the progression of invasive aspergillosis. The evolving demographics of these challenging fungal infections have therefore created a circumstance where careful application of metabolite profiling technology may provide timely and much needed insight into the physiology of fungal disease, particularly when applied in conjunction with other biochemical and genomic tools.

2. METABOLITES AND BIOSYNTHETIC PATHWAYS INVOLVED IN VIRULENCE

Fungal secondary metabolites can contribute directly to the virulence of the producing fungi. The following section discusses metabolites known to be required for virulence, as well as others whose role in virulence is still being determined. A number of these are products of *A. fumigatus*; however the products of other pathogenic fungi are discussed as well. These

metabolites and pathways form a potential core focus for metabolite profiling studies of fungal pathogenesis.

2.1 Pigments

Fungi produce a variety of pigments, many of which contain melanin. It has been known since the early 1960s that melanins exist in fungi. However, more recent understanding indicates that they can play an important role in fungal pathogenesis. In phytopathogens such as *Colletotrichum lagenarium* and *Magnaporthe grisea*, melanins are essential for infectivity as they allow the enormous pressures to build in appressoria that enable the fungus to penetrate plant leaves (reviewed in Howard and Valent, 1996; Money, 1997; Thines *et al.*, 2000). In human pathogenic fungi such as *Cryptococcus neoformans, A. fumigatus,* and others, melanins are thought to protect the pathogen from the immune system, although a mechanical role has yet to be elucidated.

In general, melanins are macromolecules formed by oxidative polymerization of phenolic or indolic compounds. In fungi several different types of melanins have been identified to date. The two most important types are DHN-melanin (named after one of the pathway intermediates, 1,8-dihydroxynaphthalene) and DOPA-melanin (named after one of the precursors, L-3,4-dihydroxyphenylalanine). Both types of melanin have been implicated in pathogenesis (Jacobson, 2000; Langfelder *et al.*, 2003; Brakhage and Jahn, 2002; Haase and Brakhage, 2004). While this chapter will focus on the DHN-melanin pigments, it is worth noting that the human pathogen *C. neoformans* produces pigments based on the DOPA-melanin pathway, which also appear to be involved in pathogenesis (reviewed in Jacobson, 2000; Langfelder *et al.*, 2003).

2.1.1 The DHN-melanin biosynthesis pathway

The canonical DHN-melanin biosynthesis pathway (Figure 2) is derived from genetic and biochemical evidence obtained from *Verticillium dahliae* and *Wangiella dermatitidis* (Wheeler and Bell, 1988).

The polyketide synthase (PKS) converts malonyl-CoA to the first detectable intermediate of the pathway, 1,3,6,8-tetrahydroxynaphthalene (1,3,6,8-THN). Following this, 1,3,6,8-THN is reduced by a specific reductase enzyme to produce scytalone. It was discovered that a specific reductase inhibitor, tricyclazole, produced the same defect as a mutation in the reductase gene, namely the accumulation of flaviolin, a shunt product of

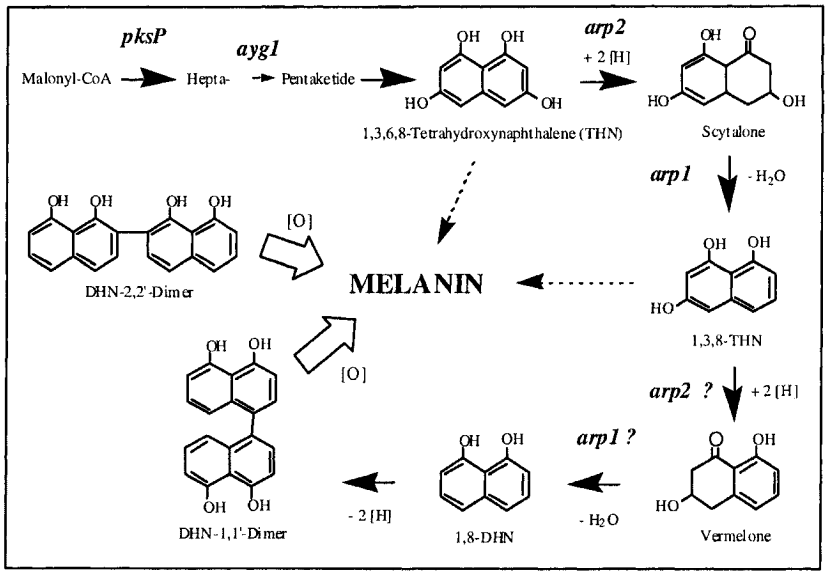


Figure 2. Dihydroxynaphthalene biosynthesis pathway of *Aspergillus fumigatus*.

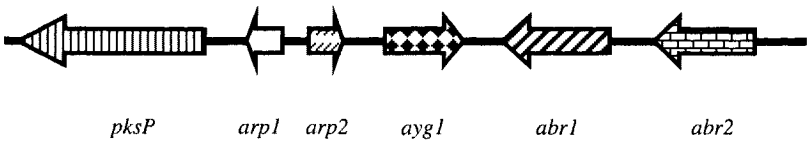


Figure 3. Dihydroxynaphthalene biosynthesis gene cluster of *Aspergillus fumigatus*.

1,3,6,8-THN. Scytalone is dehydrated enzymatically to 1,3,8-trihydoxynaphthalene, which is in turn reduced, possibly by a second reductase, to vermelone. This reductase can also be inhibited by tricyclazole. A further dehydration step, possibly also catalysed by scytalone dehydratase, leads to the intermediate 1,8-dihydroxynaphthalene (DHN), after which this pathway was named. Subsequent steps are thought to involve a dimerization of the 1,8-DHN molecules, followed by polymerisation, possibly catalyzed by a laccase. This is a general model for DHN-melanin biosynthesis but the pathway may vary in individual fungi, e.g. in *A. fumigatus* (see below). Interestingly, several by-products of the fungal DHN-melanin pathway have been shown to have antibacterial or immunosuppressive properties. While in many other fungi the structural genes are distributed throughout the genome, a cluster of six genes was discovered in *A. fumigatus*, all shown to be involved in DHN-melanin biosynthesis (Figure 3) (Langfelder *et al.*, 1998;

Tsai *et al.*, 1998; Tsai *et al.*, 1999; Langfelder *et al.*, 2003). The *pksP* gene encodes a type I polyketide synthase. *pksP* mutants of *A. fumigatus* have white conidia while the wild-type conidia are gray-green in color. For PKSP of *A. fumigatus*, Fujii *et al.* (2000) demonstrated that the polyketide formed was YWA1, a heptaketide naphthopyrone, which has a slightly different structure from 1,3,6,8-THN, which is observed in the canonical pathway (Figure 2). This result was unexpected since only 1,3,6,8-THN but not naphthopyrone (YWA1) could be detected in *A. fumigatus* by TLC chromatography. This result could be explained when Tsai *et al.* (2001) found that the product of the *A. fumigatus ayg1* gene, AYG1, is able to convert YWA1 (i.e., the product of either WA or PKSP) to 1,3,6,8-THN, by chain shortening, a reaction which apparently does not occur in the canonical pathway.

The second step in DHN-melanin biosynthesis involves a reduction of the 1,3,6,8-THN to scytalone, which is catalysed by a reductase enzyme. Such enzymes have been described in several fungi including *A. fumigatus* Previously, Tsai *et al.* (1999) had shown that *A. fumigatus* must also possess a second reductase gene, not present in the cluster, since an *arp2* deletion strain growing on agar plates did not accumulate 2-hydroxyjugalone (2-HJ, the shunt product of 1,3,8-trihydroxy naphthalene) when the strain was supplemented with scytalone. When the reductase inhibitor tricyclazole was added to cultures, in addition to scytalone, the shunt product 2-HJ was observed in abundance. However, no additional reductase has been found to date in *A. fumigatus*.

The *arp1* gene encodes scytalone dehydratase which catalyses the dehydration of scytalone to 1,3,8-trihydroxynaphthalene (Tsai *et al.*, 1997). Deletion of this gene resulted in *A. fumigatus* colonies with pink conidia. A laccase-encoding gene has been found in *A. fumigatus* (*abr2*) but it is not yet clear at which point of the pathway it is involved (Tsai *et al.*, 1999). In summary, many of the initial steps of DHN-melanin biosynthesis are well understood, but most of the later reaction steps require further investigation. The involvement of the latter two genes in the formation of the gray-green spore color of *A. fumigatus* was shown by disruption of each gene which led to altered conidial color phenotypes (Tsai *et al.*, 1999). It is possible that homologues of these genes can also be found in other fungi.

2.1.2 Melanin in pathogenesis

The proposed functions of fungal melanins include protection against UV irradiation, enzymatic lysis, oxidants, and in some instances extremes of temperatures. Also, melanins have been shown to bind metals, function as a physiological redox buffer, and thereby possibly acting as a sink for harmful

unpaired electrons. They provide structural rigidity to cell walls, and store water and ions, thus helping to prevent desiccation (reviewed by Butler and Day, 1998; Jacobson, 2000).

For most pathogenic fungi, including *W. dermatitidis*, *A. fumigatus*, *S. schenckii* and in dematiaceous fungi in general, the DHN melanin pathway represents a significant factor in fungal infectivity (reviewed by Haase and Brakhage, 2004). A number of results indicate that both DHN melanin and DOPA melanin are able to quench reactive oxygen species (ROS), one of the defense mechanisms of the human immune system and also able to prevent phagocytosis to some degree. In *A. fumigatus*, complete absence of DHN-melanin, as in the case of *pksP* mutants, resulted in a severe reduction in virulence. The *pksP* mutant of *A. fumigatus* is significantly more sensitive to hydrogen peroxide and sodium hypochlorite than the wild-type strain. Also, the mutant strain is more susceptible to damage by murine macrophages *in vitro*. As in other cases, it could be shown that DHN-melanin is able to quench ROS derived from human granulocytes (Jahn *et al.*, 1997; 2000). These results indicated that conidial DHN-melanin of *A. fumigatus* is involved in protecting conidia from the host immune response in which ROS are important for eliminating fungal conidia (Latgé, 1999; Langfelder *et al.*, 2003).

However, although the green pigment of the non-pathogenic *A. nidulans* conidia does not seem to be synthesized via the DHN-melanin pathway, it also acts as protective agent against oxidant based host defence mechanisms and thus contributes to the relative resistance of conidia against neutrophil attack, as described for *A. fumigatus*. Therefore, resistance against ROS does not explain in and of itself why *A. fumigatus* conidia can be pathogenic whereas this is rarely the case for *A. nidulans* conidia. An attractive hypothesis is that besides the pigment the *pksP* gene product of *A. fumigatus* is involved in the production of another compound which is immunosuppressive. This hypothesis is further supported by the notion that the presence of a functional *pksP* gene in *A. fumigatus* conidia is associated with an inhibition of the fusion of phagosomes and lysosomes in human MDM (Jahn *et al.*, 2002). Other pathways involving polyketide synthases have been shown to synthesize two different active products (reviewed in Langfelder *et al.*, 2003).

In *C. neoformans* the production of melanin and other virulence factors is regulated via the cAMP-signaling pathway (reviewed in Lengeler *et al.*, 2000). A similar link between cAMP-dependent signaling and virulence was also found for *A. fumigatus* (Rhodes *et al.* 2001; Liebmann *et al.* 2003).

However, the presence of melanins *per se* does not define a human pathogenic fungus because several non-pathogenic fungi also are able to synthesize melanins. Therefore, additional virulence factors are required

such as the ability of the fungus to grow at 37 °C, or possibly the production of melanins at specific stages of the infectious process or the presence of melanins in certain cell types or organs like conidia and appressoria, respectively.

2.2 Gliotoxin

A. fumigatus is also known to produce a secondary metabolite named gliotoxin (Figure 1). Gliotoxin was shown to have severe effects on different activities required for a fully active immune system, e.g. gliotoxin inhibits phagocytosis of macrophages at concentrations of 20-50 ng/ml. Also, B-cell activation was blocked (Sutton *et al.*, 1994). Re-treatment of normally resistant mice with a single injection of a sublethal dose of gliotoxin was sufficient to make them susceptible to infection and subsequent death, after challenge with *A. fumigatus* conidia. Moreover, animals infected with a non-gliotoxin producing strain survived significantly longer than those infected with a gliotoxin producer (Sutton *et al.*, 1996). Gliotoxin prevented the onset of O_2^- generation by the human neutrophil NADPH oxidase in response to PMA (Yoshida *et al.*, 2000). Gliotoxin markedly inhibited both perforin-dependent and Fas ligand-dependent cytotoxic T-lymphocyte (CTL)-mediated cytotoxicity (Yamada *et al.*, 2000). Interestingly, gliotoxin specifically inhibited transcription factor NF-kappaB (Pahl *et al.*, 1996). Evidence for the necessity of gliotoxin during the infectious process by rigorous genetic analyses, e.g., by deletion of a biosynthesis gene involved, has not been provided yet.

3. SYSTEMS LEVEL ANALYSIS OF METABOLIC AND TRANSCRIPTIONAL PROFILES

Methods to assess gene expression and metabolite levels on a genomic scale provide the opportunity to correlate patterns of global gene expression with the production of specific metabolites. It is clear from the discussion in Section 2 that fungal virulence is an integrated "metabolo-genomic process": Extracellular signals result in cascades that up-regulate genes, which produce secondary metabolites, which in turn modulate the extracelluar environment, (i.e. the host). This section describes a model study aimed at deciphering the complex inter-relationships between metabolite production trends and gene expression events, and suggests how information gleaned from such studies can be used to investigate subtleties of fungal physiology. Association analysis of transcript and metabolite profiles taken from the

same engineered strains of *A. terreus* was used to determine gene expression patterns that correlate with the yield of lovastatin and (+)-geodin (Figure 4), two secondary metabolites produced by the filamentous fungus, which constitute a simple, model metabolite profile. Lovastatin is a potent hydroxymethylglutaryl coenzyme A (HMG-CoA) reductase inhibitor (Endo *et al.*, 1976) that is used clinically to reduce serum cholesterol levels. (+)-Geodin is derived from the anthraquinone emodin, an intermediate in the biosynthesis of many natural products. It is important to keep in mind that these studies were executed on genetically engineered strains cultured *in vitro*, however given the growing role of *A. terreus* in human infection (eg. Baddley, *et al.*, 2003), this study can be treated as a model for the systems level investigation of fungal pathophysiology.

3.1 Metabolite and gene expression data sets

In order to perform association analysis, we required profiling data sets in which the levels of metabolite(s) and global gene expression patterns vary. To generate diversity, a collection of *A. terreus* strains was engineered to produce lovastatin at varying titers by transformation with a variety of fungal regulatory proteins (Askenazi *et al.*, 2003). Secondary metabolite levels produced by the strains were analyzed by high-pressure liquid chromatography-electrospray mass spectrometry (LC-MS). In addition to lovastatin and related monacolins, secondary metabolite profiling identified a variety of (+)-geodin related compounds, with (+)-geodin itself being the most abundant secondary metabolite in broths from control strains. Quantitative lovastatin and (+)-geodin yields from engineered strains were determined relative to levels from appropriate reference strains using a simplified HPLC assay focusing specifically on the two metabolites of interest.

To identify gene expression patterns that correlate with the production of these metabolites, representative transformants from each set of manipulated strains and appropriate reference strains were used to generate transcriptional profiles. Since limited sequence information is available for the *A. terreus* genome, we monitored genome-wide expression patterns using a genomic fragment microarray of 21,000 elements, providing approximately 88% coverage.

Hierarchical clustering (average linkage with Pearson correlation coefficients) of the transcriptional profiling data sets shows that strains that display similar metabolite profiles are significantly more related to each other based upon transcriptional data as well (Figure 4). For example, strains that produce high levels of lovastatin and decreased levels of (+)-geodin

cluster together and separately from strains that produce decreased levels of both metabolites.

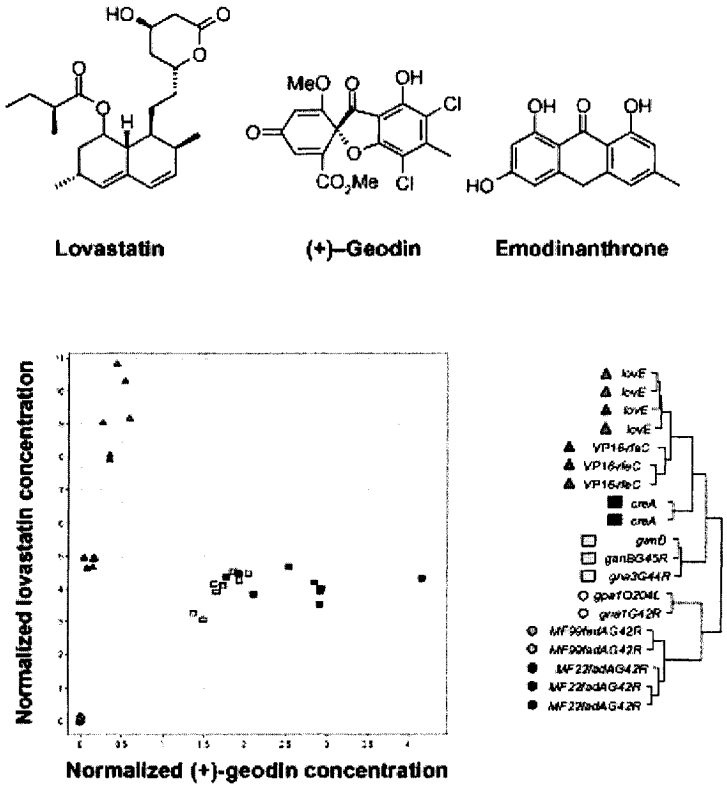


Figure 4. a) Metabolite structures. b) Scatter plot of normalized metabolite titers compared with hierarchical clustering of the transcriptional profiling datasets

3.2 Association analysis

To quantify these observed clustering relationships, association analysis was performed using the combined metabolic and transcriptional data sets in order to identify genes with expression patterns that correlate specifically with secondary metabolite production. Secondary metabolite and gene expression values were expressed as ratios that reflect a value from an engineered strain relative to that of a reference strain. Two statistical approaches were subsequently employed to define the relationships between

gene(s) present on hybridizing elements and secondary metabolite levels: Pearson product-moment correlation coefficients were calculated from transcriptional profiling ratio values and metabolite ratios, as well as association according to Goodman and Kruskal's gamma (Agresti, 1990; Goodman and Kruskal, 1954), using the same ratios binned into categories of up, down, and unchanged (ordinal). For these data sets, measures of association that use either ordinal or continuous data representations converge on a common set of elements, and sequence information was obtained for many microarray elements showing expression patterns that significantly associated with lovastatin and/or (+)-geodin production.

3.3 Identification of biosynthetic clusters and metabolic trends

This approach enabled the rapid identification of genes required for biosynthesis of these secondary metabolites. The *A. terreus* lovastatin biosynthetic cluster is a 64 kb genomic region predicted to encode 18 proteins, a subset of which are known to be required for lovastatin production (Hutchinson *et al.*, 2000); this cluster therefore represented a control, using genes and metabolites already known to be associated with each other. Array elements containing *lovA*, *lovB*, *lovC*, *lovD*, *lovF*, *lvrA*, and multiple open reading frames were identified by this approach to be positively associated with lovastatin production; the independent discovery of the regulated lovastatin biosynthetic genes by association analysis nicely validated the method.

In addition, the approach sheds light upon the biosynthesis of (+)-geodin, a less studied molecule, serving here as a representative of genes and metabolites with unknown associations. Association analysis identified the previously unknown polyketide synthase (PKS) required for (+)-geodin production (the emodinanthrone PKS), demonstrated that expression of a known (+)-geodin biosynthetic gene, encoding the dihydrogeodin oxidase, correlates with (+)-geodin production, and predicted several novel (+)-geodin biosynthetic genes (Curtis *et al.*, 1972; Fujii *et al.*, 1987; Fujimoto *et al.*, 1975; Gatenbeck and Malmstrom, 1969). For the identification of the PKS required for (+)-geodin production, the combination of observed association scores, protein sequence homology to a known PKS class, and chemical similarities to other related polyketide metabolites led to the prediction that several contiguous (+)-geodin-associated array elements encode the emodinanthrone PKS. These elements show significant homology to filamentous fungal enzymes required for pigment biosynthesis (Mayorga and Timberlake, 1992; Fulton *et al.*, 1999). These pigmented natural products are non-reduced fungal polyketides (Bingle *et al.*, 1999;

Nicholson *et al.*, 2001), and the chemical structure of emodinanthrone, a (+)-geodin precursor, clearly defines it as a member of this class. The function of the identified PKS was verified by gene disruption studies.

Association analysis further identified many genes that encode proteins either predicted or known to play a role in the production of secondary metabolites other than lovastatin and (+)-geodin. In addition, analysis of gene expression patterns that correlate generally with metabolite production provides insight into the physiological states that promote the biosynthesis of those secondary metabolites. For example, a collection of genes expected to be expressed during growth phase, or involved in the generation of ATP (e.g., glycolytic and tricarboxylic acid enzymes, proteins involved in oxidative phosphorylation) are present on elements that negatively correlate with secondary metabolite production.

4. OUTLOOK

The examples presented in this chapter summarize only briefly the current state of knowledge regarding the pathways for biosynthesis of pathogenic fungal secondary metabolites. Similarly, the model study of integrated transcriptional-metabolite profiling in *A. terreus* represents only a limited application of the current suite of metabolite profiling technologies: central metabolites, flux values, and the large number of additional *A. terreus* secondary metabolites are all ignored for the sake of clarity and demonstration. Despite these simplifications, one can readily extrapolate to the types of integrated studies that will shed light on the complex physiology of fungal pathogenesis. For example, profiling studies executed with fungal biomass cultured *in vivo* as part of animal infection model can provide even more information regarding the specific physiology of pathogenesis. Results from an *in vivo* metabolite profiling study using a murine model of filamentous fungal infection showed a wide variety of secondary metabolites to be detectable in the infected tissue that are also abundant in *in vitro* cultures of the fungus. Fully integrated *in vivo* profiling experiments in the future will hopefully provide useful information for the development of therapeutics targeted against specific features of pathogenic fungal physiology.

The pharmaceutical industry continues the effort to discover novel antifungal therapeutics that overcome the toxicity of the current treatments such as amphotericin B, and simultaneously to extend their spectrum of action to newly emerging pathogens such as *A. terreus*. Metabolite profiling is positioned to contribute in a unique way to this effort, integrating the body of existing knowledge regarding metabolic virulence factors with new

discoveries regarding the genetically coordinated production of those factors.

REFERENCES

Agresti A. Categorical Data Analysis. John Wiley and Sons, New York (1990).

Askenazi M *et al.* Integrating transcriptional and metabolite profiles to direct the engineering of lovastatin producing fungal strains. *Nat. Biotechnol.*, 21: 150-156 (2003).

Baddley JW *et al.* Epidemiology of Aspergillus Terreus at a University Hospital. *J. Clinical Microbiol.*, 41:5525-5529 (2003).

Bingle LEH, Simpson TJ, Lazarus CM. Ketosynthase domain probes identify two subclasses of fungal polyketide synthase genes. *Fungal Genet. Biol.*, 26: 209-223 (1999).

Brakhage AA, Jahn B. Molecular mechanisms of pathogenicity of *Aspergillus fumigatus*. In: *Molecular Biology of Fungal development*. Osiewacz HD (Ed.), Marcel Dekker Inc., pp. 559-582 (2002).

Buckingham J. *Dictionary of Natural Products on CD-ROM*, vol. 10:1, Chapman and Hall/CRC Press, Boca Raton, FL (2001).

Butler MJ, Day AW. Fungal melanins: a review *Can. J. Microbiol.*, 44: 1115-1136 (1998).

Calvo AM, Wilson RA, Bok JW, Keller NP. Relationship between secondary metabolism and fungal development. *Microbiol. Molec. Biol. Rev.*, 66: 447-459 (2002).

Curtis RF, Hassal CH, Perry DR. The biosynthesis of phenols XXIV. The conversion of the anthraquinone question into the benzophenone, sulochrin, in cultures of *Aspergillus terreus*. *J. Chem. Soc. Perkin Trans. I*, 2: 240-244 (1972).

Demain AL, Fang A. The natural functions of secondary metabolites. *Adv. Biochem. Eng./Biotechnol.*, 69: 1-39 (2000).

Endo A, *et al.* Competitive inhibition of 3-hydroxy-3-methylglutaryl coenzyme A reductase by ML236A and ML236B, fungal metabolites having hypocholesterolemic activity. *FEBS Lett.*, 72:323-326 (1976).

Fujii I, *et al.* Purification and properties of dihydrogeodin oxidase from *Aspergillus terreus*. *J. Biochem.*, 101:11-18 (1987).

Fujii I, Mori Y, Watanabe A, Kubo Y, Tsuji G, Ebizuka Y. Enzymatic synthesis of 1,3,5,8-tetrahydroxynaphthalene solely from malonyl coenzyme A by a fungal iterative type I polyketide synthase PKS1. *Biochem.*, 39: 8853-8858 (2000).

Fujimoto H, Flash H, Franck B. Biosyntheses der seco-anthrachinone geodin und dihydrogeodin aus emodin. *Chem. Ber.*, 108: 752-753 (1975).

Fulton TR *et al.* A melanin polyketide synthase (PKS) gene from *Nodulisporium* sp. that shows homology to the *pks1* gene of *Colletotrichum lagenarium*. *Mol. Gen. Genet.*, 262: 714-720 (1999).

Gatenbeck S, Malmstrom L. On the biosynthesis of sulochrin. *Acta Chem. Scand.*, 23: 3493-3497 (1969).

Goodman LA, Kruskal WH. Measures of association for cross classifications. *J. Am. Stat. Assoc.*, 49: 732-764 (1954).

Haase G, Brakhage AA. Melanized fungi infecting humans. Function of melanin as a pathogenicity factor. In: The Mycota. Domer JE, Kobayashi GS (Eds.) Vol. XII, *Human Fungal Pathogens*. Springer Verlag, pp. 67-88 (2004).

Howard RJ, Valent B. Breaking and entering: host penetration by the fungal rice blast pathogen *Magnaporthe grisea*. *Annu. Rev. Microbiol.*, 50: 491-512 (1996).

Hutchinson CR *et al.* Aspects of the biosynthesis of non-aromatic fungal polyketides by iterative polyketide synthases. *Antonie Van Leeuwenhoek* ,78: 287-295 (2000).

Jacobson ES. Pathogenic roles for fungal melanins. *Clin. Microbiol. Rev.*, 13: 708-717 (2000).

Jahn B, Koch A, Schmidt A, Wanner G, Gehringer H, Bhakdi S, Brakhage, AA. Isolation and characterisation of an *Aspergillus fumigatus* mutant strain with pigmentless conidia and reduced virulence. *Infect. Immun.*, 65: 5110-5117 (1997).

Jahn B, Boukhallouk F, Lotz J, Langfelder K, Wanner G, Brakhage, AA. Interaction of human phagocytes with pigmentless conidia. *Infect. Immun.*, 68: 3736-3739 (2000).

Jahn B, Langfelder K, Schneider U, Schindel C, Brakhage AA. PKSP dependent reduction of phagolysosome fusion and intracellular kill of *Aspergillus fumigatus* conidia by human macrophages. *Cell. Microbiol.*, 4: 793-804 (2002).

Langfelder K, Jahn B, Gehringer H, Schmidt A, Wanner G, Brakhage AA. Identification of polyketide synthase gene (*pksP*) of *Aspergillus fumigatus* involved in conidial pigment biosynthesis and virulence. *Med. Microbiol. Immunol.*, 187: 79-89 (1998).

Langfelder K, Streibel M, Jahn BJ, Haase G, Brakhage AA. Melanin biosynthesis and virulence of human pathogenic fungi. *Fungal Genet. Biol.*, 38: 143-158 (2003).

Latgé J-P. *Aspergillus fumigatus* and Aspergillosis. *Clin. Microbiol. Rev.*, 12: 310-350 (1999).

Lengeler KB, Davidson RC, D'Souza C, Harashima T, Shen W-C, Wang P, Pan X, Waugh M, Heitmann J. Signal transduction cascades regulating fungal development and virulence. *Microbiol. Mol. Biol. Rev.*, 64: 746-785 (2000).

Liebmann B, Gattung S, Jahn B, Brakhage AA. (2003) cAMP signaling in *Aspergillus fumigatus* is involved in the regulation of the virulence determinant-encoding gene *pksP* and the defense against killing by macrophages. *Molec. Genet. Genomics*, 269: 420-435 (2003).

Mayorga ME *et al.* A novel anti-invasin antifungal compound with activity against fluconazole-resistant Candida albicans. *Abstracts of the Interscience Conference on Antimicrobial Agents and Chemotherapy*, 43:247 (2003).

Mayorga ME, Timberlake WE. The developmentally regulated *Aspergillus nidulans* wA gene encodes a polypeptide homologous to polyketide and fatty acid synthases. *Mol. Gen. Genet.*, 235: 205-212 (1992).

Money NP. Mechanism linking cellular pigmentation and pathogenicity in rice blast disease. *Fungal Genet. Biol.*, 22: 151-152 (1997).

Mosquera J *et al* In vitro interaction of terbinafine with itraconazole, fluconazole, amphotericin B, and 5-flucytosine against *Aspergillus* spp. *J. Antimicrob. Chemother.*, 50:189-194 (2002).

Nicholson TP *et al.* Design and utility of oligonucleotide gene probes for fungal polyketide synthases. *Chem. Biol.*, 8: 157-178 (2001).

Odds FC, Brown AJP, Gow NAR. Antifungal agents: mechanisms of action. *Trends Microbiol.*, 11: 272-279 (2003).

Pahl HL, Krauss B, Schulze-Osthoff K, Decker T, Traenckner EB, Vogt M, Myers C, Parks T, Warring P, Muhlbacher A, Czernilofsky AP, Baeuerle PA. The immunosuppressive fungal metabolite gliotoxin specifically inhibits transcription factor NF-kappaB. *J. Exp. Med.*, 183: 1829-1840 (1996).

Rhodes JC, Oliver BG, Askew DS, Amlung TW. Identification of genes of *Aspergillus fumigatus* up-regulated during growth on endothelial cells. *Med. Mycol.*, 39: 253-260 (2001).

Steinbach WJ *et al.* Advances against Aspergillosis. *Clin. Infect. Dis.*, 37(supp.3):55-56 (2003).

Summers EF *et al.* MM-86553, a novel anti-invasin antifungal compound, acts synergistically with Amphotericin B against *Candida albicans*. *Abstracts of the Interscience Conference on Antimicrobial Agents and Chemotherapy*, 43:248-9 (2003).

Sutton P, Newcombe NR, Waring P, Müllbacher A. *In vivo* immunosuppressive activity of gliotoxin, a metabolite produced by human pathogenic fungi. *Infect. Immun.*, 62: 1192-1198 (1994).

Sutton P, Waring P, Mullbacher A. Exacerbation of invasive aspergillosis by the immunosuppressive fungal metabolite, gliotoxin. Immunol. *Cell. Biol.*, 74: 318-322 (1996).

Thines E, Weber RW, Talbot NJ. MAP kinase and protein kinase A-dependent mobilization of triacylglycerol and glycogen during appressorium turgor generation by *Magnaporthe grisea*. *Plant Cell*, 12: 1703-1718 (2000).

Tsai, H-F, Washburn RG, Chang YC, Kwon-Chung KJ. *Aspergillus fumigatus arp1* modulates conidial pigmentation and complement deposition. *Mol. Microbiol.*, 26: 175-183 (1997).

Tsai H-F, Yun CC, Washburn RG, Wheeler MH, Kwon-Chung KJ. The developmentally regulated *alb1* gene of *Aspergillus fumigatus*: Its role in modulation of conidial morphology and virulence. *J. Bacteriol.*, 180: 3031-3038 (1998).

Tsai H-F, Wheeler MH, Chang YC, Kwon-Chung, KJ. A developmentally regulated gene cluster involved in pigment biosynthesis in *Aspergillus fumigatus arp1* modulates conidial pigmentation and complement deposition *J. Bacteriol.*, 181: 6469-6477 (1999).

Tsai H-F, Fujii I, Watanabe A, Wheeler MH, Chang YC, Yasuoka Y, Ebizuka Y, Kwon-Chung KJ. Pentaketide-melanin biosynthesis in *Aspergillus fumigatus* requires chain-length shortening of a heptaketide precursor. *J. Biol. Chem.*, 276: 29292-29298 (2001).

Wheeler MH, Bell AA. Melanins and their importance in pathogenic fungi. In: McGinnis, MR (Ed.) *Current Topics in Medical Mycology,* Springer Verlag, New York, N.Y. pp. 338-387 (1988).

Yamada A, Kataoka T, Nagai K. The fungal metabolite gliotoxin: immunosuppressive activity on CTL-mediated cytotoxicity. *Immunol. Lett.*, 71: 27-32 (2000).

Yoshida LS, Abe S, Tsunawaki S. Fungal gliotoxin targets the onset of superoxide-generating NADPH oxidase of human neutrophils. *Biochem. Biophys. Res. Commun.*, 268: 716-723 (2000).

Index

Adrenoleukodystrophy, 355
AfCS. *See* Alliance for Cellular Signaling
Alliance for Cellular Signaling, 360
Alzheimer's, psychiatric lipidome, 355
Amphotericin B, 368
Analyte determination, 12
Antifungal drug development, 367–369
Anti-sense oligonucleotides, use in cancer therapy, 326
Apoptosis, stable isotope labeled metabolic network, sensitivity, 329–331
Apoptosis resistant cells, oxidative pentose cycle metabolism, 330–331
Apoptosis sensitive cells
  de novo fatty acid synthesis, lack of, 329–330
  lack of *de novo* fatty acid synthesis, 329–330
  non-oxidate pentose cycle metabolism, 329–330
Approaches to scientific inquiry, reductionist, systems theory, contrasted, 1–2
Austrian Genomics of Lipid-associated Disorders consortium, 360

Biochemical markers, 51–52
Biodiversity assessment, metabolic profiling and, 36–37
Biology Work Bench, database, 200
Biomarker discovery, differential metabolic profiling, 137–157
  clinical applications, 150–152
  data mining, 146–148
  data processing, quantification, 143–146
  disease biomarkers, 150–151
  drug discovery, development, 151–152
  mass spectrometry
    biomarker discovery using, 141–149
    instrumentation, 143
  metabolic profiling approaches, 139–141
  sample collection, handling, 142
  sample preparation, 142–143
  statistics, 146–148
  validation, 148–149
Biomarkers
  disease, overview, 46–48
  neurodegenerative diseases, 48–52
Biosynthetic clusters, transcriptional profiles, fungal virulence, 377–378
Breeding, plants, metabolic profiling and, 34–35

Cancer
  genetic, proteomic targets, therapies, 325–327
  transformed metabolic network, 327–328
Capillary electrophoresis, 83–101
  application, 94–99
  capillary dimensions, 89
  detection, 88–89
  electrolyte system, 90–91
  field strength, 89–90
  injection, 87
  instrumentation, 86–89
  metabolome profiling, 98–99
  micellar electrokinetic chromatography, 84–86
  on-line sample proconcentration, 91–93
    dynamic pH injection, 92–93
    dynamic pH junction-sweeping, 93
    field-enhanced sample stacking, 91–92
    sweeping, 92
    transient-isotachophoresis, 93
  optimizing parameters, 89–91
  principles, 84–86
  role of, 99–100
  target metabolites, 94–98
  temperature, 90
  zone electrophoresis, 84
Capillary zone electrophoresis, 83
Caspofungin, 368
Cellular metabolism, modelling, 195–197
  model-based methods, 196
  types of models, 195–196
Central metabolic pathways, regulation, yeast as reference model, 14
Central nervous system disorders, 45–61
  biomarkers, 45–61
  clinical biomarkers, 51

disease, overview, 46–48
disease signatures, identifying, 53–54
genetic markers, 48–49
information flow, metabolomics in, 52–53
motor neuron diseases, 56
neurodegenerative diseases, 48–52
neuroimaging biomarkers, 49–50
personalized approach to therapies, 58–59
psychiatric disorders, 57
therapeutic targets, identifying, 54
use of metabolomics, 52–57
Classifications, metabolic profile-based, methodological issues, 173–194
experimental design, 178–184
analytical concerns, 181
biological variability, 181–182
controls, fuzzy vs tight, 180–181
genders/cohorts, 182–184
exploratory analyses, 186–188
high abundance state markers, metabolomics with, 175–176
informatics approaches, 184–191
initial cuts, 184–185
model optimization, 188–191
algorithm, choice of, 189
model simplification, 189
pattern recognition, 189–190
reliability of models, increasing, 190–191
robust metabolic profiles, 185–186
serotype, defining, 176–177
Coenzyme Q10, 129
COMET. *See* Imperial College Consortium on Metabonomic Toxicology
Comparative metabolome profiling, with two-dimensional thin layer chromatography, 63–81. *See also* Two-dimensional thin layer chromatography
Conjugated toxins, use in cancer therapy, 326
Contrast between reductionist approach to scientific inquiry, systems theory, 1–2
Control coefficients, metabolic control analysis, direct experimental determination, 234–235
Co-regulation, metabolic, models of, 279–284
CZE. *See* Capillary zone electrophoresis

Data sets, kinetic models using, 215–242
kinetic modelling, biological systems, 221–226
metabolic control analysis, 233–236
control coefficients, direct experimental determination, 234–235
definitions, 233–234
kinetic models, 235–236
model validation, 227–232
examples, 228–229
by nuclear magnetic resonance spectroscopy, 229–232
nuclear magnetic resonance, *in vivo* enzyme kinetics by, 225–226
silicon cell, linking modules, 237–238
in situ kinetic parameters, determination of, 224–226
structural modeling, biological systems, 216–221
elementary mode analysis, 220–221
in vivo kinetic parameters, determination of, 224–226
De novo fatty acid synthesis
apoptosis resistant cells, 330–331
apoptosis sensitive cells, 329–330
Detailed kinetic models using metabolomics data sets, 215–242
kinetic modelling, biological systems, 221–226
metabolic control analysis, 233–236
control coefficients, direct experimental determination, 234–235
definitions, 233–234
kinetic models, 235–236
model validation, 227–232
examples, 228–229
by nuclear magnetic resonance spectroscopy, 229–232
nuclear magnetic resonance, *in vivo* enzyme kinetics by, 225–226
silicon cell, linking modules, 237–238
in situ kinetic parameters, determination of, 224–226
structural modeling, biological systems, 216–221
elementary mode analysis, 220–221
in vivo kinetic parameters, determination of, 224–226
Differential metabolic profiling, biomarker discovery, 137–157
clinical applications, 150–152
data mining, 146–148
data processing, quantification, 143–146
disease biomarkers, 150–151
drug discovery, development, 151–152
mass spectrometry

biomarker discovery using, 141–149
instrumentation, 143
metabolic profiling approaches, 139–141
sample collection, handling, 142
sample preparation, 142–143
statistics, 146–148
validation, 148–149
Disease signatures, identifying, 53–54
Drug development
lipidomics in, 349–365
metabolomics in (*See under* specific condition or drug)
Dynamic pH injection, capillary electrophoresis, 92–93
Dynamic pH junction-sweeping, capillary electrophoresis, 93
Dyslexia, 355

EcoCyc, database, 200
Electrochemistry in metabolic profiling, 119–135
electrochemical measurement, 127–129
genomics, 129–130
liquid chromatography-electrochemical-array, 121–130
parallel electrochemical array-mass spectrometry, xenobiotic toxicity studies, 122–127
analytical conditions, 122
biological samples, 122–123
pattern recognition analysis, 125–127
proteomics, 129–130
serial electrochemical-mass spectrometry, 131–132
EMP database. *See* Enzymes and Metabolic Pathways database
Enzymes and Metabolic Pathways database, database, 200
Excreted metabolites, yeast, role of, 15
External metabolites, yeast, metabolic profiling, 13
External signals, yeast as reference model, metabolite sensors, 14
Extraction of internal metabolites, yeast as reference model, 11–12

Fast sampling, 11
Fatty acid synthesis, *de novo*
apoptosis resistant cells, 330–331
apoptosis sensitive cells, 329–330
Field-enhanced sample stacking, capillary electrophoresis, 91–92
Fluconazole, 368
Fluxome profiling in microbes, 307–322
analytical fluxome profiling, 309–310
challenges, 315–318
model-independent comparative profiling, 311–315
complex media, application to, 313–314
experimental proof-of-concept, 312–313
^{2}H-tracers, application to, 313–314
learning methods, unsupervised *versus* supervised, 314–315
Fungal metabolism modelling, 195–214
cellular metabolism, modelling, 195–197
model-based methods, 196
types of models, 195–196
fungal models, 204–211
functional properties, 205–207
network properties, 205
reaction deletion analysis, 209–211
topological properties, 207–209
genome-scale models, 197–204
current models, properties, 198–199
genome-scale models, applications of, 202
metabolic network reconstruction, 199–201
model development, 201
integrative analysis, 211–212
Fungal virulence, 367–381
antifungal drugs, 367–369
biosynthetic pathways, 369–374
demographics of infection, evolving, 367–369
gliotoxin, 374
pigments, 370
DHN-melanin biosynthesis pathway, 370–372
melanin in pathogenesis, 372–374
transcriptional profiles, 374–378
association analysis, 376–377
biosynthetic clusters, 377–378
gene expression data sets, 375–376

Gas chromatography-mass spectrometry, 103–106
nutritional research, 113
pharmaceutical research, 113
Genome-scale models, fungal metabolism, 197–204
analysis of metabolic networks, 201
current models, properties, 198–199
genome-scale models, applications of, 202

metabolic network reconstruction, 199–201
model development, 201
Gliotoxin, 368, 374

Hierarchial network model, 250
High performance liquid chromatography-mass spectrometry, 340–341
HPLC. *See* High performance liquid chromatography

Identification of disease signatures, 53–54
Immunoliposome-encapsulated drugs, use in cancer therapy, 326
Imperial College Consortium on Metabonomic Toxicology, 360
Information flow, metabolomics in, 52–53
In silico route to systems biology, 2, 4–5
In situ kinetic parameters, kinetic models using metabolomics data sets, determination of, 224–226
Integrative biochemical profiling, metabolites, and proteins, 269–276
Integrative functional genomics, 196–197
capillary electrophoresis, 83–101
in central nervous system disorders, 45–61
classifications, metabolic profile-based, methodological issues, experimental design, 173–194
developments in, overview, 1–7
differential profiling, for biomarker discovery, 137–157
electrochemistry, application of, 119–135
fluxome profiling in microbes, 307–322
fungal metabolism, 195–214
with gas chromatography-mass spectrometry, 103–118
kinetic models, using metabolomics data sets, 215–242
with liquid chromatography-mass spectrometry, 103–118
metabolite, transcript profiling, parallel, 291–306
networks, metabolic, 243–264
systems perspective, 265–289
pathogenic fungal physiology, 367–381
pharmaceuticals, 337–348
lipidomic approaches, 349–365
metabolic pathway flux, 323–335
in plants, 31–44
using nuclear magnetic resonance, 159–171
using two-dimensional thin layer chromatography, 63–81
yeast as reference model, integrative functional genomics using, 9–29
In vivo enzyme kinetics, kinetic models using metabolomics data sets, nuclear magnetic resonance, 225–226
In vivo kinetic parameters, kinetic models using metabolomics data sets, determination of, 224–226

Kinetic models using metabolomics data sets, 215–242
kinetic modelling, biological systems, 221–226
metabolic control analysis, 233–236
control coefficients, direct experimental determination, 234–235
definitions, 233–234
kinetic models, 235–236
model validation, 227–232
examples, 228–229
by nuclear magnetic resonance spectroscopy, 229–232
nuclear magnetic resonance, *in vivo* enzyme kinetics by, 225–226
silicon cell, linking modules, 237–238
in situ kinetic parameters, determination of, 224–226
structural modeling, biological systems, 216–221
elementary mode analysis, 220–221
in vivo kinetic parameters, determination of, 224–226
Kyoto Encyclopedia of Genes and Genomes, database, 200

Lipid class modules, combined, 354–355
Lipid consortiums, lipidomics, 359
Lipid Map Consortium, National Institutes of Health, 360
Lipidome, dividing into modules, 353–354
Lipidomics
classifications, 350–351
defined, 349–350
pharmaceuticals, 349–365
vs. conventional approaches, 351–352
Lipid transport, 357
Liquid chromatography-electrochemical-array, 121–130
Liquid chromatography-mass spectrometry, 106–108

contemporary applications of, 110–113
functional genomics, 112–113
high throughput metabolite profiling, 108–110
medical research, 113
metabolism research, engineering, 110–112
nutritional research, 113
pharmaceutical research, 113

Macular degeneration, 355
Mass spectrometry, 340–341
Max Plank Insitute, Germany, 359
MEKC. *See* Micellar electrokinetic chromatography
Melanin, in pathogenesis, 372–374
Metabolic networks, 243–264, 277–285
characterization of, 244–247
hierarchial network model, 250
metabolic network utilization, 251–256
models, 247–256
random network models, 247–248
regulation of metabolic reactions, 256–260
scale-free network model, 248–250
structure, 244–247
topological modularity, 251
utilization of metabolic reactions, 256–260
Metabolome analyses, 196–197
capillary electrophoresis, 83–101
in central nervous system disorders, 45–61
classifications, metabolic profile-based, methodological issues, experimental design, 173–194
developments in, overview, 1–7
differential profiling, for biomarker discovery, 137–157
electrochemistry, application of, 119–135
fluxome profiling in microbes, 307–322
fungal metabolism, 195–214
with gas chromatography-mass spectrometry, 103–118
kinetic models, using metabolomics data sets, 215–242
with liquid chromatography-mass spectrometry, 103–118
metabolite, transcript profiling, parallel, 291–306
networks, metabolic, 243–264
systems perspective, 265–289
pathogenic fungal physiology, 367–381
pharmaceuticals, 337–348
lipidomic approaches, 349–365
metabolic pathway flux, 323–335
in plants, 31–44
using nuclear magnetic resonance, 159–171
using two-dimensional thin layer chromatography, 63–81
yeast as reference model, integrative functional genomics using, 9–29
MetaCyc, database, 200
Methodological issues, metabolic profile-based classifications, 173–194
experimental design, 178–184
analytical concerns, 181
biological variability, 181–182
controls, fuzzy vs tight, 180–181
genders, cohorts, 182–184
exploratory analyses, 186–188
high abundance state markers, metabolomics with, 175–176
informatics approaches, 184–191
initial cuts, 184–185
model optimization, 188–191
algorithm, choice of, 189
model simplification, 189
pattern recognition, 189–190
reliability of models, increasing, 190–191
robust metabolic profiles, 185–186
serotype, defining, 176–177
Micellar electrokinetic chromatography, 83, 84–86
Microbes, fluxome profiling in, 307–322
analytical fluxome profiling, 309–310
challenges, 315–318
model-independent comparative profiling, 311–315
complex media, application to, 313–314
experimental proof-of-concept, 312–313
^{2}H-tracers, application to, 313–314
learning methods, unsupervised *versus* supervised, 314–315
Microbial lipidome, 355
Milk lipidome, 355
Model-independent comparative profiling, microbe fluxome profiling, 311–315
complex media, application to, 313–314
experimental proof-of-concept, 312–313
^{2}H-tracers, application to, 313–314
learning methods, unsupervised *versus* supervised, 314–315
Monoclonal antibodies, use in cancer therapy, 326
Motor neuron diseases, 56

MPW database. *See* Netabolic Pathways Database
MS. *See* Mass spectrometry
Multiple sclerosis, 355

National Institutes of Health, Lipid Map Consortium, 360
Nestle Research Center, Lausanne, 357, 359
Netabolic Pathways Database, database, 200
Neurodegenerative disease, biomarkers, 45–61
  biochemical markers, 51–52
  clinical biomarkers, 51
  disease, overview, 46–48
  disease signatures, identifying, 53–54
  genetic markers, 48–49
  information flow, metabolomics in, 52–53
  motor neuron diseases, 56
  neuroimaging biomarkers, 49–50
  personalized approach to therapies, 58–59
  psychiatric disorders, 57
  therapeutic targets, identifying, 54
  use of metabolomics, 52–57
Neurodegenerative diseases, biomarkers, 48–52
Neuroimaging biomarkers, 49–50
NMR. *See* Nuclear magnetic resonance
Non-oxidate pentose cycle metabolism, apoptosis sensitive cells, lack of *de novo* fatty acid synthesis, 329–330
Nuclear magnetic resonance
  spectrometry, 167
  spectroscopy, 339–340
    kinetic models using metabolomics data sets, 229–232
    liquid samples, 339–340
    solid samples, 340
  toxicology research, 159–171
    advantages of, 164
    nuclear magnetic resonance, mass spectrometry, integration of, 167
    nuclear magnetic resonance data, analysis of, 161–163
    serum, nuclear magnetic resonance-based metabonomics of, 167–168
    tissue extracts, metabonomics investigations of, 168
    urine, examples of metabonomics research on, 165–166
    whole tissue, metabonomics investigations of, 168

On-line sample proconcentration, capillary electrophoresis, 91–93
  dynamic pH injection, 92–93
  dynamic pH junction-sweeping, 93
  field-enhanced sample stacking, 91–92
  sweeping, 92
  transient-isotachophoresis, 93
Oxidative pentose cycle metabolism, apoptosis resistant cells, *de novo* fatty acid synthesis, 330–331

Panomics route to systems biology, 2–4
Parallel electrochemical array-mass spectrometry, in xenobiotic toxicity studies, 122–127
  analytical conditions, 122
  biological samples, 122–123
  pattern recognition analysis, 125–127
Parallel metabolite, transcript profiling, 291–306
  combined data sets, bioinformatics on, insights, 299–301
  comparison, technology platforms available, 297–299
  correlative approach in biology, 294–296
  technology platforms, 292–294
Parkinson's disease, psychiatric lipidome, 355
Personalized approach to central nervous system therapies, 58–59
Pharmaceuticals, metabolomics in. *See under* specific condition or drug
Plants, functional diversity assessment, 31–44
  biodiversity assessment, 36–37
  breeding, 34–35
  non-targeted biochemical analyses, novel strategies, 32–33
  physiology, 34–35
  production chain, quality assessment in, 37–39
  quality traits, 31–44
  systems level understanding, role of metabolomics, 39–40
Plastoquinone, vitamin K1, 129
Production chain, plants, quality assessment in, 37–39
Proteins, integrative biochemical profiling, 269–276
Proteomics, 129–130
Psychiatric disorders, 57
Psychiatric lipidome, 355
  function of, 355

Quenching of metabolites, yeast as reference model, 11

Radioisotopes, in leukemia, 326
Random network models, 247–248
Redox active metabolite, 129
Reductionist approach to scientific inquiry, systems theory inquiry, contrasted, 1–2
Retinitis pigmentosa, 355

Scale-free network model, 248–250
Schizophrenia, psychiatric lipidome, 355
Scientific inquiry, reductionist approach to, systems theory inquiry, contrasted, 1–2
Serial electrochemical-mass spectrometry, 131–132
Serotype, defining, 176–177
Serum, nuclear magnetic resonance-based metabonomics of, 167–168
Signal transduction pathways, yeast as reference model, internal metabolites, 14–15
Silicon cell, kinetic models using metabolomics data sets, 237–238
Skin lipidome, 355
Small molecule inhibitors, use in cancer therapy, 326
Small molecule receptor antagonists, cancer therapy, 326
Stroke, 355
Sweeping, capillary electrophoresis, 92
Systems biology approach to metabolome
  capturing metabolome-wide changes, strategies, 3–4
  future developments, 5–6
  overview, 1–6
  panomics route to systems biology, 2–4
  role of metabolomics, 1–7
  in silico route to systems biology, 2, 4–5
Systems perspective, metabolic networks from, 265–289
  co-regulation, metabolic, models of, 279–284
  differential metabolic networks, 284–285
  integrative biochemical profiling, metabolites, 269–276
  metabolic networks, 277–285
  proteins, integrative biochemical profiling, 269–276
Systems theory, reductionist approach to scientific inquiry, contrasted, 1–2

Thin layer chromatography, two-dimensional, 63–81
  advantages of, 77–78
  bacterial applications, 72–77
  bacterial taxonomy, metabolome comparisons, 75
  culture conditions, 66
  culture extraction, 67–68
  differential comparisons, controls, stressed bacteria, 70–71
  labelling metabolites on chromatography plates, 70–71
  limitations of, 77–78
  metabolite labeling conditions, 67
  methodology, 66–71
  mutational changes, 74–75
  spot quantitation, 70–71
  stress effects, 72–74
Tissue extracts, metabonomics investigations of, 168
TLC. *See* Thin layer chromatography
Traditional view concerning function of matabolites, overview of, 2
Transcript, metabolite profiling, parallel, 291–306
  combined data sets, bioinformatics on, insights, 299–301
  comparison, technology platforms available, 297–299
  correlative approach in biology, 294–296
  technology platforms, 292–294
Transcriptional profiles, fungal virulence, 374–378
  association analysis, 376–377
  biosynthetic clusters, 377–378
  gene expression data sets, 375–376
Transient-isotachophoresis, capillary electrophoresis, 93
Two-dimensional thin layer chromatography, 63–81
  advantages of, 77–78
  bacterial applications, 72–77
  bacterial taxonomy, metabolome comparisons, 75
  culture conditions, 66
  culture extraction, 67–68
  differential comparisons, controls, stressed bacteria, 70–71
  labelling metabolites on chromatography plates, 70–71
  limitations of, 77–78
  metabolite labeling conditions, 67
  methodology, 66–71

mutational changes, 74–75
spot quantitation, 70–71
stress effects, 72–74

Urine, examples of metabonomics research on, 165–166
drug toxicity, in mice, 166
ethanol toxicity, in rats, 165–166

Virulence, fungal, 367–381
antifungal drugs, 367–369
biosynthetic pathways, 369–374
demographics of infection, evolving, 367–369
gliotoxin, 374
pigments, 370
DHN-melanin biosynthesis pathway, 370–372
melanin in pathogenesis, 372–374
transcriptional profiles, 374–378
association analysis, 376–377
biosynthetic clusters, 377–378
gene expression data sets, 375–376
Vitamin K1, 129
Voriconazole, 368

What Is There, database, 200
Whole tissue, metabonomics investigations of, 168

Xenobiotic toxicity studies, parallel electrochemical array-mass spectrometry in, 122–127
analytical conditions, 122
biological samples, 122–123
pattern recognition analysis, 125–127

Yeast as reference model, 9–29
excreted metabolites, role of, 15
functional genomics, metabolomic studies in, 13–18
metabolic profiling, 10–13
analysis methods, 10–12
concentration step, 12
extraction, internal metabolites, 11–12
fast sampling, 11
internal metabolites, 13
preparation of sample, 12
quenching of metabolites, 11
regulation, 13–16
central metabolic pathways, 14
external signals, metabolite sensors, 14
signal transduction pathways, internal metabolites, 14–15