

Computational Biology

Marc L. Pusey
Ramazan Savaş Aygün

Data Analytics for Protein Crystallization



 Springer

Computational Biology

Volume 25

Editors-in-Chief

Andreas Dress
CAS-MPG Partner Institute for Computational Biology, Shanghai, China

Michal Linial
Hebrew University of Jerusalem, Jerusalem, Israel

Olga Troyanskaya
Princeton University, Princeton, NJ, USA

Martin Vingron
Max Planck Institute for Molecular Genetics, Berlin, Germany

Editorial Board

Robert Giegerich, University of Bielefeld, Bielefeld, Germany
Janet Kelso, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
Gene Myers, Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
Pavel Pevzner, University of California, San Diego, CA, USA

Advisory Board

Gordon Crippen, University of Michigan, Ann Arbor, MI, USA
Joseph Felsenstein, University of Washington, Seattle, WA, USA
Dan Gusfield, University of California, Davis, CA, USA
Sorin Istrail, Brown University, Providence, RI, USA
Thomas Lengauer, Max Planck Institute for Computer Science, Saarbrücken, Germany
Marcella McClure, Montana State University, Bozeman, MO, USA
Martin Nowak, Harvard University, Cambridge, MA, USA
David Sankoff, University of Ottawa, Ottawa, ON, Canada
Ron Shamir, Tel Aviv University, Tel Aviv, Israel
Mike Steel, University of Canterbury, Christchurch, New Zealand
Gary Stormo, Washington University in St. Louis, St. Louis, MO, USA
Simon Tavaré, University of Cambridge, Cambridge, USA
Tandy Warnow, The University of Illinois at Urbana-Champaign, Urbana, IL, USA
Lonnie Welch, Ohio University, Athens, OH, USA

The *Computational Biology* series publishes the very latest, high-quality research devoted to specific issues in computer-assisted analysis of biological data. The main emphasis is on current scientific developments and innovative techniques in computational biology (bioinformatics), bringing to light methods from mathematics, statistics and computer science that directly address biological problems currently under investigation.

The series offers publications that present the state-of-the-art regarding the problems in question; show computational biology/bioinformatics methods at work; and finally discuss anticipated demands regarding developments in future methodology. Titles can range from focused monographs, to undergraduate and graduate textbooks, and professional text/reference works.

More information about this series at <http://www.springer.com/series/5769>

Marc L. Pusey · Ramazan Savaş Aygün

Data Analytics for Protein Crystallization

 Springer

Marc L. Pusey
iXpressGenes, Inc.
Huntsville, AL
USA

Ramazan Savaş Aygün
University of Alabama in Huntsville
Huntsville, AL
USA

ISSN 1568-2684

Computational Biology

ISBN 978-3-319-58936-7

ISBN 978-3-319-58937-4 (eBook)

<https://doi.org/10.1007/978-3-319-58937-4>

Library of Congress Control Number: 2017957692

© Springer International Publishing AG 2017

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature

The registered company is Springer International Publishing AG

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my parents who sponsored my education throughout their lives and for their continuous support and motivation, to my siblings for their support,

To my teachers, educators and mentors starting from the elementary school to dissertation work, and

To my lovely wife, Emel, and our beautiful children Dilay, Enes, and Akif.

Ramazan Savaş Aygün

Preface

Protein crystallization usually requires many experiments that check combinations of various factors such as pH, ionic strength, etc. for a successful crystalline outcome. Nevertheless, as crystalline outcome especially for proteins to difficult crystallize such as membrane proteins in the presence of lipids and detergents is rare, many trials have been set up. These protein crystallization trials are usually analyzed by an expert using a microscope. Going over thousands of unsuccessful trials for a few successful (but important) outcomes has been tedious. In recent years, automated robotic high-throughput systems are proposed to conduct many experiments and fast detection of crystalline conditions. Initially, these high-throughput systems were costly and accessible only by major research laboratories. The significant cost of these systems made these research systems available only big research laboratories. Recent advancements in computational aspects and analysis of protein crystallization and decreasing cost of hardware architectures make automated systems available to also small research laboratories. Moreover, new protein crystallization techniques such as trace fluorescence labeling do not only reduce the time for preparation and analysis of crystallization experiments but also help to develop fast and accurate computational methods for protein crystallization analysis. This book covers how to build low-cost but fairly accurate protein crystallization analysis system thus enabling small research groups to build their own robotic high-throughput systems and crystallization analysis systems.

This book covers various aspects of computational aspects of protein crystallization. Previously, the computational aspects were usually covered as supplementary information in major crystallization books or sometimes they were ignored. This book unites important aspects of data analytics for protein crystallization into a single book. The methods and programs were developed as a part of collaborative research by iXpressGenes, Inc. and the University of Alabama in Huntsville funded through NIH-STTR grants. These projects funded two Ph.D. students, six M.S. students, and two part-time students along with other ongoing contributors. We have developed a number of systems for analyzing protein crystallization process while comparing our work with the state-of-art techniques. This

book also covers relevant research that will help readers understand different dimensions of protein crystallization analysis.

This book is relevant to researchers who would like to know about computational aspects and data analytics components of protein crystallization. While the book is relevant to the community of structural biology, it also serves computer scientists who would like to get into the protein crystallization field.

This data analytics book on protein crystallization analysis covers the complete cycle of data analysis for protein crystallization. It starts from background information on protein crystallization, setting up screens by analyzing prior crystallization trials, building robotic setups, classifying crystallization trial images by effective feature extraction, analyzing crystal growth in time series images, segmenting crystal regions in images, providing focal stacking methods for crystallization images captures at varying fields of depth, and visualization of trials. The book is organized as follows:

“Chapter 1: Introduction to Protein Crystallization” gives information about how protein crystallization experiments are conducted in a wet lab. Besides traditional experiments, we also cover trace fluorescence labeling that helps data analytics.

“Chapter 2: Scoring and Phases of Crystallization” covers scoring and categorization of crystallization image trials. Researchers came up with their own way of categorization in the literature. This chapter presents a variety of ways for categorization.

“Chapter 3: Computational Methods for Protein Crystallization Screening” presents computational methods for determining cocktails to be tested based on the results of prior experiments and their scoring. While commercial screens enable setting up plates with many successful cocktails, the analysis of unsuccessful trials has been left to the experts. This chapter provides approaches for setting up plates for successful crystalline outcomes.

“Chapter 4: Robotic Image Acquisition” presents the hardware and software architectures for a basic high-throughput system.

“Chapter 5: Classification of Crystallization Trial Images” presents overview of features used in protein crystallization image classification. As feature extraction has been the bottleneck of high-throughput systems, this chapter categorizes features and analyzes their running-time for real-time systems.

“Chapter 6: Crystal Growth Analysis” presents spatiotemporal analysis of protein crystal growth. This chapter analyzes the formation of new crystals as well as the growth of crystals in size.

“Chapter 7: Focal Stacking for Crystallization Microscopy” presents how to generate in focus crystallization images from a set of images that are captured at varying depths of field of a microscope. Since crystals usually float in a 3D well, some crystals may be out of focus and focal stacking may be necessary for proper analysis.

“Chapter 8: Crystal Image Region Segmentation” presents how to extract regions of crystals as thresholding or binarization has been one of the challenging issues in image segmentation.

“Chapter 9: Visualization of Crystallization Trial Experiments” introduces how plates can be visualized before/after scoring, temporal visualization of wells under different lighting conditions, and enabling/updating scoring by experts.

“Chapter 10: Other Structure Determination Methods” provides alternate methods to obtain a 3D structure (neutron diffraction, cryogenic electron microscopy, nuclear magnetic resonance, and X-ray free electron laser diffraction) and methods suitable for more general structural information (chemical cross-linking, fluorescence resonance energy transfer, and circular dichroism).

“Chapter 11: Future of Computational Protein Crystallization” provides overview of methods in progress and future trends for protein crystallization.

Huntsville, AL, USA
August 2017

Marc L. Pusey
Ramazan Savaş Aygün

Acknowledgements

This book would not have been possible without funding obtained from National Institutes of Health (GM-090453) and (GM116283) grants.

There are a number of sincere students, researchers, and faculty members who contributed to different parts of research mentioned in this book. The degrees of graduate students who were funded through this research are mentioned in parenthesis: Samyam Acharya (M.S.), Bidhan Bhattarai (M.S.), Imren Dinc (Ph.D.), Semih Dinc, Midusha Shrestha (M.S.), Madhav Sigdel (Ph.D.), Madhu Sigdel (M.S.), Mahesh Kumar Juttu (M.S.), Suraj Subedi (M.S.), and Truong Xuan Tran (Ph.D. in progress). In addition, several students contributed to this project during NSF-REU program at the University of Alabama in Huntsville. These NSF-REU students are Trevor Chan, James Rothenfleu, Nancy Gordillo-Herrejon, Jennifer Li, Edmond Malone, and Hilarie Pilkinton. Rujesh Shrestha and Hari Pradhan contributed as part-time students. Semih Dinc (Ph.D.) voluntarily contributed to this project as he was pursuing his Ph.D. degree.

Ms. Crissy L. Tarver, Ph.D. student in structural biology, was instrumental in preparing many of the samples and carrying out crystallizations using the methods and protocols outlined.

Contents

1	Introduction to Protein Crystallization	1
1.1	Introduction	1
1.1.1	The Protein Molecule	1
1.2	The Phase Diagram	2
1.3	The Second Virial Coefficient	5
1.3.1	Second Virial Coefficient Thought Experiments	6
1.3.2	But the Protein Still Does Not Crystallize!	7
1.4	Practical Considerations When Crystallizing Proteins	7
1.4.1	Other Factors Affecting Protein Crystallization	7
1.4.2	The Importance of the Protein	8
1.5	The Protein Crystallization Screening Process	8
1.5.1	Screening Methods	10
1.5.2	Experimental Design in Introducing the Protein to Precipitant	10
1.5.3	Screening Data Analysis	11
1.6	Introducing the Protein to the Precipitant—How to Do It?	13
1.6.1	Dialysis	13
1.6.2	Liquid–Liquid Diffusion	13
1.6.3	Vapor Diffusion	14
1.6.4	Batch Method	15
1.7	Following the Crystallization Experiment	15
1.7.1	Methods for Viewing the Crystallization Screening Results	16
1.8	Results Interpretation	16
1.9	Crystallization of Complexes	17
1.10	Crystallization of Integral Membrane Proteins	17
1.11	Summary	18
	References	18

2	Scoring and Phases of Crystallization	21
2.1	Introduction	21
2.2	Why Score Crystallization Drop Results?	22
2.3	Our Scoring Scale	22
2.4	Our Scoring Procedure	22
2.4.1	What You See Is Not Always Simply Classified	24
2.4.2	Hierarchical Categories	28
2.5	Even if You Are Not Going to Process Your Scored Data...	30
2.6	Summary	31
	References	31
3	Computational Methods for Protein Crystallization Screening	33
3.1	Introduction	33
3.2	Overview of Experimental Design Methods for Screening	34
3.3	Using Neural Networks for Experimental Design	35
3.4	Genetic Algorithm for Protein Crystallization Screening	37
3.5	Associative Experimental Design	39
3.6	Optimization of Cocktails	41
3.6.1	Elimination of Prohibited Combinations	42
3.6.2	Prioritization of Reagents	43
3.6.3	Ranking of Prioritized Conditions	43
3.6.4	Optimizing Concentration Values	45
3.7	Experiments and Evaluation	46
3.7.1	Proteins for Preliminary Experiments	46
3.7.2	Results for Preliminary Data	47
3.7.3	Expanded Screen Analysis	49
3.7.4	Evaluation of Ranked Results	51
3.8	Summary	52
	References	55
4	Robotic Image Acquisition	57
4.1	Introduction	57
4.2	Components of a Robotic Setup	61
4.2.1	Well Plates	61
4.2.2	Fluorescence Microscopy	61
4.3	Image Acquisition	64
4.4	Image Processing and Segmentation	64
4.4.1	Image Preprocessing	65
4.4.2	Segmentation	67
4.5	Feature Extraction	70
4.5.1	Intensity Features	70
4.5.2	Region Features	71
4.6	Accuracy and Timing Analysis	75
4.6.1	Multilayer Perceptron Neural Network (MLP)	76

4.6.2	Max-Class Ensemble Method	76
4.6.3	Computation Time	79
4.7	Summary	79
	References	80
5	Classification of Crystallization Trial Images	83
5.1	Introduction	83
5.1.1	Challenges of Protein Crystallization Classification	84
5.1.2	Factors for Classification	86
5.1.3	Feature Analysis for Building Real-Time Classifiers	88
5.2	Data Preprocessing	92
5.2.1	Feature Normalization	92
5.2.2	Dimensionality Reduction and Feature Selection	92
5.2.3	Image Processing	93
5.3	Classifiers	94
5.4	Feature Sets	96
5.4.1	Intensity Features	96
5.4.2	Histogram Features	96
5.4.3	Texture Features	98
5.4.4	Region Features	99
5.4.5	Graph Features	101
5.4.6	Shape-Adaptive Features	101
5.5	Analysis of Feature Sets	102
5.5.1	Data	103
5.5.2	Evaluating Features for Hierarchical Classification	105
5.5.3	First-Level (3-Class) Classification	105
5.5.4	Second-Level Classification	109
5.6	Timing Analysis for Classification	112
5.7	Deep Learning for Protein Crystallization Images	115
5.8	Discussion	116
5.9	Summary	119
	References	120
6	Crystal Growth Analysis	125
6.1	Introduction	125
6.2	Is it a Protein—Rule of Thumb	127
6.2.1	Protein—Get it While it is Fresh	128
6.3	Temporal Analysis of Time Series Images	128
6.3.1	Stages of Temporal Analysis	129
6.3.2	Sample Dataset and Experimental Setup	131
6.4	Identifying Trials for Spatiotemporal Analysis	132
6.4.1	Image Thresholding	132
6.4.2	Canny Edge Detection	133
6.4.3	Merging Results of Thresholding and Canny Edge Detection	134

6.4.4	Evaluation	135
6.5	Spatiotemporal Analysis of Protein Crystal Growth	135
6.5.1	Identifying Crystallographically Important Regions	136
6.5.2	Image Registration and Alignment	138
6.5.3	Spatiotemporal Features	138
6.6	Determining Crystal Growth	141
6.7	Detection of New Crystals	142
6.8	Detection of Crystal Size Increase	144
6.9	Discussion	145
6.9.1	Trace Fluorescent Labeling	145
6.9.2	Spatiotemporal Analysis	146
6.10	Summary	147
	References	148
7	Focal Stacking for Crystallization Microscopy	151
7.1	Introduction	151
7.2	Typical Viewing Area ~ 2 mm in Diameter	152
7.2.1	Objective Characteristics	153
7.2.2	Depth of Field	153
7.2.3	Drop Depth and Your Crystal Probably Isn't Where You Are Looking	154
7.3	Take Multiple Images to See Through the Drop	154
7.4	Auto-Focusing	155
7.4.1	Active Auto-Focusing	155
7.4.2	Passive Auto-Focusing	155
7.5	Focal Stacking	156
7.5.1	Pixel-Based Focal Stacking (PBFS)	158
7.5.2	Neighborhood-Based Focal Stacking (NBFS)	158
7.5.3	Transformation-Based Focal Stacking	158
7.6	Focal Stacking for Trace Fluorescently Labeling Microscopy	159
7.6.1	Modification of Harris Corner Response Measure (HCRM)	159
7.6.2	Calculating Representative HCRM Value	161
7.6.3	Generating Focused Image	162
7.7	Handling High-Resolution Images	164
7.8	Handling Varying Illumination	165
7.9	Evaluation of Focal Stacking Methods	168
7.9.1	Low-Resolution Image	169
7.9.2	High-Resolution Image	172
7.9.3	Varying Illumination Images	173
7.9.4	Comparison of Different Methods	173
7.10	Summary	175
	References	176

8	Crystal Image Region Segmentation	177
8.1	Introduction	177
8.2	Image Binarization Methods and Limitations	178
8.3	Supervised Thresholding	180
8.3.1	Building the Training Set	181
8.3.2	Correctness Measurement	181
8.3.3	Feature Extraction	182
8.4	Framework of Super-Thresholding	185
8.5	Priori Approach	186
8.6	Posteriori Approach	187
8.7	Evaluation of Super-Thresholding	188
8.7.1	Results	189
8.7.2	Discussion	193
8.8	Summary	194
	References	195
9	Visualization	199
9.1	Introduction	199
9.2	Plate Visualization	200
9.3	Well View	204
9.4	Scoring Crystallization Trials	205
9.5	Multiple Crystallization Trial Analysis	206
9.5.1	Time Course Analysis	206
9.5.2	Support for Sequential View	206
9.5.3	Multiple Light Source Support	206
9.6	Chemical Space Mapping	207
9.7	Summary	208
	References	209
10	Other Structure Determination Methods	211
10.1	Introduction	211
10.2	Neutron Diffraction (ND)	212
10.3	Nuclear Magnetic Resonance (NMR)	213
10.4	Cryogenic Electron Microscopy (Cryo-EM)	214
10.5	X-Ray Free Electron Laser (XFEL)	215
10.6	Other Approaches	217
10.6.1	Chemical Cross linking	217
10.6.2	Fluorescence Resonance Energy Transfer	218
10.6.3	Circular Dichroism (CD)	219
10.7	Summary	220
	References	220

11 Future of Computational Protein Crystallization	223
11.1 Introduction	223
11.2 Challenges and Future Directions	224
11.3 Summary	226
Index	227

Acronyms

ACC	Accuracy
AED	Associative experimental design
ANN	Artificial neural network
blob	Binary large object
BYS	Bayesian
CD	Circular dichroism
CNN	Convolutional neural network
CR	Carboxyrhodamine
CR-SE	Carboxyrhodamine succinimidyl ester
Cryo-EM	Cryogenic electron microscopy
CWT	Complex wavelet transform
DCT	Discrete cosine transform
DFT	Discrete Fourier transform
DT	Decision tree
EDF	Extended depth of field
EDF-CWT	Extended depth of field-Complex wavelet transform
EDF-RW	Extended depth of field-Real-valued wavelet transform
F1	F-Score
FHI	Full Harris image
FN	False negative
FocusALL-HR	FocusALL for high-resolution images
FocusALL-VI	FocusALL for varying illumination images
FP	False positive
FRET	Fluorescence resonance energy transfer
FS	Feature set
GLCM	Gray-level co-occurrence matrix in Chap. 4
GLCM	Green-level co-occurrence matrix in Chap. 5
GS	Grid screening
HCRM	Harris corner response measure
HRHT	Hampton research high throughput

IF	Incomplete factorial
IFE	Incomplete factorial experiments
IMP	Integral membrane protein
JACC	Jaccard similarity
LCP	Lipidic cubic phase
LED	Light-emitting diode
MBR	Minimum bounding rectangle
MCC	Matthew's correlation coefficient
MDA	Mean decrease in accuracy
MDA-RF	Random forest feature selection with mean decrease in accuracy
MLP	Multilayer perceptron neural network
mRMR	Minimal-redundancy-maximal-relevance
MW	Molecular weight
NBFS	Neighborhood-based focal stacking
ND	Neutron diffraction
NMR	Nuclear magnetic resonance
PBFS	Pixel-based focal stacking
PCA	Principle components analysis
Pcp	Pyroglutamate amino peptidase
PHI	Partial Harris image
RF	Random forest
ROI	Region of interest
RPE	Retinal pigment epithelial
RrP42	Archaeal exosome protein
RW	Real-valued wavelet transform
SA-DCT	Shape-adaptive discrete cosine transform
SaIPP	Staphylococcus aureus IPPase
SDS	sodium dodecylsulfate
SMS	Sparse matrix sampling
SVM	Support vector machine
TFL	Trace fluorescent labeling
TN	True negative
TP	True positive
TR	Texas Red, Molecular Probes/Invitrogen cat. # T-10244
Tt106	Nucleoside kinase
Tt189	Nucleoside diphosphate kinase
Tt82	HAD-superfamily hydrolase
ULWD	Ultra long working distance
WPF	Windows Presentation Framework
XFEL	X-ray free electron laser

Chapter 1

Introduction to Protein Crystallization

Abstract This chapter reviews the basics of the protein crystallization process. As amply proven by the protein structure initiative, protein crystallization can be carried out without any basic knowledge about the specific protein or how it behaves in solution. However, when the goal is not just processing as many proteins as can be produced, but is directed toward a better understanding of a specific biological moiety, a better understanding of what is being done, what one is observing, and how they all relate to the crystal nucleation and growth process is an invaluable aid in translating the observed screening results to a successful outcome. Informed observation is a key component to increased success. Similarly, there are a plethora of approaches that can be taken to screening for crystals, and knowing the strengths and weaknesses of each is key to matching them to the immediate goals to be achieved.

1.1 Introduction

Proteins are crystallized for several reasons. That of most importance to this work is for use in determining the proteins structure by diffraction methods, such as X-ray or Neutron crystallography. While a discussion of these methods is beyond the scope of this treatise, the important point is that the ability to use them is dependent upon the quality of the crystal. Diffraction data has been obtained from crystals of $< 1 \mu m$ in size [5], and thus with increasingly powerful X-ray sources, the size of the crystals is becoming less important. However, while size is not important, the quality of the crystal packing is very much so. Another reason for crystallizing proteins is as a purification step [18], an application that will likely grow with advances in biotechnology.

1.1.1 *The Protein Molecule*

Every protein is different, but that does not stop us from being able to describe their general properties, which both add to the difficulty in their crystallization and

lend tremendous utility in being able to obtain them in a crystalline state. To start with, they are very large compared to what the small molecule world works with, having monomeric molecular weights that can exceed 100,000+ kDa. When we start adding complexes to the consideration, either through self-association or with other macromolecules, these MWs can exceed 1,000,000 kDa.

The protein molecule is a linear polymer of amino acids, of which there are 20 that are coded for genetically and common to all life on Earth. It is the specific sequence of these amino acids and the particular geometry with which the polymeric chains fold that gives a protein its specific properties. A protein molecule's shape is not rigid, but flexes, in which flexibility is usually associated with its biological function. Conformational flexibility can be a major impediment to obtaining crystals, particularly diffracting to higher resolution, for a protein under study.

Amino acids are not the only components commonly found in protein structure. Many proteins undergo post-translational modification, the most common being glycosylation. These covalently attached carbohydrate groups are typically not well ordered and can be a major source of difficulty when trying to obtain crystals. It is possible to remove them, either by judicious choice of an expression system, mutation of the glycosylation sites, or chemical and/or enzymatic removal. However, the presence of the glycosylation may be key to the protein's biological function, with that information being lost in its absence.

Among the many other possible post-translational modifications are phosphorylation, lipidation, S-nitrosylation, acetylation, and methylation. All play important roles in protein function, and their presence, or absence, may affect one's ability to obtain crystals. A comprehensive review of post-translational modifications is beyond the scope of this work. However, we note that reductive methylation, where charged basic amino acid side chains are chemically modified to give hydrophobic derivatives, is employed as a tool for obtaining crystals from difficult proteins [39].

The size of the protein molecule, and its irregular shape, results in crystals that have considerable solvent channels. Protein crystals may be from ~25% to >60% solvent by volume, present as channels through the crystal. As a result, the crystals are typically very fragile, and great care needs to be taken when getting them from the growth conditions to diffraction analysis. On the plus side, the presence of these solvent channels means that the crystallographer can diffuse materials into the crystal, an important consideration when carrying out ligand binding or drug development studies.

1.2 The Phase Diagram

Any discussion about protein crystallization must start with the phase diagram. Understanding this, and what it tells about the crystallization process, is key to understanding what one is experimentally seeing and how the results should be interpreted. Further, even in the case of crystallization screening failures, the most common outcome, it gives us a tool for interpreting the results and some guidance for

subsequent experiments. A typical phase diagram is shown in Fig. 1.1. The X-axis, the crystallization variable, can be any factor that affects the protein in solution, such as temperature, precipitant concentration, pH, etc. The Y-axis is the protein or, more generally, the crystallizing solute, concentration. The phase diagram is divided into three zones or regions: soluble protein, the metastable zone, and the labile zone. Only one of the dividing lines shown between these regions is fixed and thermodynamically defined, the solubility. This represents the equilibrium between the solute in solution and that in the insoluble phase at the defined X-axis conditions and with all other factors held constant. The crystallographic state has the lowest solubility, which explains why we often observe crystals growing out of an amorphous (nonstructured) precipitate. In kinetic terms, the solubility concentration is where the attachment and dissociation rates of the solute to the insoluble material (crystal) surface are equal.

The metastable zone falls between the solubility line and the nucleation line. While it is often thought that this is a no nucleation zone, strictly speaking this is not true. Nucleation rates may be low, essentially zero over the short term, but if one defines the nucleation line as where nucleation begins, then one finds that the width of the metastable zone progressively narrows with increasing time. It then becomes intuitively obvious that the nucleation rate increases the further out one goes from the solubility line, and thus the position of the nucleation line is dependent upon how long one waits. Within this region, the nucleation rate is extremely low, and a crystal in this solution tends to grow without formation of new nuclei. Proteins have stability concerns over the longer term (from hours to weeks or months, depending upon the protein), which make waiting for the nucleation line to move sufficiently close to result in nucleation over a prolonged period is often not practicable.

Several notional crystallization pathways are shown in Fig. 1.1. Pathway A, the ideal case, starts with the system in the solubility zone, and progressing through the metastable to just within the labile zone as the drop comes to equilibration. Once in the vicinity of the metastable zone boundary, it comes to equilibrium with respect to the crystallization variable (X-axis) and a (in this case) single crystal is nucleated. The nucleated crystal removes solute (protein) as it grows and follows the line going down to the solubility, with the single crystal growing. The timescale is not indicated in this pathway. From start to crystal nucleation may be on the order of a few days. However, crystal growth rates are a function of the solute concentration, and thus the rate of solute removal, crystal growth, becomes progressively slower as one approaches the solubility line. Pathway B starts with the experiment being set up in the metastable zone and passing into the labile zone before the precipitant equilibrium concentration is reached. A single crystal is shown being nucleated once the labile zone is entered, but as the precipitant concentration continues to increase additional crystals are also nucleated. Once precipitant equilibrium is attained, the crystals grow and the pathway follows the vertical pathway to B*. However, as it is still in the labile zone, additional nucleation events occur, yielding still more crystals that continue to grow until the system passes into the metastable zone.

Pathway C starts in the soluble zone, terminating either just inside the metastable zone or still in the soluble protein zone. In both cases, the results will show as a clear solution. Pathway D starts well into the metastable zone and rapidly passes into the

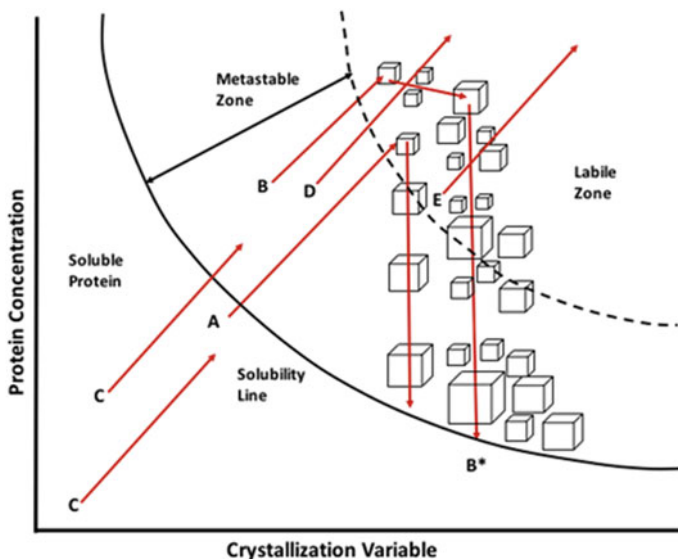


Fig. 1.1 Phase diagram

labile zone. Crystals may or may not be formed here, as it is possible that desolubilization kinetics manifesting as random self-associations can overtake the orderly assembly of a faceted crystal surface, with the results showing as a precipitated solution, possibly with some crystalline material buried within. Pathway E starts in the labile zone and typically shows up as amorphously precipitated protein when the drop is first set up. Other often occurring results for pathways D, E, and often also A and B, are the growth of non-faceted crystals, often manifesting as “urchins”, dendritic crystals, and spheroids. The presence of these outcomes infers that the protein can potentially form an ordered faceted crystal if the conditions are sufficiently adjusted; there is hope that the protein in its present form can be crystallized.

Several important points can be derived from Fig. 1.1. First, in the case of pathway $B \rightarrow B^*$, the crystals are nucleating at different supersaturation levels, and this may have an effect on the resultant diffraction qualities. Second, results that give a clear solution (line C) or an amorphous precipitate (lines D and E) are not indications of not being a crystallization condition. An increase in the protein concentration or crystallization variable may bring line C to crystallization conditions, while a decrease in these parameters may bring lines D and E to a successful conclusion. Third, reality and practical experience inform us that just being in the metastable zone does not guarantee that a crystal will be nucleated. There are other parameters at play, concerning the solution composition and how that affects the ability of the protein to form an ordered array of contacts—the crystal lattice. Crystallization screening experiments are carried out because we do not know *a priori* where we are in the solubility diagram when first mixing a test protein with any given precipitant

solution, and careful observation and interpretation of the observed outcomes can often bring the investigator to a successful outcome. For a more in-depth review of the phase diagram, the reader is referred to [30].

Most protein crystallization experiments are designed such that the solution hopefully transits a path shown by lines A or B in Fig. 1.1, where the solution just enters the labile zone, and then proceeds on to growing a crystal. While this is the ideal, the paths more often followed are shown by lines C, D, and E. Even the presumably worst-case scenario, where one starts in the labile zone (line E), may not be a disaster. It is entirely possible, and frequently observed, that one has an amorphous precipitate immediately upon setting up the experiment, only to find after some time that crystal nucleation has occurred and the amorphous protein has been replaced by crystals. One also frequently observes crystals growing in a cleared zone in an amorphous precipitate, indications that the crystal is “feeding” off of the precipitate for growth.

Assuming that the protein can be crystallized, a simple “fix” can be tried to obtain crystals from the pathway C. If sufficient protein is available, we have found it useful to increase the protein concentration 3 to 5 times that used in the initial screening experiment. Crystallization trials are then set up using just those conditions that were scored as a clear solution (see Chap. 2). This effectively shifts the pathway C to where their endpoints are more likely to be in the labile zone, possibly resulting in crystals. By the same token, the protein concentration can also be reduced in cases of pathways D and E, again shifting them to potentially more favorable endpoints.

1.3 The Second Virial Coefficient

No discussion of protein crystallization screening can be had without including the insights provided by light scattering studies from the Wilson laboratory [14]. Whereas light scattering studies on nucleating protein solutions had been carried out using dynamic light scattering (photon correlation), Wilson used static light scattering (the scattering intensity). The second virial coefficient B_{22} is a measure of two body, protein–protein, interactions. The data was plotted according to the equation:

$$\frac{Kc}{R(90)} = \frac{1}{M} + 2B(22)c, \quad (1.1)$$

where a plot of Kc/R_{90} versus C yields the second virial coefficient B_{22} as the slope of the line. The B_{22} values at crystallization conditions were determined for 10 proteins, and all were found to fall between -1×10^{-4} and -8×10^{-4} mol ml g^{-2} , known as the crystallization slot. Values above this gave clear solutions, while those below this range gave precipitated protein. The interpretation given of the results is that solution conditions giving B_{22} values in the crystallization slot are “moderately poor”, such that orderly interactions enabling formation of a crystalline lattice can occur. Note, however, that being within the slot does not mean that lattice formation will occur, although it was strongly postulated that solution conditions

giving B_{22} values outside (below) the crystallization slot had a very low probability for successful crystal formation.

1.3.1 Second Virial Coefficient Thought Experiments

The conclusions drawn from the light scattering experiments can be rethought as shown in Fig. 1.2. Here, the Y-axis is the B_{22} , or second virial coefficient, value and the X-axis is the crystallization variable being manipulated to desolubilize the protein. The crystallization slot region of the B_{22} , which is necessary but not sufficient for a successful outcome, is indicated. Starting with a solution at point A, with zero precipitant, there is a range of possible outcomes as the concentration of the solubility variable is varied. Note that these outcomes are dependent upon the protein, the other (fixed) components present, and the nature of the solubility variable.

Three possible trajectories are shown, varying X to arrive at points B, C, or D. All three points are outside and below the crystallization zone, and the end product in all three cases is assumed to be amorphously precipitated (i.e., non-crystalline) protein. While the trajectories shown are drawn as straight lines, in reality they may be curved, as shown by the dashed line going from A to C. However, all trajectories pass through the crystallization zone, and this must be true for any such experiment. We can reason that in principle we should be able to control the crystallization variable such that we can stop in the crystallization zone and increase our chances of success. That this is not possible is because in some instances, such as the A→B pathway, the control required to keep the system in the crystallization slot long enough for nucleation to occur is difficult if not impossible to achieve. This control is graphically more feasible for the A→C and A→D pathways, both of which provide sufficient range for us to set our limit conditions to within the slot.

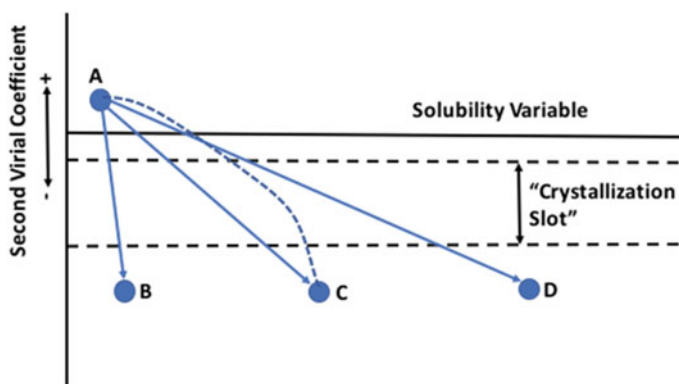


Fig. 1.2 Solubility

1.3.2 But the Protein Still Does Not Crystallize!

The reason things are not so simple is the crystal packing process itself. The second virial coefficients are definable as physical interactions between molecules [16, 40]. This says that the attractive interactions between two molecules need to fall within a certain range for structure formation. Short-range attractive interactions are group specific, which gives a basis for structure formation [41]. However, there may be more than one set of interactions giving interactions having the appropriate strength. In this case, more than one interaction may form, with a multiplicity of interactions resulting in an aggregate structure comprising limited ordering on the molecular but none at the macroscale. Alternatively, the protein may have too much conformational flexibility or have extensive floppy or unstructured regions that interfere with the formation of a lattice structure.

As mentioned above in Fig. 1.1, the labile zone boundary line is not fixed, but over time moves toward the solubility line. For this reason, experiments to define its position need to note the chosen endpoint time. Proteins are not rigid structures, but flex and shift in their conformation. This conformational flexibility occurs both normally and is often critical to their biological function. Further, they have a finite lifetime both in vivo and in vitro. As a result, they degrade over time, both in storage and crystallization experiments. One of the best ways of storing proteins is in a precipitated form, preferably as a crystalline precipitate, for which one first has to determine crystallization conditions.

1.4 Practical Considerations When Crystallizing Proteins

Proteins are typically stabilized by lower temperatures, and protein chemists almost reflexively store their solutions at 4° C. While protein solubility diagrams are not routinely determined, for those that have been determined, it has been found that solubility usually goes down with decreasing temperature. Temperature has been long recognized as an important variable in protein crystallization, and there are several vendors selling variable temperature incubators specifically for protein crystallization purposes. Despite the advantages of lower temperatures, it is perhaps a testament to human psychology that most crystallization experiments are carried out around ambient temperatures, not 4° C. If one is to spend hours setting up, then observing crystallization experiments, the preference is obviously to do so at more comfortable temperatures.

1.4.1 Other Factors Affecting Protein Crystallization

Factors other than just temperature and the precipitant concentration affect the protein crystallization process. Physical factors other than temperature include the vibra-

tional level of the experiment, the surfaces where the solution touches including the presence or not of added nucleants, the pressure, and the rate at which the system is brought to equilibrium. Considerations based on the protein molecule itself include the purity of the protein, the presence or absence of naturally occurring or experimenter-induced post-translational modifications, ligands, substrates, or inhibitors, the molecular shape such as its inherent symmetry and the presence or not of unstructured floppy regions, and the self-association state of the protein. Some of these are properties inherent to the protein itself, while others are directly affected by the experimenter, such as limited proteolysis and chemical or genetic modification. Finally, there are those factors that are most often manipulated by the experimenter to effect crystallization, including the solution composition such as precipitant(s), ionic strength, and pH; the presence or not of specific ions, metals, or cross-linking species; the protein concentration and purity; the presence or not of added solubility modifiers such as detergents and amphophiles. All of the above factors, and more, will affect the final outcome and the quality of the crystal obtained.

1.4.2 The Importance of the Protein

The protein crystallization process begins with the protein. How the protein is prepared is a dominant factor in the eventual success, or not, of the crystallization process. Planning for crystallization begins at the genetic level: which amino acids to modify, which domains to keep or leave off, etc. The source of the protein is important; Is it purified directly from the source or is it obtained by recombinant methods? If recombinant methods then what expression system is used? Does it provide post-translational modification? Are all necessary disulfide bonds formed? Once a supply of the raw protein is on hand, how is it purified? This is often a trial-and-error process when working with native proteins from their source, and often a rote procedure when purifying recombinant proteins having an affinity purification tag. At the end, how pure is the final product? Some protein can tolerate the presence of considerable impurities, while others are highly intolerant. Discussion of purification methods and approaches is beyond the scope of this treatise. However, a final consideration of the purification process is that variations from batch to batch of protein are often present, and these may dramatically affect the ability of the protein to crystallize as well as the quality of the crystals obtained.

1.5 The Protein Crystallization Screening Process

Given the myriad factors affecting crystallization, how does one typically set about determining crystallization conditions? The method by which they are found is called the screening process. The purpose of screening experiments is the determination of the main factors of importance to the system under investigation. Once identi-

fied, subsequent optimization tests are carried out to refine the effects of the main factors [24, 25]. To this end, the current macromolecular screening methodology often fails, as major factors are only taken to be revealed when crystals are obtained. A consequence is that screening is often expanded to many hundreds or thousands of conditions to maximize combinatorial chemical space coverage and the chances of a randomly acquired successful (crystalline) outcome. This drives a reduction in screening trial volumes to enable additional experiments and thus the need for robots to set up and then monitor them. The cost of the latter puts the “survey everything” approach out of the reach of many smaller sized structure groups. The reduction in screening solution volumes also reduces the chances of finding hit conditions that may be present in the screening experiments. From the classical definition, the crystal nucleation rate is a function of the number of nuclei appearing in a given volume per unit of time. As it is a stochastic process, experiments to determine nucleation rates are typically set up multiple replicates, as one may have some that nucleate almost immediately, some eventually, and some never over the course of the experiment. As a result, setting up screening experiments where the volume is minimized, and having only one experiment per condition, may lead to many missed crystallization hits.

There is a practical limit to the screening process. For many labs, this is taken to be the amount of protein and/or number of screening kits on hand combined with the dispensing precision of their screening plate dispensing system. However, Segelke [35] has calculated that for a protein having a 2% chance of successful crystallization about 300 experiments would provide a sufficient screening base. He determined that a better approach would be to focus more on protein purification, to have the best possible material for screening. The best screening method was postulated to be random sampling to maximize the chemical space surveyed. Another argument in favor of a more limited screening approach is the effort involved. The results from each screening experiment need to be tracked over time, and as a practical consideration it is considerably less effort to follow 3 or 4 plates, giving a 2% chance of success, than it is to track 7 or 8 plates which may only give a 1–1.5% increase in the success rate.

The screening process is composed of two components: carrying out the screening experiment itself and the subsequent analysis of the results obtained. For many, the only good endpoint of the analysis is if there is a crystal, and as a result the rich lode of experimental data that has been acquired is ignored. The bulk of this treatise is directed toward approaches that can be taken to the subsequent analysis, to recovering additional information that may lead to eventual success where one initially had none. The actual screening experiments, which provides that information for analysis, can and have been carried out in a variety of methods. Virtually any parameter control method that will modify the desolubilization of protein in solution has been tried. At their most fundamental, however, they comprise a means of introducing the protein solution to a potential crystallization solution, aka the screening solution or cocktail.

1.5.1 Screening Methods

Protein crystallization conditions were initially devised based on an understanding of and familiarity with the protein under study. Prior to structural biology, they were often developed as a purification step, under the philosophy that crystallization represented the ultimate in purity. As the use of X-ray diffraction proliferated, newer approaches have been developed to facilitate screening for obtaining crystals for structural studies.

Despite the plethora of methods available for the controlled desolubilization of proteins to obtain crystals, only a few methods now dominate the field. In large part, this is because they must accommodate several requirements: They must be easily scaled to a large number of experiments, involving a wide range of solution conditions; the experimental results should be easily accessible for imaging and analysis; it should be relatively easy to set up a large number of diverse solutions in parallel, using either robotic or manual methods; and when crystals are obtained, they should be readily accessible for subsequent diffraction analysis. Because of these practical considerations, methods based on standard SBS size plates, 127.75 mm × 85.48 mm—height variable, now dominate the field. The plates are made having an array of experimental formats, from 24 wells to 1536 wells (or more), for use with sitting drop, hanging drop, or dialysis experiments, having depressions for one or more sitting drops connected to each reservoir, and designed for counter diffusion experiments. The following discussion is limited to these more commonly employed methods.

1.5.2 Experimental Design in Introducing the Protein to Precipitant

Crystallization of screening approaches, the method by which combinatorial chemical space is surveyed for crystallization conditions, can be broken down into four methods: sparse matrix, random, grid (complete factorial), and incomplete factorial. Variations on these approaches, for example, random sampling screens that vary precipitating agents over a limited grid screen range, aka the footprint screen [37] are also extant. Sparse matrix screens [17], which are based upon known crystallization conditions chosen to cover a broad range of crystallization space, predominate in the commercial world. Random screens [35] avoid one of the pitfalls of sparse matrix screens by enabling the inclusion of neglected areas or components of crystallization space.

The approach that has been most successful is the sparse matrix screen, where combinations of chemicals are used. The most significant, which set the pattern for virtually all subsequent developments, was the devising of a screen based on analysis of published crystallization conditions that had been developed to that time for proteins. Developed by Jancarik and Kim [17], this set of 50, then expanded to include

another 48, screening conditions is perhaps the most used screen to date. Virtually all commercial screen manufacturers produce their own version of this screen. The typical implementation is to use the first 48 conditions of the first screen and the 48 conditions of the follow-on screen as a 96-condition block of solutions, which format is directly compatible with the standard 96-well plate format for crystallization screening.

Random screens may be thought of as a hybrid of incomplete factorial and sparse matrix screens [35]. They avoid the pitfalls of sparse matrix screens by the inclusion of neglected regions of chemical space. Like sparse matrix, they attempt to cover as broad a range of chemical space as possible. Like IF screens, they attempt to do this coverage in a balanced method, with no one or group of components being over represented. However, they suffer in that one must be equipped with the ability to readily compose the myriad solutions required on demand.

Grid screens systematically explore variations in the components of crystallization solutions. This may be carried out in one or two dimensions, one dimension being the concentration of a precipitant and the second of another precipitant or the solution pH. Although grid screens are commercially available, because of the limited chemical space covered, this is generally not a method of first choice for the crystallization of a new protein. Typically, lead conditions derived by other methods are refined to improved crystallization conditions using a grid screen.

The use of the incomplete factorial approach to screening for protein crystallization conditions was first put forth by Dr. C. W. Carter [8] for the crystallization of tryptophan tRNA synthetase. A 35-condition screen was used to define the important variables, and then a complete factorial screen was carried out using four factors to obtain crystals. The six variables and number of levels for each were precipitating agent (7), pH (4), temperature (3), divalent cation (3), anion (4), and monovalent cation (4). The method was shown to quickly be able to identify the factors important to the proteins crystallization. The incomplete factorial approach was used in subsequent crystallizations from this group [3, 4, 7], and elaborated on in a review [6]. The program used to design the screen, INFAC, was available from Dr. Carter upon request.

1.5.3 Screening Data Analysis

Analysis of screening results begins once the screens have been set up. It is advantageous to look at the experiments as soon as possible after set up, to get a time 0 view of the effects of mixing the protein with precipitant solution. Thereafter, one should periodically review the results obtained. The purpose is to track and follow emerging crystallizations, or developments in the precipitation in the well. A rule of thumb is that salt crystals appear quickly, while protein crystals appear over time, although this is not always the case. However, generally even quickly appearing protein crystals will often grow over a few days, while salt crystals do not.

A number of systems have been developed and/or marketed for the routine imaging of crystallization screening plates, using a diverse range of imaging methods. Some have included software for finding and identifying crystals in the experimental wells, which is not a trivial process when using white light imaging. The ability of the software to distinguish crystals from white light images then becomes paramount, and this is not a trivial process. The myriad shapes and arrangements that occur (see Chap. 2) confound the image analysis process, and this before we consider separating protein from salt crystals. However, the philosophical goal of this approach is that crystal = good, no crystal = not good, and ignore those wells where there were no crystals. In taking this approach, one discards a trove of data reflecting the protein's response to the imposed solution conditions, which data may eventually point to crystallization conditions. Thus, the analysis of the results, beyond crystal/no crystal, can be a powerful tool in proceeding to the goal of a diffracting crystal, and thus a protein structure.

Several approaches to analysis of the screening results have been put forth. The first of these was dependent upon use of the IF method, where use of a statistically balanced approach enables the user to extract more than simple yes/no information from the screening results. Several descriptions of use of the IF method have been published [1, 8, 9, 31, 33, 34, 38]. However, this method has not significantly (as determined by cited usage) caught on. One reason may be the absence of designed incomplete factorial screens, complete with appropriate results analysis software. In the absence of these tools, each would-be user is left to develop their own screen, then their own software approach for the statistical analysis of the results obtained. The 1536 condition screen implemented by the Hauptmann-Woodard Institute [21] contains an IF component, but analysis of the data is by visual methods using graphically presented results [26, 36].

Saridakis [32] first suggested the use of a genetic algorithm for the analysis of crystallization screening results obtained by more standard screening methods. This approach is a "stochastic multiparameter optimization technique" which has found utility in a range of applications, particularly when simultaneous optimization of a number of parameters is required. In this case, the optimization process is evolutionary, using recombination, mutation, and selective pressure, over a number of generations.

Methods of better visualizing the results, beyond simple viewing with transmission microscopy, are highly useful when carrying out crystallization screening. A number of methods have been described in the literature [11, 15, 18, 20, 22, 23, 27, 29], many of which are commercially available. We use the trace fluorescent labeling (TFL) method [11, 29], where a fluorescent probe emitting in the visible region of the spectrum is covalently attached to the protein. The procedure is designed so that $\sim 0.2\%$ of the protein molecules are modified, with all free dye removed after the binding reaction is carried out. An advantage of this method over the others is that more than one color of the probe can be used, making this very useful for the crystallization of macromolecular complexes. All fluorescent images shown in this treatise are acquired using TFL protein.

1.6 Introducing the Protein to the Precipitant—How to Do It?

Once the means for exploring crystallization parameter space has been determined, one is now left with the method to be taken for exploring that space. There are a number of physical approaches that have been taken to setting up protein crystallization experiments, and not surprisingly each comes with its own advantages and disadvantages. The results obtained from each may vary. The method chosen will determine the ease with which the experimenter can review the results. Some, particularly those that are based on standard plate geometries, are well suited to automated methods of imaging and image data storage.

1.6.1 *Dialysis*

Historically crystallization was typically carried out by dialysis or by batch methods as a final purification step. In the batch method, the protein is introduced to the precipitant conditions, possibly by slow addition of the factors or adjustment of pH, and left to sit or is then subjected to a temperature change to induce crystallization. In today's parlance batch, crystallization is carried out by mixing the protein with the crystallization cocktail at a specified ratio of solutions, typically under an oil layer to reduce evaporation of the small solution volume. Dialysis is where the protein solution is placed within a dialysis bag, the material of the bag being a semi-permeable membrane, which is then closed off at both ends and placed in a larger volume of the solution to be used for crystallization. The membrane used is chosen such that the precipitant chemicals can pass through but not the protein. The solution is typically stirred and the protein solution comes to equilibrium with the external dialysis solution, the end result hopefully being the crystallization of the protein. Micro-versions of the dialysis approach are extant, with dialysis buttons having volumes down to 5 μL available (www.hamptonresearch.com). In this case, the protein solution is placed within the button cavity, and then covered over by a dialysis membrane which is secured in place with an O-ring. The button is placed within a precipitant solution in a crystallization plate (typically 24-well size). Advantages are that the external solution can be changed over time, and one can readily track progress of the experiment using standard microscopy systems. A disadvantage is that this approach does not lend itself to high, or even moderate, screening rates.

1.6.2 *Liquid–Liquid Diffusion*

Dialysis is a version of liquid–liquid diffusion screening, whereby a precipitant solution is allowed to slowly diffuse into the protein solution to hopefully effect crys-

tallization. The second implementation of this approach is called capillary counter diffusion [12]. While there are a number of ways to physically set up a capillary counter diffusion crystallization experiment, the basis is to put the protein solution in a capillary, followed by a precipitant solution such that there is a starting interface between the two. The capillaries are typically set up such that the lighter, less dense, solution, usually the protein, is above the precipitant. The precipitant slowly diffuses into the protein, and when crystallization occurs one often has a distribution of increasingly larger crystals the further one goes from the starting interface. It is not unusual to obtain crystals that totally fill the diameter of the capillary, and diffraction data can be directly acquired without having to handle the crystal at all [13]. As with dialysis, this technique is not well suited for large-scale screening trials, but more for terminal crystal production for diffraction analysis.

1.6.3 Vapor Diffusion

Vapor diffusion is the most popular method used for protein crystallization screening. The principle is relatively straightforward. Protein solution is mixed at some volume ratio, usually 1 to 1, with the precipitant solution and then sealed in a chamber in the presence of the precipitant solution at full concentration, known as the reservoir. Water is generally the only volatile component in the system, and it moves from the protein:precipitant mixture, through the vapor phase, to the reservoir solution. This concentrates the protein and precipitant solution concentrations in the crystallization drop. Referring to Fig. 1.1, all paths describe a possible vapor diffusion scenario. There are two geometries extant shown in Fig. 1.3, hanging drop and sitting drop, which basically describes the orientation of the crystallization trial solution with respect to the precipitant solution. In hanging drop, the earliest implementation of this technique, the protein and precipitant are mixed on a (usually) glass surface, typically a microscope coverslip, which is then inverted over and sealed above a reservoir solution, such that the drop is enclosed in the same volume. With careful pipetting, several drops can be placed on a single coverslip to survey different volume ratios. Sitting drop vapor diffusion involves incorporation of a support surface within the well volume, which may have stations far from 1 to 5 drops. Again, the protein and precipitant solution are mixed in each station and the wells are typically sealed by a transparent film. There are advantages and disadvantages to each, most notably limiting hanging drop size and stability. By virtue of its more reproducible positioning format, the sitting drop approach better lends itself to robotic methods for both setting up and imaging.

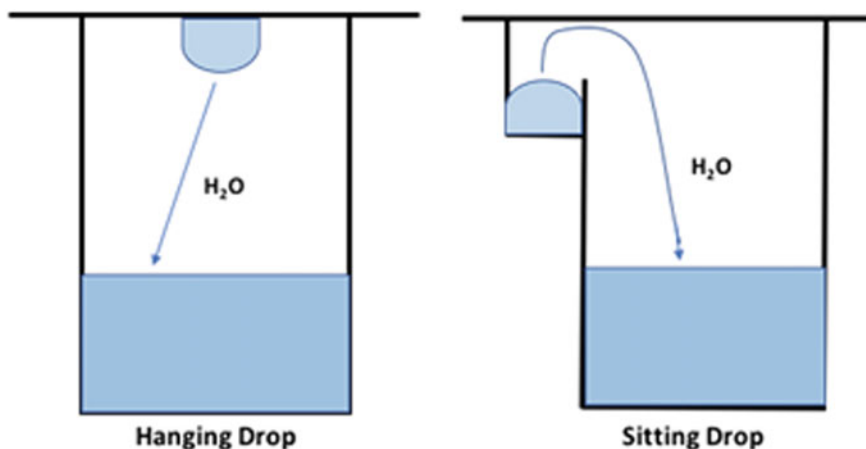


Fig. 1.3 Vapor diffusion approaches to macromolecule crystallization

1.6.4 Batch Method

The batch method is where one simply mixes the protein with precipitant, at some volume ratio of the solutions, and waits for a result. It should be pointed out that the starting protein + precipitant droplet in a vapor diffusion experiment is a batch experiment. Whereas in the vapor diffusion the protein droplet varies in size and composition, in a batch experiment the conditions are fixed at the outset. Batch experiments are typically carried out under a layer of oil to reduce evaporation, and because of this have an advantage in that one can set them up at a more leisurely pace, without worrying about evaporation during the setup process. The typical oil overlayer is paraffin oil, which is not water permeable. However, the composition of this overlayer can be modified by mixing in silicon oil, typically at a 1:1 ratio. The silicon oil is water permeable, and this mixture allows the slow loss of water from the drop, making this a vapor diffusion experiment with the atmosphere as the reservoir.

1.7 Following the Crystallization Experiment

Once the screen has been set up, the next task is to periodically review the experiments to determine what progress, if any, has been made toward the macromolecule crystallization goal. Finding crystals may seem to be a simple proposition, but this is often not the case. Crystals may be anywhere in the drop, and it pays to be very careful when examining each drop. Crystals may not be of the protein, but of other solution components—the dreaded salt crystals. The typical tool for this is a low-power microscope, preferably of the dissecting type, having a large support region

for the crystallization plates, a comfortable working distance, transmission illumination, and a zoom feature to facilitate examination of regions of interest at higher magnification. With practice, one can reduce the amount of time spent on manually reviewing a plate to 5–10 min. The astute reader will realize that this can be the downside to using as many screens as possible in their quest for a crystal. At some point, one has to look at the plates, preferably multiple times as the evolution of features can be a strong pointer to their being macromolecule, and not salt crystals which tend to appear quickly.

1.7.1 Methods for Viewing the Crystallization Screening Results

The logical response to having large numbers of plates or crystallization drops to review was to automate the process. Over the years, a number of crystal plate imaging systems have appeared on the market. Initially, they just used transmission microscopy, with or without image processing, but in recent years, detection methods have diversified to the use of UV fluorescence [10, 19], visible fluorescence [11, 22], two-photon fluorescence [23, 27], and second-order nonlinear optical imaging (SONICC) [20]. Prices and functionality vary, with some systems requiring manual movement of the plate, some manual loading with subsequent automatic imaging of the plate, and some having plate “hotels” that double as isothermal crystallization incubators from which plates are robotically retrieved, scanned, and then returned, on a programmed interval basis.

Having images periodically recorded of the plates does not alleviate the review process. The crystal grower must still look at every image, if possible with reference to previous images for that crystallization well. This is not an exercise that should only be carried out at the end of some arbitrary plate incubation period. While crystals may often develop over a time period of weeks or months, we have found it useful to identify those conditions that give crystals within the first week or so after setup. These conditions can then be used as the basis for optimization and/or production experiments to generate crystals for an upcoming data collection trip.

1.8 Results Interpretation

Once the screening plates have been set up, the next problem is to interpret the results obtained and find the macromolecule crystals if they are present. This topic is covered in greater depth in Chap. 2. While one or more crystals growing in the center of an otherwise clear crystallization drop are easily observed, the investigator is still left with the problem of distinguishing whether they are protein or salt crystals, a common occurrence given that components of the protein solution may react to form

insoluble salts with components of the crystallization solution. The development of fluorescence-based methods was a response in part to the problem of distinguishing protein versus salt crystals. Historically, the most direct method to determine if the crystal was protein or salt was the “crush test”, where one applies light pressure, or even simply touches the crystal. If it is a salt crystal, it will remain intact, whereas the considerably more sensitive protein crystals will fall apart, often as a brown precipitate. Subsequently, the use of the dye methylene blue was introduced, originally marketed by Hampton Research under the name “Izit”. A small amount of the dye is added to the crystal containing droplet. Macromolecule crystals have large water channels running through them, whereas salt crystals do not. If the crystal takes up the dye, becoming intensely blue as a result, then it is protein. A fluorescence variation on this has been described [15]. Also, crystals under white light illumination may be obscured by their location in the well, or by being buried in precipitate. In a fluorescence or SONICC-based system, the crystals stand out from the solution, making their presence readily detectable when reviewing the well images. All protein crystallizations carried out in our laboratory use trace fluorescent labeling (TFL) as an aid in interpreting the results [11, 29]. Most of the images in this treatise show images using TFL’s crystals, many with their corresponding white light images for comparison purposes.

1.9 Crystallization of Complexes

Crystallization of protein complexes presents another round of problems. While the crystals may obviously (according to the standard tests) be protein, one is now faced with the problem of whether or not all of the components of the complex are present. Visible fluorescence offers a facile means of making this determination. Each component can be TFL with a different colored fluorescent dye prior to assembling the complex in solution. The presence of each of these dyes in the crystal then verifies the presence of that component in the crystal. This can be carried out for as many colors (components) as one can clearly distinguish in the presence of the others.

1.10 Crystallization of Integral Membrane Proteins

A still more complex problem is the crystallization of integral membrane proteins (IMPs). This group of proteins represents a higher level of difficulty in their production, purification, and crystallization. IMPs, having extensive hydrophobic regions, are not soluble without added detergents. The detergents must be present to extract the proteins from the membranes and to keep them soluble throughout the purification and crystallization process. Supply of IMPs, whether from the source organism or by recombinant production, is often very limited relative to soluble proteins. Further, the protein stability and the ease with which it can be crystallized is a function

of the detergent(s) employed, and one must often switch from a purification to a crystallization detergent. There is an ever-expanding range of detergents available for use with IMPs, and in addition to the usual buffer and salt selection process, the crystallographer must also optimize this additional and most critical component as well. Result analyses are further confounded by the possibility that the detergent may be coming out of solution in an ordered phase. The methods used for crystallizing IMPs include the recent addition of the lipidic cubic phase (LCP) method [2, 28].

1.11 Summary

Protein crystallization is not a trivial process. The practitioner must balance a large number of variables, both in the solution compositions employed and final protein purity obtained, in their quest for a successful outcome. While it is possible to blindly carry out a rote series of steps to obtain crystals, it is to the experimenter's advantage to understand the physical chemistry of what they are trying to accomplish and to interpret their results in light of that understanding. The results obtained will be dependent upon the pathway taken, both in purification and crystallization methodology employed, and it is usually beneficial to use all of the tools at ones disposal in the quest for well-diffracting crystals. The final results, the crystal diffraction resolution, will be highly dependent upon how well those variables are managed.

References

1. Abergel, C., Moulard, M., Moreau, H., Loret, E., Cambillau, C., & Fontecilla-Camps, J. C. (1991). Systematic use of the incomplete factorial approach in the design of protein crystallization experiments. *Journal of Biological Chemistry*, 266(30), 20131–20138.
2. Ai, X., & Caffrey, M. (2000). Membrane protein crystallization in lipidic mesophases: Detergent effects. *Biophysical Journal*, 79(1), 394–405.
3. Bell, J. B., Jones, M. E., & Carter, C. W. (1991). Crystallization of yeast orotidine 5'-monophosphate decarboxylase complexed with 1-(5'-phospho- β -D-ribofuranosyl) barbituric acid. *Proteins: Structure, Function, and Bioinformatics*, 9(2), 143–151.
4. Betts, L., Frick, L., Wolfenden, R., & Carter, C. W. (1989). Incomplete factorial search for conditions leading to high quality crystals of *Escherichia coli* cytidine deaminase complexed to a transition state analog inhibitor. *Journal of Biological Chemistry*, 264(12), 6737–6740.
5. Boutet, S., Lomb, L., Williams, G. J., Barends, T. R. M., Aquila, A., Doak, R. B., Weierstall, U., DePonte, D. P., Steinbrener, J., Shoeman, R. L., Messerschmidt, M., Barty, A., White, T. A., Kassemeyer, S., Kirian, R. A., Seibert, M. M., Montanez, P. A., Kenney, C., Herbst, R., Hart, P., Pines, J., Haller, G., Gruner, S. M., Philipp, H. T., Tate, M. W., Hromalik, M., Koerner, L. J., Bakel, N. v., Morse, J., Ghonsalves, W., Arnlund, D., Bogan, M. J., Caleman, C., Fromme, R., Hampton, C. Y., Hunter, M. S., Johansson, L. C., Katona, G., Kupitz, C., Liang, M., Martin, A. V., Nass, K., Redecke, L., Stellato, F., Timneanu, N., Wang, D., Zatsepin, N. A., Schafer, D., DeFevers, J., Neutze, R., Fromme, P., Spence, J. C. H., Chapman, H. N., & Schlichting, I. (2012). High-Resolution protein structure determination by serial femtosecond crystallography. *Science* 337(6092), 362–364.

6. Carter, C. W. (1997). [5] Response surface methods for optimizing and improving reproducibility of crystal growth. *Methods in Enzymology*, 276, 74–99.
7. Carter, C. W., Baldwin, E. T., & Frick, L. (1988). Statistical design of experiments for protein crystal growth and the use of a precrystallization assay. *Journal of Crystal Growth*, 90(1–3), 60–73.
8. Carter, C. W., & Carter, C. W. (1979). Protein crystallization using incomplete factorial experiments. *Journal of Biological Chemistry*, 254(23), 12219–12223.
9. DeLucas, L. J., Bray, T. L., Nagy, L., McCombs, D., Chernov, N., Hamrick, D., et al. (2003). Efficient protein crystallization. *Journal of Structural Biology*, 142(1), 188–206.
10. Dierks, K., Meyer, A., Oberthü, D., Rapp, G., Einspahr, H., & Betzel, C. (2010). Efficient UV detection of protein crystals enabled by fluorescence excitation at wavelengths longer than 300 nm. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, 66(4), 478–484.
11. Forsythe, E., Achari, A., & Pusey, M. L. (2006). Trace fluorescent labeling for high-throughput crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 62(3), 339–346.
12. García-Ruiz, J. M. The Uses of crystal growth in gels and other diffusing-reacting systems. *Key Engineering Materials* 58 (1991), 87–106.
13. Gavira, J. A., Toh, D., Lopéz–Jaramillo, J., & García–Ruiz, J. M. (2002). Ab initio crystallographic structure determination of insulin from protein to electron density without crystal handling. *Acta Crystallographica Section D: Biological Crystallography*, 58(7), 1147–1154.
14. George, A., & Wilson, W. W. (1994). Predicting protein crystallization from a dilute solution property. *Acta Crystallographica Section D: Biological Crystallography*, 50(4), 361–365.
15. Groves, M. R., Müller, I. B., Kreplin, X., & Müller-Dieckmann, J. (2007). A method for the general identification of protein crystals in crystallization experiments using a noncovalent fluorescent dye. *Acta Crystallographica Section D: Biological Crystallography*, 63(4), 526–535.
16. Hill, T. L. (1959). Theory of solutions. II. osmotic pressure virial expansion and light scattering in two component solutions. *The Journal of Chemical Physics*, 30(1), 93–97.
17. Jancarik, J., & Kim, S.-H. (1991). Sparse matrix sampling: A screening method for crystallization of proteins. *Journal of Applied Crystallography*, 24(4), 409–411.
18. Judge, R. A., Johns, M. R., & White, E. T. (1995). Protein purification by bulk crystallization: The recovery of ovalbumin. *Biotechnology and Bioengineering*, 48(4), 316–323.
19. Judge, R. A., Swift, K., & González, C. (2005). An ultraviolet fluorescence-based method for identifying and distinguishing protein crystals. *Acta Crystallographica Section D: Biological Crystallography*, 61(1), 60–66.
20. Kissick, D. J., Wanapun, D., & Simpson, G. J. (2011). Second-order nonlinear optical imaging of chiral crystals. *Annual Review of Analytical Chemistry*, 4(1), 419–437.
21. Luft, J. R., Collins, R. J., Fehrman, N. A., Lauricella, A. M., Veatch, C. K., & DeTitta, G. T. (2003). A deliberate approach to screening for initial crystallization conditions of biological macromolecules. *Journal of Structural Biology*, 142(1), 170–179.
22. Lukk, T., Gillilan, R. E., Szebenyi, D. M. E., & Zipfel, W. R. (2016). A visible-light-excited fluorescence method for imaging protein crystals without added dyes. *Journal of Applied Crystallography*, 49(1), 234–240.
23. Madden, J. T., DeWalt, E. L., & Simpson, G. J. (2011). Two-photon excited UV fluorescence for protein crystal detection. *Acta Crystallographica Section D: Biological Crystallography*, 67(10), 839–846.
24. Mason, R. L., Gunst, R. F., & Hess, J. L. (2003). *Statistical Design and Analysis of Experiments: With Applications to Engineering and Science* (2nd ed.), Wiley series in probability and statistics New York: Wiley.
25. Myers, R., & Montgomery, D. (2009). *Response Surface Methodology: Product and Process Optimization Using Designed Experiments*. 1995 (4th ed.). New York: Wiley.
26. Nagel, R. M., Luft, J. R., & Snell, E. H. (2008). AutoSherlock: A program for effective crystallization data analysis. *Journal of Applied Crystallography*, 41(6), 1173–1176.

27. Padayatti, P., Palczewska, G., Sun, W., Palczewski, K., & Salom, D. (2012). Imaging of protein crystals with two-photon microscopy. *Biochemistry*, *51*(8), 1625–1637.
28. Pebay-Peyroula, E., Rummel, G., Rosenbusch, J. P., & Landau, E. M. (1997). X-ray structure of bacteriorhodopsin at 2.5 angstroms from microcrystals grown in lipidic cubic phases. *Science*, *277*(5332), 1676–1681.
29. Pusey, M., Barcena, J., Morris, M., Singhal, A., Yuan, Q., & Ng, J. (2015). Trace fluorescent labeling for protein crystallization. *Acta Crystallographica Section F: Structural Biology Communications*, *71*(7), 806–814.
30. Rupp, B. (2015). Origin and use of crystallization phase diagrams. *Acta Crystallographica Section F: Structural Biology Communications*, *71*(3), 247–260.
31. Saijo, S., Sato, T., Tanaka, N., Ichianagi, A., Sugano, Y., & Shoda, M. (2005). Precipitation diagram and optimization of crystallization conditions at low ionic strength for deglycosylated dye-decolorizing peroxidase from a basidiomycete. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, *61*(8), 729–732.
32. Saridakis, E. (2011). Novel genetic algorithm-inspired concept for macromolecular crystal optimization. *Crystal Growth and Design*, *11*(7), 2993–2998.
33. Sedzik, J. (1994). Design: A guide to protein crystallization experiments. *Archives of Biochemistry and Biophysics*, *308*(2), 342–348.
34. Sedzik, J. (1995). Regression analysis of factorially designed trials – a logical approach to protein crystallization. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* *1251*(2), 177–185.
35. Segelke, B. W. (2001). Efficiency analysis of sampling protocols used in protein crystallization screening. *Journal of Crystal Growth*, *232*(1), 553–562.
36. Snell, E. H., Nagel, R. M., Wojtaszyk, A., O'Neill, H., Wolfley, J. L., & Luft, J. R. (2008). The application and use of chemical space mapping to interpret crystallization screening results. *Acta Crystallographica Section D: Biological Crystallography*, *64*(12), 1240–1249.
37. Stura, E. A., Nemerow, G. R., & Wilson, I. A. (1992). Strategies in the crystallization of glycoproteins and protein complexes. *Journal of Crystal Growth*, *122*(1), 273–285.
38. Tran, T. T., Sorel, I., & Lewit-Bentley, A. (2004). Statistical experimental design of protein crystallization screening revisited. *Acta Crystallographica Section D: Biological Crystallography*, *60*(9), 1562–1568.
39. Walter, T. S., Meier, C., Assenberg, R., Au, K.-F., Ren, J., Verma, A., et al. (2006). Lysine methylation as a routine rescue strategy for protein crystallization. *Structure*, *14*(11), 1617–1622.
40. Wills, P. R., Comper, W. D., & Winzor, D. J. (1993). Thermodynamic nonideality in macromolecular solutions: Interpretation of virial coefficients. *Archives of Biochemistry and Biophysics*, *300*(1), 206–212.
41. Wills, P. R., Jacobsen, M. P., & Winzor, D. J. (2000). Analysis of sedimentation equilibrium distributions reflecting nonideal macromolecular associations. *Biophysical Journal*, *79*(4), 2178–2187.

Chapter 2

Scoring and Phases of Crystallization

Abstract The practice of scoring of protein crystallization screening results is more honored in the breach than in the observance. However, as we hope to show in the balance of this treatise, it can lead to a means for extracting more information than immediately apparent from a crystallization experiment. Scoring has advantages beyond simple good scientific note-keeping practice; the act of objectively examining one's results, with some thought added, can lead to a deeper appreciation of what led to those results, be it at the protein, screening solution, or mechanics of setting up the plate level. The first goal is to have a system which reflects an increase in the desirability of the results obtained with the numerical score. The scoring scale does not have to be complex or extensive; a 10-point scale is elaborated on herein. However, the scale should clearly distinguish between classes of desirable outcomes.

2.1 Introduction

The opening mantra of this chapter, and in fact for all successful protein crystallization experiments, is that there is no substitute for careful visual observation of crystallization plates. Even in the absence of a formal analysis methodology, such as those outlined in subsequent chapters, an alert and careful observer will note patterns emerging in the results, either from well to well or within the droplets of a given well if that approach is taken. This chapter is written to give examples of how we interpret crystallization results. Other interpretation schemes may be used, but the primary importance is that one develops a familiarity with the results that are, or could, be obtained.

Tracking protein crystallization results, particularly in smaller laboratory's, is often a matter of circling the found outcomes of interest with a Sharpie™. Notes may be taken, but since the outcome of interest is a nicely faceted crystal then why bother noting that this precipitation was gummy in appearance, that one was lightly granular, and the one next to it was heavy and brown, while in between were several clear wells? Thus, while over the years a number of scoring scales have been put forth, they are rarely used when all that was deemed necessary was to circle the hits on the plate with a Sharpie™.

2.2 Why Score Crystallization Drop Results?

There are several reasons why one should score their crystallization screening results, not the least of which is that it is good scientific note keeping. In the absence of any formal post screening analysis, knowing what happened as the protein was placed in solution with a number of different chemicals, over a range of concentrations and pH's, may still serve as a basis in guiding subsequent optimization strategies when a hit is finally obtained. Careful note taking, with respect to the solution compositions, can serve to rule out the inclusion of specific chemicals, or suggest changes in the protein concentrations used. However, the rationale that serves as the basis of this treatise is that the scores can be used in the analysis of the results obtained, for potentially extracting crystallization conditions where one previously had none, or for expanding on the known conditions and identifying those that can more reproducibly yield crystals. This latter point is of particular interest if one is going to extensively work with the protein, such as for binding studies, and reliable crystalline conditions are needed.

2.3 Our Scoring Scale

A practical scoring scale needs to reflect an improvement in outcomes with an increase in the score. Many scales begin with a score of 0 for a clear solution. Referring to Fig. 1.1, we see that clear solutions can occur on either side of the solubility line, and in fact are not the worst outcome that can occur. That distinction is reserved for a heavy precipitate, and even here there are two types that can occur; one where the protein is still "intact" and can be redissolved, and second where the protein is partially denatured and cannot be redissolved. Distinguishing between these two precipitant types is not always easy, although having a heavy brown precipitate is typically taken as an indicator of the second type. Regardless, it then is apparent that changes in solution conditions that take an outcome from a precipitate to a clear solution are not detrimental, but an improvement in the outcome. The scoring scale that we have found best is provided revised score column of Table 2.1. This scale is the same as given by [3]. Figures 2.1, 2.2, and 2.3 gives an illustration of these scores.

2.4 Our Scoring Procedure

We follow a defined procedure for scoring crystallization screening results. All plates use trace fluorescently labeled protein, which enables us to follow what the protein is doing in response to the crystallization screening solution being tested [8, 11]. We use Corning 3553 CrystalEX™ sitting drop crystallization plates having 3 drop

Table 2.1 Scores for protein crystallization images

General category	Hampton's score	Revised score as in [3]	
Non-crystals	1	2	Clear drop
	2	3	Phase separation
	3	0	Heavy precipitate
	3	1	Light precipitate
Likely leads	4	4	Birefringent precipitate or Microcrystals
	–	4	Bright spots (Not present in Hampton's category)
Crystals	5	5	Urchins, spheroids, dendrites - non-faceted crystals
	6	6	Needles
	7	7	Plates - 2D crystals
	8	8	3D crystals <math>< \mu\text{m}</math>
	9	9	3D crystals >math>> 200 \mu\text{m}</math>

positions/precipitant well. The three positions are set up at protein:precipitant ratios of 1:1, 2:1, and 4:1 (vol:vol) or, alternatively, 1:2, 1:1, and 2:1. Assuming the same endpoint precipitant concentrations at equilibrium the varying drop ratio's give an indication of the effects of protein concentration on the outcome, and we frequently have results that progress from precipitate or small crystals to large single crystals across the three drops.

Typically, the plates are fluorescently imaged on a regular basis, and scoring is carried out between the 6th and 8th week after setup. The first step is to manually go through the plates well by well, using a standard low power microscope typically used for crystal plate viewing, and note down a score for each well on a scoring sheet. The score written down at this point is for what is observed under white light. The second step is to review the scores written down with respect to the most recent fluorescent images. The scores at this point are adjusted as necessary based upon what the fluorescent image reveals. Thus, objects that scored as a crystal are downgraded to what the background conditions show if they do not fluoresce, which indicates that they are not protein but salt crystals. It is at this point that we identify outcomes having a score of 4, the bright spots. The score for any given crystallization drop is that of the highest scoring object within that drop. This often necessitates careful examination of the drop contents, zooming in on features of interest and focusing through the solution. One small faceted crystal within a drop containing precipitate, granular precipitate, or apparently non-faceted crystals, will result in the drop being scored as an 8. Similarly, if a cluster of crystals or rods that might otherwise be scored as a 5 has one or more that protrude out sufficiently that it could be cleaved off and

mounted for diffraction as a single crystal then the score is that of the piece that can be cleaved off.

2.4.1 What You See Is Not Always Simply Classified

The first question when scoring a crystallization plate is “what is this”? While a faceted crystal, a long rod, or a spikey urchin may be obvious, there are other outcomes that are not so clearly defined. A good resource for interpreting one’s results are the images on the Terese Bergfors website: <http://xray.bmc.uu.se/terese>. A guide to the scoring used in our work is shown in Figs. 2.1, 2.2, and 2.3.

As can be inferred from Fig. 2.1, there are no hard and fast rules for scoring what one observes. While most outcomes are easily scored, this is not always the case. For example, as shown in Fig. 2.1, Panels A, B, and C, what distinguishes a heavy vs. a light precipitate? This particularly when there may be occasions where the

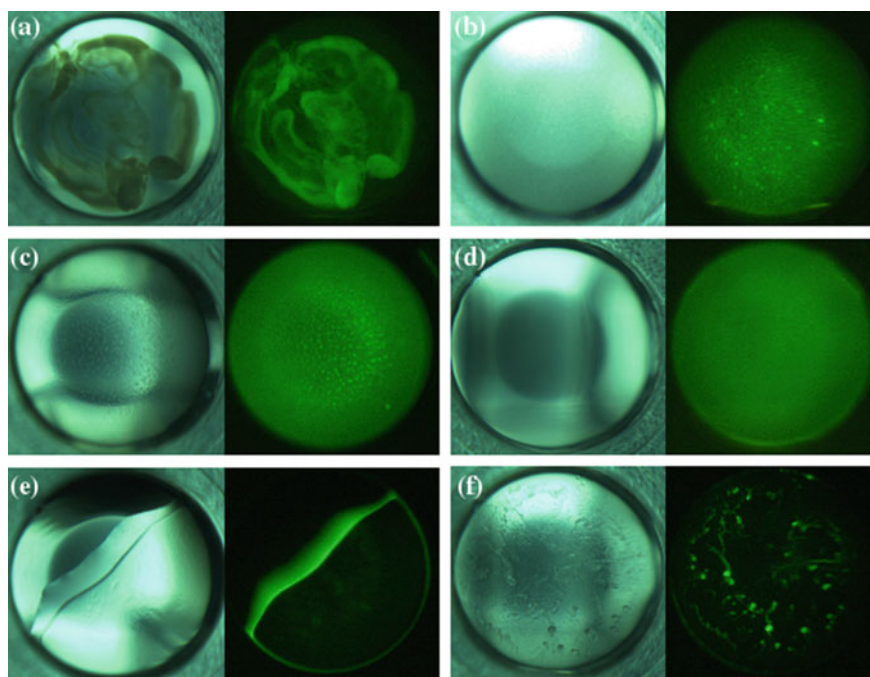


Fig. 2.1 Outcomes and their scores. Each panel has a white light and its corresponding fluorescence image. Panels A and B, score = 0, heavy precipitate, however note the presence of bright spots in Panel B, which would result in this being scored 4; Panel C, score = 1, light precipitate; Panel D, score = 2, clear solution; Panels E and F, score = 3, phase transition, although the presence of the bright spots in Panel F would result in this being scored 4. The protein in all images is canavalin, purified from Jack Bean

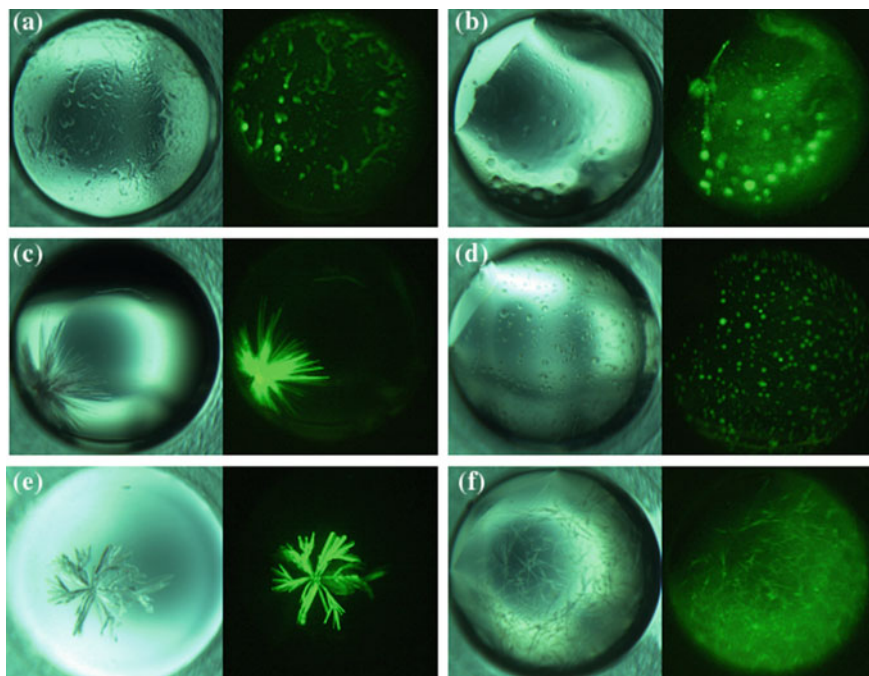


Fig. 2.2 Outcomes and their scores, each panel showing a white light and its corresponding fluorescence image. Panels A and B would have a starting score of 3, but due to the bright spots this would be increased to a score of 4. In the case of B the larger areas of intensity correspond to observable structures, and these may also be scored as 5. Panels C, D, and E, scores = 5, non-faceted crystals, with C being “urchins”, D being spheroids, and E dendrites. Panel F, score = 6, needles. Proteins are A, B - Canavalin; C, D, F - *Klebsiella pneumoniae* Inorganic pyrophosphatase, E - Tt36

precipitate is not clearly observable, unless one is using a non-white light imaging method. This is often a judgement call on the observers part. However, when making the distinction, one should attempt to be consistent in making that distinction for a given set of plates where the outcomes will be analyzed together. Also note that Panel A shows a heavy brown precipitate while B shows a heavy “white” precipitate. The nature of the precipitate in these cases is likely totally different, with one, B, likely being readily soluble while the other is likely not. Also note that the precipitate in B has a number of “bright spots” distributed throughout. As a result, the drop shown in panel B would be given a score of 0 during the manual examination, then that would be revised to a score of 4 during the fluorescence review.

The next difficulty comes in distinguishing a very light precipitate from a clear solution. While this is readily apparent in Fig. 2.1, Panels C and D, in many cases very light precipitates may not be visible. This is again a case where consistency in a set of plates is more important than accuracy. On occasion the results from other drops for that crystallization condition may suggest whether the solution is a

light precipitate or a clear solution. However, precipitate does often show up when observing the plates with fluorescent illumination.

Panels E and F of Fig. 2.1 show drops that would be scored as phase transitions during the initial white light scoring. Although not the case here, phase transitions are often spherical in shape, and can on occasion be mistaken for spheroids. In the case of Panel E the protein, which is fluorescent, has clearly separated from other components of the solution. This underscores the importance of having a means other than simple transmission microscopy for examining one's results. One may not know what solution components are separating out, whether they are the protein or some other (probably polymer) component. This distinction is important as we are most concerned with what the protein is doing in reaction to the solution conditions. Figure 2.1 Panel E shows the fluorescent image for a protein phase separation.

The white light image for Fig. 2.1, Panel F, shows an apparently "gummy" precipitate. We had initially scored this type of result as a heavy precipitate, but have shifted to scoring it as a phase transition. In large measure this is due to this type of precipitate often having bright spots, as shown in the fluorescent image, giving it a score of 4. As bright spot results can often be optimized to crystallization conditions, and as they are often associated with this type of precipitate, then we felt that it should be upgraded in the scoring to reflect the increased likelihood that these conditions may be on the path to crystallization.

The one novel scoring point, for our laboratory, is for the bright spots. Bright spots do not show crystalline features when viewed at higher magnification, and we currently assume them to be failed crystal nucleation's, where non-specific self-association kinetics have overtaken the orderly crystal self-assembly process. This is a score that is not made during the initial manual analysis phase but instead assigned when resolving those results with the fluorescent images. As we use trace fluorescent labeling for all of our crystallizations, the first pass image interpretation mantra is that intensity = structure. This is because the fluorescence intensity is a function of the density of the probe concentration, and the greatest protein, and thus probe, concentration will be had in the crystalline state. As shown in Fig. 2.2, Panels A and B, the bright spots can show up in a variety of "background" outcomes. We have found that in ~30% of the cases these conditions can be optimized to obtain crystals, and thus this score represents a major source of previously unknown lead conditions.

The most common crystalline outcome is often non-faceted crystals, which we assign a score of 5. Some of the most often observed of these are shown in Fig. 2.2, panels C, D, and E. Panel E shows what is typically referred to as an "urchin". These are often also observed as a more linear spray of needles commonly referred to as a "shaving brush". Panel D shows spheroids. Smaller spheroids can often be mistaken for phase separations. A distinguishing characteristic however is the presence of surface features or roughness on the spheroids. Panel E shows a dendritic crystal form. These may present as the stick-like crystal shown, as a snowflake-like feathery structure, or some intermediate form. This score is essentially a "catch-all" for any outcome that is crystalline, does not have clear facets, does not fit into one of the other categories, or includes clusters of crystals. The clusters can be stacks of 2D or agglomerations of 3D crystals. These sometimes have protruding single crystals,

and our rule of thumb is that if we think we can cleave a clear faceted region off then the structure, and thus the well, is given the score for that part.

Needles, a score of 6, often show up as shown in Fig. 2.2, Panel F. They can be clearly resolvable as individual needles, or they may be present as a dense cluster, or any outcome in between. Careful examination of the fluorescent image for Panel F shows an interesting light pipe phenomena often associate with needle (and rod) shaped crystals, where the ends fluoresce at a higher intensity than the body. This phenomenon is sometimes also found with 2D plates, and to a lesser extent with 3D crystals, where the edges are often more intense. We distinguish needles from rods by the presence, or not, of facets. If the ends of the crystals are clearly flat or have facets under higher magnification then they are scored as rods, 3D crystals, and not needles. This also holds true for the body of the crystal; if it is faceted then it is a rod, not a needle. Outcomes having a score of 5 or 6 are not suitable for diffraction analysis, but can be used as a source of material for seeded crystal growth [7].

2D Plates, having a score of 7, are shown in Fig. 2.3, Panel A. The last questionable distinction is between plates and 3D crystals. Again, this is often a judgement call. In our hands, it becomes a 3D crystal when there are clearly visible faceted edges,

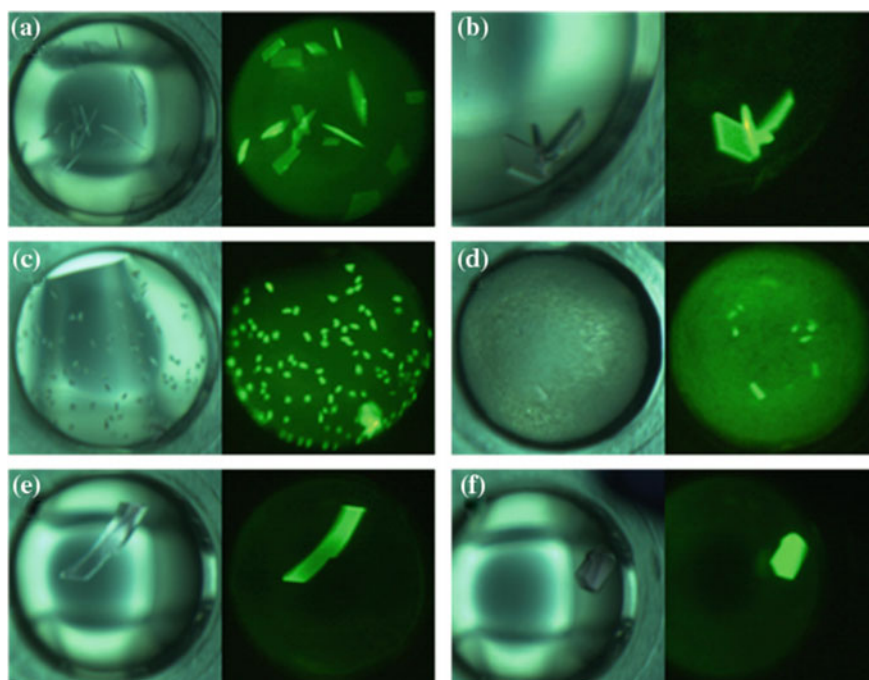


Fig. 2.3 White light and corresponding fluorescent images of scored plate outcomes. Panels A and B, score = 7, plate crystals. Panels C and D, score = 8, small 3D crystals size $\leq 200 \mu\text{m}$, with the crystals in Panel D being surrounded by precipitated protein. Panels E and F, large 3D crystals, size $\geq 200 \mu\text{m}$. All crystals are of *Klebsiella pneumoniae* inorganic pyrophosphatase

similar to the distinction between needles and rods. This is illustrated in Fig. 2.3, Panel B, where the plate-like crystals show a distinct edge, both in the white light and fluorescent image. Note that the body of the plate-like crystals, in both Panels A and B, fluoresces at a lower intensity. It is advantageous to have a microscope with a scale attached to gain some estimate of the thickness of the crystal. When imaging with TFL the edge typically shows up as a more intense fluorescence, particularly when it is partially or wholly oriented towards the viewing direction, while the body shows up with a lower fluorescence intensity when it is perpendicular to the viewing axis. However, the presence of straight fluorescent edges bordering a weaker fluorescence signals the presence of 2D plates where one may not have observed them using white light imaging.

Small 3D crystals, having a score of 8, are illustrated in Fig. 2.3, Panels C and D. Smaller crystals, or crystalline appearing material, should be examined under high magnification to look for the presence of faceted edges, which distinguishes them from non-faceted crystalline material having a score of 5. Panel D shows small crystals that are buried in precipitate. While they may not be apparent to cursory examination under transmission microscopy, they become readily apparent when viewed using fluorescence illumination.

The last category, and not surprisingly the least common, is large 3D crystals $\geq 200\mu\text{m}$ in size. Two examples of these are shown in Fig. 2.3, Panels E and F. Obtaining crystals of this size was once the goal of crystallographers several decades ago. However, as the X-ray technology has advanced these are not as desirable, except maybe as an experimental trophy. One potential benefit is that they do show that one can obtain larger crystals of that particular protein, thus suggesting that at some future date one could carry out neutron diffraction studies on it.

2.4.2 Hierarchical Categories

Classifying protein crystallization trial images into a *number of categories* is one of the main tasks in analysis. However, the key point in such analysis is to determine the categories. The number of categories is usually determined based on the purpose of the analysis. In the literature, we have observed that typically the number of categories is between 2 and 10. Since the most common goal is to detect the presence of a crystal, the use of two categories as crystals and non-crystals is not rare (Zuk and Ward [18], Cumba et al. [6], Cumba et al. [4], Zhu et al. [17], Berry et al. [2], Pan et al. [9], Po and Laine [10]). Additional categories are typically obtained by using sub-categories of these two main categories and erroneous/mistake/unclear or doubtful categories. Clear, precipitate, and crystal categories are three categories used by Yang et al. [15]. The five categories analyzed by Bern et al. [1] are empty, clear, precipitate, microcrystal hit, and crystal categories. Another group of five categories is formed as clear drop, creamy precipitate, granulated precipitate, amorphous state precipitate, and crystal categories by Saitoh et al. [12]. An example of six categories includes experimental mistake, clear drop, homogeneous precipitant, inhomogeneous precip-

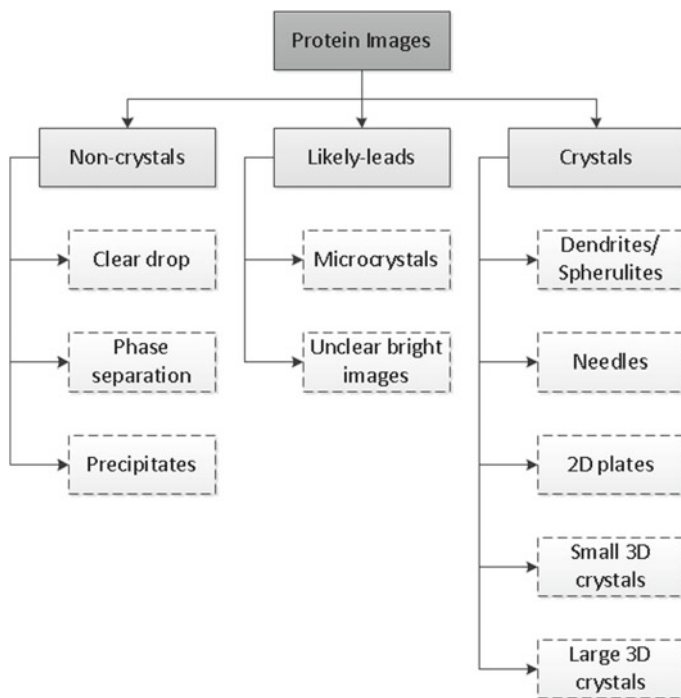


Fig. 2.4 Hierarchy of crystallization categories [13]

itant, microcrystals, and crystal categories (Spraggon et al. [14]). Another example of six categories includes phase separation, precipitate, skin effect, crystal, junk, and unsure categories (Cumba et al. [5]). 10 categories used by [16] are (clear, precipitate, crystal, phase, precipitate and crystal, precipitate and skin, phase and crystal, phase and precipitate, skin, and junk categories).

There is not a perfect system that would classify crystallization trial images into any number of categories. However, binary categorization as crystals vs. non-crystals should be avoided where possible. The costly misclassification occurs when a crystal is classified as a non-crystal. To detect such an error, the expert analyzes non-crystal images in addition to crystal images categorized by the system to avoid missing crystals. This would suggest checking all images in the experiment, thus losing the value of a classifier. In our work, we have added one more category in between crystals and non-crystals as likely-leads. Classifying crystals as likely-leads is not a major problem as long as crystals are not labeled as non-crystals and false positive rate where non-crystals labeled as crystals is low.

Depending on the depth of analysis of protein crystallization images, the classification could be performed roughly as non-crystals, crystals, and likely-leads or for the sub-categories of these categories as mentioned in the previous section. For analyzing protein crystallization trial images, we generally use two levels of hierarchical

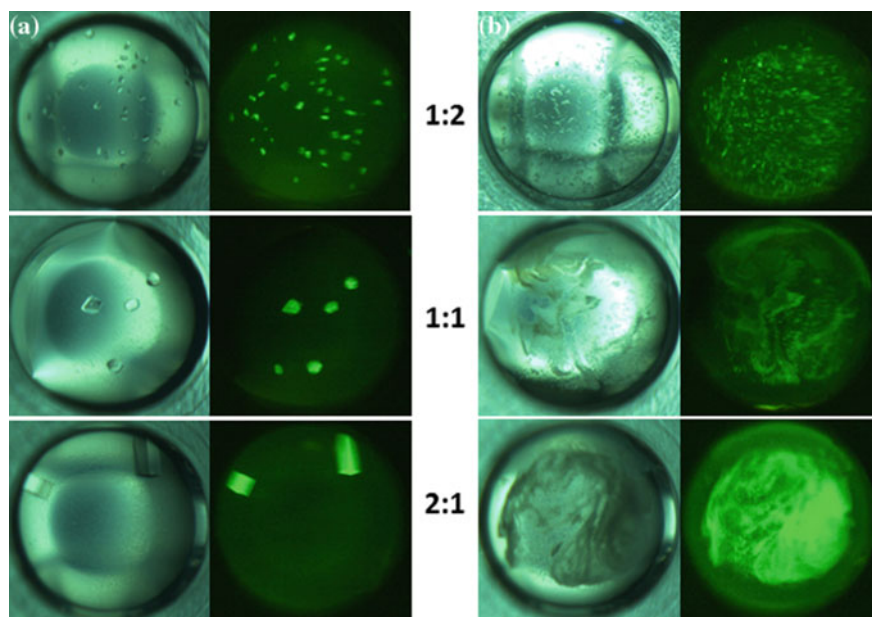


Fig. 2.5 Variations in crystallization drop outcome with changing protein:precipitant drop volume ratios. Column A shows a progression of smaller to larger, while column B shows a progression of crystals to precipitate

categorization as shown in Fig. 2.4. This hierarchy helps develop classifiers for the first level and then classifiers for each category of the first level. Such a hierarchy enables developing classifiers at two levels.

2.5 Even if You Are Not Going to Process Your Scored Data...

Careful scoring requires attention to and consideration of what is being observed. Insights beyond just the assignment of a numerical score can be obtained, which can be utilized to direct subsequent crystallization optimization experiments. For example, random, or limited grid, screens where one or two components are varied would be expected to show a trend in the results with the changes in the components. This is the reason for using such a screen, and justifies its utility. Similarly however we can also observe trends in screen set-ups where we vary the protein:precipitant drop ratio. Two examples of this are shown in Fig. 2.5, both coming from the same screening plate for the protein concanavalin A. In both cases the protein:precipitant drop volume ratios are 1:2, 1:1, and 2:1. A first pass assumption is that at equilibrium the drop and well precipitant concentrations are equivalent to those of the reservoir

solution for all three cases, although this assumption does not take into account the protein contribution to the crystallization drop vapor pressure. For the 1:2 drop the initial precipitant concentration is high relative to the protein and the equilibrium protein concentration would be low. In the results shown for Panel A we find that this results in many smaller crystals. At a 1:1 ratio, the “standard” crystallization drop, we find fewer but larger crystals, while at 2:1 the starting precipitant concentration is low and the equilibrium protein concentration higher, likely higher than the starting protein concentration. In this case fewer nucleation events occur and there is more protein available to feed the growth of larger crystals. These results can be mapped directly onto the phase diagram in Fig. 1.1. The results in Panel B show an opposite effect. Crystals are nucleated at higher precipitant and lower protein concentrations. As the protein:precipitant ratio increases the results go to increased precipitation and no crystals, the opposite of what is observed for the wells in Panel A. The precipitants in Panel A are ethylene glycol and Polyethylene glycol 8,000, while those in Panel B are sodium chloride and sodium citrate. The results give a good indication of how one should adjust the protein and/or the precipitant concentration in both cases for obtaining optimized crystallization results.

2.6 Summary

While it may appear to be tedious at the outset, the scoring of one’s crystallization screening experiments can be a very productive exercise. Firstly, one has a more detailed description of the results obtained for the experiment(s), which can always be referred to in subsequent work. Secondly, the act of scoring directs the mind towards considerations of how or why those results were obtained, possible new approaches that can be tried to obtain crystals, and insights into how one can optimize found crystallization conditions. Time spent carefully observing, and scoring, one’s crystallization results can speed up progress towards the ultimate goal, well diffracting crystals.

References

1. Bern, M., Goldberg, D., Stevens, R. C., & Kuhn, P. (2004). Automatic classification of protein crystallization images using a curve-tracking algorithm. *Journal of Applied Crystallography*, 37(2), 279–287.
2. Berry, I. M., Dym, O., Esnouf, R., Harlos, K., Meged, R., Perrakis, A., et al. (2006). Spine high-throughput crystallization, crystal imaging and recognition techniques: current state, performance analysis, new technologies and future aspects. *Acta Crystallographica Section D: Biological Crystallography*, 62(10), 1137–1149.

3. Brodersen, D. E., Andersen, G. R., & Andersen, C. B. F. (2013). Mimer: an automated spreadsheet-based crystallization screening system. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, 69(7), 815–820.
4. Cumbaa, C., & Jurisica, I. (2005). Automatic classification and pattern discovery in high-throughput protein crystallization trials. *Journal of Structural and Functional Genomics*, 6(2–3), 195–202.
5. Cumbaa, C. A., & Jurisica, I. (2010). Protein crystallization analysis on the world computing grid. *Journal of Structural and Functional Genomics*, 11(1), 61–69.
6. Cumbaa, C. A., Lauricella, A., Fehrman, N., Veatch, C., Collins, R., Luft, J., et al. (2003). Automatic classification of sub-microlitre protein-crystallization trials in 1536-well plates. *Acta Crystallographica Section D: Biological Crystallography*, 59(9), 1619–1627.
7. D’Arcy, A., Bergfors, T., Cowan-Jacob, S. W., & Marsh, M. (2014). Microseed matrix screening for optimization in protein crystallization: what have we learned? *Acta Crystallographica Section F: Structural Biology Communications*, 70(9), 1117–1126.
8. Forsythe, E., Achari, A., & Pusey, M. L. (2006). Trace fluorescent labeling for high-throughput crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 62(3), 339–346.
9. Pan, S., Shavit, G., Penas-Centeno, M., Xu, D. -H., Shapiro, L., Ladner, R., et al. (2006). Automated classification of protein crystallization images using support vector machines with scale-invariant texture and gabor features. *Acta Crystallographica Section D: Biological Crystallography*, 62(3), 271–279.
10. Po, M. J., & Laine, A. F. (2008) Leveraging genetic algorithm and neural network in automated protein crystal recognition. In *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2008*. (pp. 1926–1929). IEEE.
11. Pusey, M., Barcena, J., Morris, M., Singhal, A., Yuan, Q., & Ng, J. (2015). Trace fluorescent labeling for protein crystallization. *Acta Crystallographica Section F: Structural Biology Communications*, 71(7), 806–814.
12. Saitoh, K., Kawabata, K., & Asama, H. (2006). Design of classifier to automate the evaluation of protein crystallization states. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, ICRA 2006* (pp. 1800–1805). IEEE.
13. Sigdel, M., Dinc, I., Sigdel, M. S., Dinc, S., Pusey, M. L., & Aygun, R. S. (2017). Feature analysis for classification of trace fluorescent labeled protein crystallization images. *BioData Mining*, 10, 14.
14. Spraggon, G., Lesley, S. A., Kreuzsch, A., & Priestle, J. P. (2002). Computational analysis of crystallization trials. *Acta Crystallographica Section D: Biological Crystallography*, 58(11), 1915–1923.
15. Yang, X., Chen, W., Zheng, Y. F., & Jiang, T. (2006). Image-based classification for automating protein crystal identification. *Intelligent computing in signal processing and pattern recognition* (pp. 932–937). Berlin: Springer.
16. Yann, M. L. -J., & Tang, Y. (2016). Learning deep convolutional neural networks for x-ray protein crystallization image analysis. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*.
17. Zhu, X., Sun, S., & Bern, M. (2004). Classification of protein crystallization imagery. In *26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEMBS’04*. (Vol. 1, pp. 1628–1631). IEEE.
18. Zuk, W. M., & Ward, K. B. (1991). Methods of analysis of protein crystal images. *Journal of Crystal Growth*, 110(1), 148–155.

Chapter 3

Computational Methods for Protein Crystallization Screening

Abstract The goal of protein crystallization screening is to determine the main factors of importance to crystallize a protein under investigation. The protein crystallization screening is often expanded to many hundreds or thousands of conditions to maximize combinatorial chemical space coverage for maximizing the chances of a successful (crystalline) outcome. Available commercial screens may not generate crystalline conditions for some proteins difficult to crystallize. Nevertheless, the previous crystallization trials could be analyzed to recommend screens with crystalline conditions. This chapter presents computational methods for protein crystallization screening.

3.1 Introduction

Crystallization is usually the bottleneck process in the determination of the three-dimensional structure of a protein. One of the major difficulties in macromolecular crystallization is setting up the cocktails that yield a single large crystal for X-ray data collection [20, 21]. Physical, chemical, and biochemical factors such as the type of precipitants, type of salts, concentrations, pH value of the buffer, temperature of the environment, genetic modifications of the protein, etc., influence the crystallization process. The large chemical space along with varying concentration values of chemicals makes exhaustive trial of all possible combinations practically impossible.

For example, consider generating a screen using nine different salt concentration values, 23 different salts, nine different buffers, 26 different precipitant concentration values, 38 different precipitants, and three different protein concentration values. The concentrations and pH values are continuous data and the other features are categorical data. Full factorial design for this single protein would require setting approximately 5,521,932 different experiments based on this set of factors without considering the continuity of some of the variables, which is not feasible.

Moreover, since each protein has a unique primary structure, it is quite challenging to determine the parameters of the experiment that can yield a crystal for a protein [19, 23]. The cost in time and materials renders exhaustive trial of all

possible combinations of conditions practically impossible. As a result determining the crystallization conditions is conducted using screening experiments.

One of the major challenges for an experimenter is to analyze previous crystallization trials to determine new conditions to be tested. This chapter firstly provides an overview of how initial screens are prepared for crystallization experiments. These initial experiments are critical and ideally should provide some likely-lead conditions if not crystalline conditions. New screens could be generated by analyzing encouraging conditions from these previous experiments. Nevertheless, this is not computationally straightforward. As expected, the noncrystal conditions are significantly dominant, and while there could be at most several likely-lead conditions, they may not be suitable for X-ray crystallography. Classifiers built for these few encouraging conditions cannot learn with such unbalanced and skewed data. Similarly, regression methods cannot fit data with satisfactory R^2 [13]. Association rule mining cannot generate association rules with enough support and confidence.

Three computational methods for screening are discussed in this chapter: neural networks, genetic algorithms, and associative experimental design [13, 14]. Since these methods provided successful results by analyzing crystallization trials of some proteins, they are worthy to be analyzed here. Moreover, associative experimental design is actively used for generating novel crystalline conditions. Optimization of cocktails and its success for recommending novel conditions for a number of proteins has been provided later in this chapter.

3.2 Overview of Experimental Design Methods for Screening

If an experimenter would like to crystallize a protein, her question will be about the parameters to evaluate in the wet lab. The parameters for protein crystallization experiments are usually set by two main techniques [8, 34]: incomplete factorial experiments (*IFE*) [4, 10] or sparse matrix sampling (*SMS*) [16, 20]. Systematic grid screening (*GS*) of crystallization conditions is a complete factorial screen over a defined range.

The incomplete factorial approach aims to determine the important factors of the experiments and to significantly reduce the number of experiments compared to full factorial design experiments [10]. The *IFE* is a beneficial tool especially when there are not enough resources, such as available protein, to carry out those many experiments or it is practically discouraging to set up many experiments [22]. The *IFE* method generates balanced experiments with respect to the important factors of the experiments. In *IFE*, balanced crystallization screening experiments are generated using selected reagents, which allows analysis of a broad chemical space. One of the drawbacks of *IFE*, the occurrence of each reagent for a factor is equal in the experiments; however, in the real world, some reagents might be more favorable for the crystallization trials compared to others [24].

The sparse matrix sampling (*SMS*) method [20] utilizes a wider range of major reagents conditions (i.e., pH values, type of precipitants, type of salts, etc.) in experiments. In *SMS*, type of salts, pH, and type of precipitants and their values are selected based on past experience to have resulted in protein crystallization. The reagents appear based on their frequency in the sparse matrix [16]. The sparse matrix approach was first put forth by Jancarik and Kim (1991), and their original screen, plus a wide range of variations, has been commercialized [33]. *SMS* tries to overcome the limitations of *IFE* by increasing the occurrence of the reagents that are more favorable for the experiments based on existing experimental results. The frequency of each chemical used in *SMS* is generally calculated based on accumulated experimental results.

Grid screening of crystallization conditions is an early method that methodically varies a set of solution components over a range of conditions. This typically requires some insight into those parameters likely to produce crystals and is more often carried out as part of the end game process following the successful determination of lead conditions by sparse matrix methods. In *GS*, the experts generally focus on a small chemical space and generate finer samples for a small set of reagents, making this impractical for covering extensive chemical space.

Once the results of these methods are obtained, a set of optimization methods can be applied [22]. The details of those optimization techniques can be found in the literature. These studies in macromolecular crystallization try to generate new cocktails or optimize available cocktails, which are supposed to yield crystals. The optimization steps in the literature generally involve changing the pH, concentration, and concentrations of precipitants and salts.

3.3 Using Neural Networks for Experimental Design

Researchers at Diversified Scientific Inc., the University of Alabama at Birmingham and Interactive Analysis Inc. have utilized neural networks for protein crystallization screening [12]. Neural networks are based on a real nervous system paradigm composed of multiple neurons communicating through axon connections. Characteristics of neural networks include self-organization, nonlinear processing, and massive parallelism. The neural network exhibits enhanced approximation, noise immunity, and classification properties. The self-organizing and predictive nature of the neural networks allow for accurate prediction of never before seen crystallization conditions, even in the presence of noise. The predictive neural network is trained via back propagation using the incomplete factorial screen. If properly trained, the neural network can be used to identify or recognize important patterns of crystallization. An input pattern comprised of the incomplete factorial screen is presented to the network. The outputs are compared to the known scores. Additional neurons are added and interconnect weights (basis functions) are adjusted to minimize the error and maximize R^2 between the actual versus the predicted values. This process is continued until the average error across all the training sets is minimized. Eventually, if the correct vari-

ables and sample size are chosen to adequately represent the crystallization nature of the protein, a stable set of hidden neurons and basis function weights evolve. This neural network can then be used to predict non-sampled complete factorial conditions to be used for optimization, i.e., predicting the conditions that produce crystals from the entire “crystallization space” of possible experimental conditions based on the results from a much smaller number of actual experiments performed. This approach has a higher probability of producing accurate predictions if the small test set is statistical representative of the “crystallization space”.¹

DeLucas et al. provide the results for 9C9 (*C. elegans* protein) to display performance of neural networks [12]. For these initial experiments, the neural network was trained using experiments 1–315 from the complete set of 360 screen conditions. This partial sampling of the incomplete factorial design experiment was used to train a neural network to recognize conditions that result in crystallization. The neural network trained with all results, including failures. The 315 experiments (used for training) allowed the neural network to converge with an acceptable R^2 value of 0.604. The scoring system was modified from a linear scale with clear drops equal to 0 and crystals scored at 10, to a binary scheme. In the binary scheme any crystalline result was given a mark of 2000, the other results (i.e., clear drop, phase separation, precipitate, microcrystals/precipitate, and rosettes/spherulites) were scored 1–5, respectively. The input to the neural network is the indexed variables and the output is the predicted score. The weights of the hidden neurons are determined by back propagation. The remaining 12.5% (45 experiments) of the incomplete factorial screen results were used for verification.²

Only one experiment had a crystalline condition in the training set. Similarly, only one experiment had a crystalline condition in experiments 316–360. The proposed neural network was able to detect the crystalline experiment while generating low scores for unsuccessful experiments.

Their use of binary scheme with significantly different scores (0 and 2000) helped the neural network to adjust for crystalline conditions. It would be interesting to analyze how good neural networks cover crystallization space based on the scored input screens.

¹Reprinted from *Progress in Biophysics and Molecular Biology*, Volume 88, Issue 3, Lawrence J. DeLucas, David Hamrick, Larry Cosenza, Lisa Nagy, Debbie McCombs, Terry Bray, Arnon Chait, Brad Stoops, Alexander Belgovskiy, W. William Wilson, Marc Parham, Nikolai Chernov, Protein crystallization: virtual screening and optimization, Pages 285–309, Copyright (2005) with permission from Elsevier.

²Reprinted from *Progress in Biophysics and Molecular Biology*, Volume 88, Issue 3, Lawrence J. DeLucas, David Hamrick, Larry Cosenza, Lisa Nagy, Debbie McCombs, Terry Bray, Arnon Chait, Brad Stoops, Alexander Belgovskiy, W. William Wilson, Marc Parham, Nikolai Chernov, Protein crystallization: virtual screening and optimization, Pages 285–309, Copyright (2005) with permission from Elsevier.

3.4 Genetic Algorithm for Protein Crystallization Screening

Genetic algorithms are computational methods inspired by natural selection in biological evolution to solve constrained and unconstrained optimization problems. Interesting enough, these algorithms could be used to solve back another biochemical process, protein crystallization domain. Genetic algorithm represents input as a set of chromosomes of genes. Two fundamental operators are crossover and mutation. The selection process of parent chromosomes for crossover as well as mutation rate affects the performance of genetic algorithms. As the new population is generated after each iteration, fitness score or in other words, survivability rate of an offspring plays a critical role in selecting parents and ranking offsprings.

For protein crystallization screening, the mutation is critical to explore the large chemical space or generate novel conditions. Saridakis presents a successful way of applying genetic algorithms for protein crystallization [31]. The parameters to be optimized can be thought of as genetic loci on a virtual chromosome. Each value of the parameter is an allele. The whole “chromosome” is thus a full set of parameters with specified values (in this case, a crystallization condition). A few “successful chromosomes” (crystallization hits) are selected from a “parent generation” (a crystallization screen) and their alleles (parameter values) are recombined to form the next “generation” of “chromosomes” (candidate optimization conditions). From that second generation, the most successful conditions are again selected and the process is reiterated. Sometimes a ‘mutation’ is introduced, that is, a parameter is randomly selected, and its value randomly changed to a completely new value, ideally one that was not present in the original screen at all. Mutations can be simple, multiple, or they can be mixed with recombinations. For the protein crystallization case, a chromosome may be specified as follows:

$$C_{1a} = \{[protein]_i, precipitant_k, [precipitant]_l, temperature_m, pH_n, additive_o, [additive]_p, [ligand]_q, \dots\} \quad (3.1)$$

where i, k, l, \dots are the different discrete or continuous, numerical or descriptive, values that the respective parameters may take and the square brackets signify concentration. Thus a particular condition may for instance be:

$$C_{1a} = \{[proteinX]_{20mg/mL}, precipitant_{NaCl}, [precipitant]_{4\%(w/v)}, temperature_{20^\circ C}, pH_{4.5}, additive_{PEG4000}, [additive]_{2\%(w/v)}, \dots\} \quad (3.2)$$

Another condition may be:

$$C_{1b} = \{[proteinX]_{20mg/mL}, precipitant_{amm.phosphate}, [precipitant]_{1.5M}, temperature_{4^\circ C}, pH_{6.5}, additive_{KCl}, [additive]_{0.1M}, \dots\} \quad (3.3)$$

Assuming that the above two conditions are hits (neither of which can be optimized by conventional fine-tuning of the variables), one of the possible recombinations is as follows:

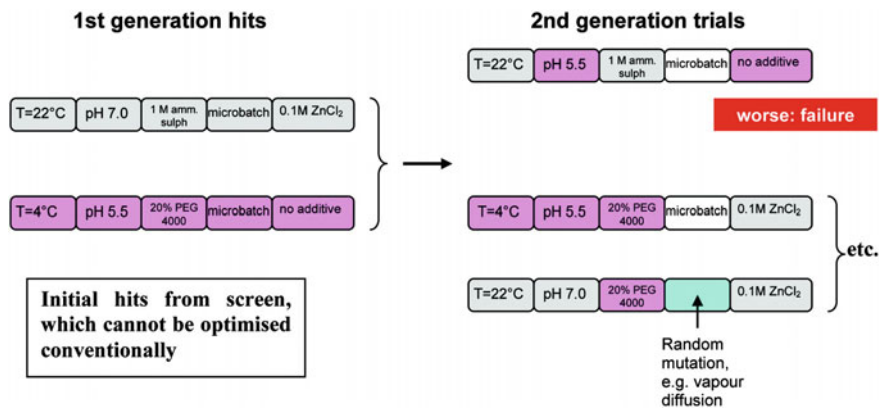


Fig. 3.1 Simplified schematic illustration for one cycle of a Genetic Algorithm-inspired procedure as applied to only two “hits” from a standard crystallization screen, used to generate three 2nd generation conditions. A mutation regarding the setup technique is also introduced. Reprinted (adapted) with permission from Crystal Growth and Design 2011 11 (7), Emmanuel Saridakis, Novel Genetic Algorithm-Inspired Concept for Macromolecular Crystal Optimization, 2993–2998. Copyright (2011) American Chemical Society ©2016 IEEE

$$C_{2a} = \{[proteinX]_{20mg/mL}, precipitant_{amm.phosphate}, [precipitant]_{1.5M}, temperature_{20^{\circ}C}, pH_{4.5}, additive_{PEG4000}, [additive]_{2\%(w/v)}, \dots\} \quad (3.4)$$

where the subscripts 1 and 2 denote the successive generations.

A few such “recombinant” (or “mutated,” or mixed) conditions are randomly generated, by hand or with the help of a computer depending on the number and complexity of the hits, and some or all are set up. Since it is easier to design a great number of second generation conditions that to actually set them up, the experimenter’s intuition may be used to select the ones that are actually going to be set up. Too much use of intuition nevertheless might lead to missing unlikely but successful conditions. The second generation trials are inspected and the procedure may stop there if interesting crystals are found in one or more drops, or the process can be reiterated to form a new generation of conditions. A simplified schematic illustration of one cycle of such a procedure is shown in Fig. 3.1.

Saridakis [31] analyzes each population after each iteration and decides experiments to set up. When it is considered that a new population may not yield successful experiments, the experimenter may stop (e.g., after the second iteration if desired). In a typical genetic algorithm, populations are generated after a number of iterations (e.g., 100). The latest population is used for experiments. The major challenge of this approach for protein crystallization screening is to use a viable fitness function. In other words, how can a new condition be given a score higher than others without experimenting? When a reliable fitness score is defined, the genetic algorithm may run a number of iterations, and the experiments may utilize the last population of conditions.

3.5 Associative Experimental Design

The associative experimental design (AED) [14, 15] analyzes possible interactions between reagents to determine new crystallization conditions. By analyzing the outcome of preliminary experiments, the *AED* generates candidate cocktails identifying screening factors that are most likely to lead to higher scoring outcomes, crystals. Thus, *AED* is not just an optimization method for crystallization conditions, since it could generate novel conditions leading to crystals.

Associative experimental design generates a new set of experiment conditions by analyzing the scores of screening experiments already carried out in the lab. Plate results are scored over the range 0 to 9, as listed in revised scores column in Table 2.1. For trace fluorescent labeling (*TFL*) [17], a score of 4 is assigned to outcomes giving “bright spot” lead conditions. For *AED* let

$$D = \{(C_1, H_1), (C_2, H_2), \dots, (C_n, H_n)\} \quad (3.5)$$

be the dataset containing the pairs that include features of the conditions C_i and their scores H_i for the i th solution in the dataset. For simplicity, this version does not include conditions that have more than one type of salt or precipitant. *AED* uses the three main components of the remaining conditions: type of precipitant, type of salt and pH value of the solution, while separating their concentrations. Let

$$C_i = \{S_i [sc_i], pH_i, P_i [pc_i]\} \quad (3.6)$$

be the set of reagents of the i th crystal cocktail where n is the number of samples in the dataset, $1 \leq i \leq n$, $S_i [sc_i]$ represents type of salt with the concentration of sc_i , pH_i value represents the pH of i th solution, and $P_i [pc_i]$ represents type of precipitant with the concentration of pc_i . Let R be a subset of D that contains the crystal cocktail pairs with a score greater than or equal to low_H and less than or equal to $high_H$:

$$R = \{(C_i, H_i) \mid (C_i, H_i) \in D, low_H \leq H_i \leq high_H, 1 \leq i \leq n\} \quad (3.7)$$

In the preliminary experiments, the low score is set to 4 ($low_H = 4$) and the high score is set to 7 ($high_H = 7$). Therefore, the samples that have a score of 8 or 9 are excluded to generate unbiased conditions for proteins. However, there is no harm to include these scores, as well. Similarly, for simplicity the samples with scores from 1 to 3 have not been included to the result set.

The *AED* analysis process consists of two major phases. In the first phase, the data is processed to reduce its size as stated before. Let

$$R_c = \{C_i \mid (C_i, H_i) \in R\} \quad (3.8)$$

denote the set of conditions of R , $SC_i = \{sc_1, sc_2, \dots, sc_k\}$ represents the all unique concentration values of the i th salt, and $PC_i = \{pc_1, pc_2, \dots, pc_k\}$ represents the all

unique concentration values of i th precipitant. Then, all C_i and C_j condition pairs are compared in R_c where $i \neq j$. If C_i and C_j have a common component, then the candidate conditions' set Z is generated based on these two sets. For example, assume that $C_i = \{S_i [SC_i], pH_i, P_i [PC_i]\}$ and $C_j = \{S_j [SC_j], pH_j, P_j [PC_j]\}$ where $S_i = S_j$ (i.e., the type of salt is common in C_i and C_j). Two new conditions Z are generated by swapping the other components among each other. Therefore,

$$Z = \{\{S_i [SC_i], pH_j, P_i [PC_i]\}, \{S_i [SC_i], pH_i, P_j [PC_j]\}\} \quad (3.9)$$

is the set of candidate crystal cocktails for the pair C_i and C_j . In a similar way, candidate cocktails can be generated where pH value or precipitant is common between the pairs as well. After generating candidate combinations using these components, conditions that are replicated or already in the screening data (i.e., have known outcomes) are removed. In the second phase of this method, unique values of concentrations are assigned to generate SC_i and PC_i , and unique type of buffers that were used in the preliminary data is assigned to generate finalized crystal cocktails. At the end, the results from two phases of the method are merged. Then, if the number of candidate conditions are more than the desired number of cocktails or there are some bad combinations which are proved empirically, an optimization method is applied to generate a set of conditions as mentioned in the following section. Examples of bad combinations are those known to result in a phase separation or where the two reagents react to form salt crystals. The steps of *AED* before optimization is provided below.

1. Data preprocessing.
2. Generate a list of cocktails score between 4 and 7.
3. Generate triplets of salt, type of precipitant, and pH value.
4. Find common reagents between each triplet pairs.
5. Generate two new cocktails by swapping different reagents.
6. Generate unique concentration values for each specific reagent.
7. Assign concentration values.

In order to increase robustness, after obtaining the preliminary results from *AED*, the family of the conditions from the cocktails having score 7, 8, or 9 for some of the proteins is generated. Basically, the cocktails in a family consist of the same type of buffer, precipitant, and salt with different concentrations. In the experiments, it was possible to get multiple crystals for a single family. In other words, the number of crystals in a family shows the robustness, the stability, and the reproducibility of that family. In Sect. 3.7.2, brief information about these family of conditions is provided.

Sample Scenario Fig. 3.2 shows the scores from four experiments using a commercial screen. The figure shows a partial graph of scores for common pH value of 6.5. These conditions generated four scores: 1, 1, 4, and 4. As it can be seen, none of the conditions lead to a good crystallization outcome for these conditions. The *AED* method determines the common reagent between solutions that could lead crystallization conditions. In this example, there are only two

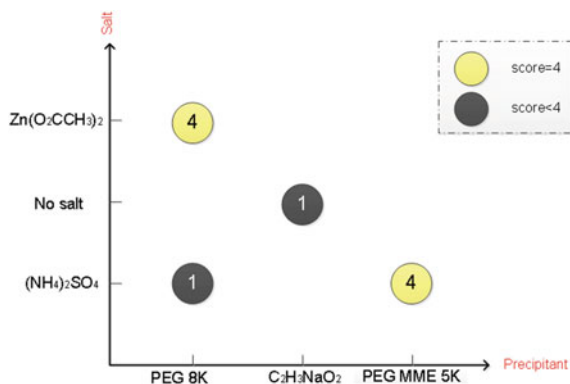


Fig. 3.2 Sample preliminary results of experiments for AED ©2016 IEEE

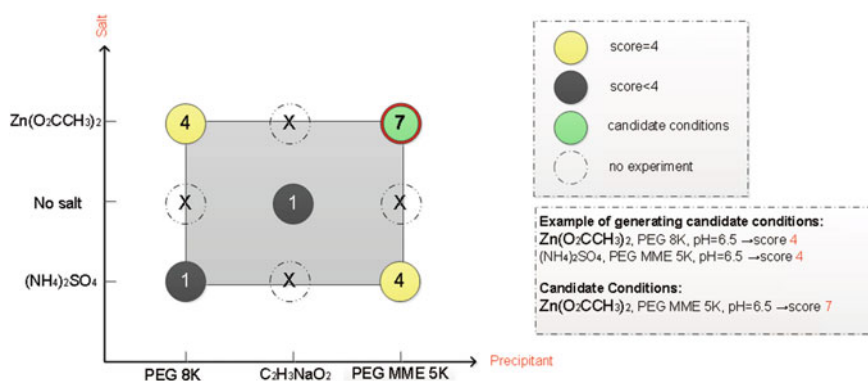


Fig. 3.3 Visual example for AED ©2016 IEEE

promising conditions (with score 4): $[Zn(O_2CCH_3)_2, PEG\ 8K, pH = 6.5]$ and $[(NH_4)_2SO_4, PEG\ MME\ 5K, pH = 6.5]$. The AED draws a rectangle where these conditions (with score 4) are the two corners of this rectangle (Fig. 3.3) and the other corners represent the candidate conditions. This scenario has two possible candidate conditions. One of them ($[(NH_4)_2SO_4, PEG\ 8K, pH = 6.5]$) already appeared in the commercial screen and yielded a low score. After conducting the experiment for the other condition ($[Zn(O_2CCH_3)_2, PEG\ MME\ 5K, pH = 6.5]$), a score of 7 was obtained after optimizations. The experiments have not been conducted for others in the figure since they were not on the corners of conditions with promising scores.

3.6 Optimization of Cocktails

Computational methods such as the *AED* may generate many candidate cocktails. The output of these methods should be optimized by eliminating prohibited combinations and prioritizing reagents based upon their performance in the input screens. After identifying initial screens from outputs of these methods, the combinations known to produce a precipitate are eliminated. These combinations are identified either from the literature (for example [5, 6, 30]) or by empirical observation based on lab experiments. This section adopts the optimizations done by the *AED* and describes optimizations with respect to the *AED* analysis.

In the *AED* analysis, the number of candidate cocktails depends on the number of cocktails that have scores in a range (e.g., from 4 to 7) in the input data. When *AED* generates more cocktails than the desired number (e.g., the number of wells in a plate) of cocktails, the experts may want to try the most promising candidate cocktails that need to be set. For example, if *AED* generates 150 candidate cocktails, the expert may want to know 96 cocktails to be tried for a 96-well plate. To resolve this problem, an optimization process is employed to eliminate cocktails having poor combinations of reagents and to prioritize the remaining conditions based on a metric. The following stages are described in the following sections:

1. Eliminate prohibited combinations.
2. Prioritize remaining combinations.
3. Optimize the concentration values.
4. Rank prioritized cocktails.

3.6.1 Elimination of Prohibited Combinations

The output from the *AED* analysis usually results in more solution combinations than were present in the initial screen(s). The *AED* analysis indicates all of the possible unique combinations, and these are reduced to the final solutions by two processes. First is to remove “prohibited” combinations of reagents, such as mixtures known, either from the literature (for example [5, 5, 6]) or by empirical observation, to produce a precipitate, to produce a phase separation (e.g., high concentrations of PEG and a salt), those known to produce a precipitate, such as mixtures of divalent cations with particular anions such as phosphate or sulfate, or those that would tend to remove one or more of the components as unique entities in the solution, such as mixing divalent cations with diacid chelators such as EDTA or citrate. Additional unfavorable pairings are added to this list, as they are empirically determined. Additionally, the output does not (yet) take into account the feasibility of attaining the final solutions on the basis of the available stock solution used for formulation. Thus, for example, stock trisodium citrate is 1.6 M. A solution calling for 0.1 M buffer, 1.6 M citrate, and possibly a third component cannot be made using the available stocks. Redundant

outputs are also removed, such as 0.1 M citrate buffer with citrate as precipitants 1 and 2.

3.6.2 Prioritization of Reagents

The second step of the optimization is a simple prioritization of the reagents for their association with better scoring outcomes. In this stage, the list of the reagents and scores is sorted with respect to the class of reagent being analyzed (buffer, precipitant, salt, etc.). For a candidate cocktail C that consists of precipitant p , buffer b , and salt s as reagents, the ratio of the average of the scores for the component of interest versus all other scores is determined by the ranking.

Let δ_p , δ_s , and δ_b represent the scores of the cocktails having precipitant p , salt s , and buffer b for a given screen file, respectively. Let Δ represent all scores of the input file. Then, the significance ratio, $\rho(\delta_r)$ for each class of reagent: precipitant, salt, and buffer, is computed as $\frac{\mu(\delta_p)}{\mu(\Delta-\delta_p)}$, $\frac{\mu(\delta_s)}{\mu(\Delta-\delta_s)}$, and $\frac{\mu(\delta_b)}{\mu(\Delta-\delta_b)}$, respectively. Those with significance ratio greater than 1 ($\rho(\delta_r) > 1$) perform better than the average while those with significance ratio less than 1 ($\rho(\delta_r) < 1$) perform worse. After identifying the components with highest significance ratios for each category, those components appearing with high significance ratios are tried in the wet lab.

Once the composition of the 96 conditions for the *AED* optimization screen has been determined, a pipetting table is generated to produce a block of 96 solutions of 1 mL volume, using the desired final concentrations for each reagent and the stock solution concentrations. In some cases, the stock reagent concentrations are not sufficiently high to produce the desired final solutions, typically indicated by a negative value for the calculated distilled water added to bring the solution to the final volume. In such cases, either the concentration of one of the precipitants is reduced or an alternative set of solutions are used.

3.6.3 Ranking of Prioritized Conditions

In screen designing, it is important to know that whether a result cocktail is close to another cocktail in the input screen data to make a judgment about its outcome or priority. Chemical distance is a useful tool to evaluate the relationship between cocktails [25]. A ranking method based on closeness of the prioritized cocktails to the crystal cocktails in the preliminary data is applied to sort the prioritized cocktails generated by *AED*. For example, in Fig. 3.4, assume that the green points indicate the crystal cocktails with scores 4, 5, 6, 7, or 8, and red points indicate the *AED* results. The candidates close to the green points may have a higher chance to yield a good crystal compared to the other candidates.

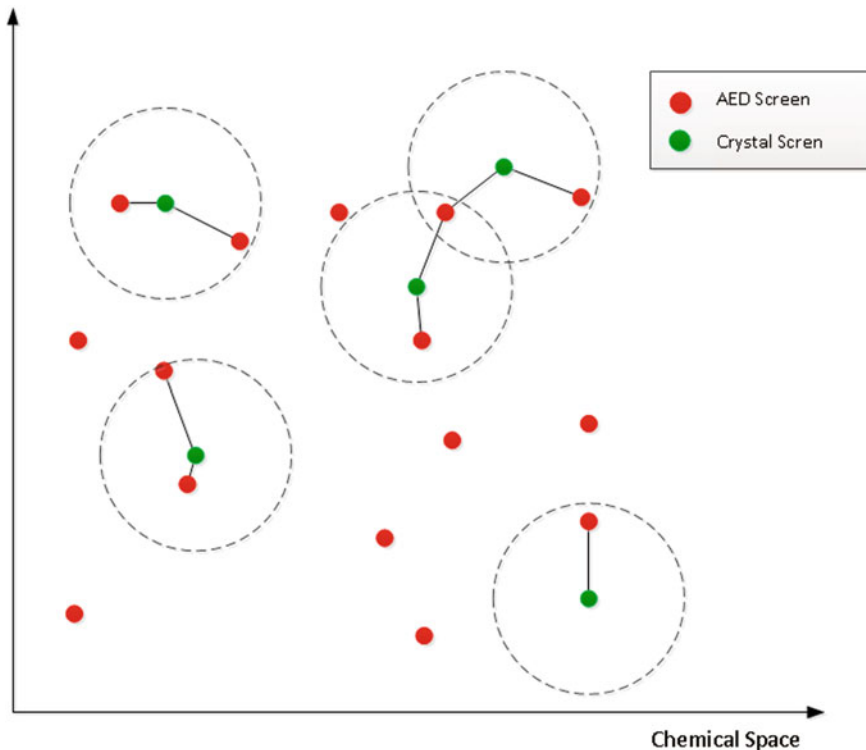


Fig. 3.4 Selecting the candidate cocktails ©2016 IEEE

For analyzing crystallization likelihood, the distance from *AED* cocktails (red points) to all crystal cocktails (green points) is calculated firstly. At this point, no score from the input list is eliminated, because even if the *AED* generates candidate cocktails using crystal conditions having score 4 to 7, it is still able to generate some cocktails that are close to 3D crystals in the chemical space. To calculate the distance between two cocktails, cocktail distance coefficient (CD_{coeff}) [9] given in Eq. 3.10 is used:

$$CD_{coeff} = \frac{1}{\text{sum}(\omega)} \left(\left(\frac{|E(pH_i) - E(pH_j)|}{14} \right) \omega_1 + BC(F_i, F_j) \omega_2 \right) \quad (3.10)$$

where $\omega = \omega_1, \omega_2, \omega_i \geq 0$, and $\text{sum}(\omega) > 0$. $E(pH_i)$ is the estimation of the pH in the cocktail, and $BC(F_i, F_j)$ is the Bray–Curtis dissimilarity measure[7] of fingerprints of the chemicals as in shown Eqs. 3.11 and 3.12:

$$BC(F_i, F_j) = \sum_k |F_{ik} - F_{jk}| / \sum_k |F_{ik} + F_{jk}| \quad (3.11)$$

and

$$F_k = \sum_{i=1}^n f_{ik}[c_i] \quad (3.12)$$

where f_{ik} is the frequency count of descriptor k from the extended-connectivity fingerprints of component i , and c_i is the molar concentrations of the i th component of the chemical. The detailed information about the calculation of the CD_{coeff} is provided in [9] and at the website.³

Once the distances are calculated, for each *AED* cocktail, the minimum distance to each crystal class and also to all crystal classes in the preliminary data are taken. In this way, a matrix of distances to each crystal in the preliminary data is obtained. By using the minimum distance to any crystal, the lists are sorted in ascending order. This analysis is performed on the prioritized candidate cocktails.

3.6.4 Optimizing Concentration Values

The goal of the optimization screen here is to test the leading combinations over several concentrations. Thus, for precipitant X in buffer Y , with additive Z , the concentration of buffer Y and additive Z are kept constant (typically at 0.1 and 0.2 M, respectively.) while the concentration of precipitant X is varied. Concentrations of X are varied over three solutions, starting at the highest concentration indicated from either the *AED* analysis or by reference to the original screen compositions, and reducing by typically 20–25% for each of the next two solutions. Thus, a 96 condition screen results in 32 unique combinations of X , Y , and Z at three different concentrations of X .

A rapid reduction in the *AED* analysis listing can be carried out using the methods given. Output conditions are listed in order of their calculated priority scores, highest to lowest. Those with the highest priority scores are the mixtures containing the components judged most likely to result in crystals, while those with the lowest are the least likely. The final screen conditions are arrived at by going through the *AED* analysis and working down the priority listing. The *AED* analysis on its own gives new and unique combinations not present in the original screens, while the prioritization process gives the reagents associated with the highest scores. Optimization screens based solely on prioritization lead to a “cookie cutter” approach to optimization screen generation, where the same mixtures of precipitants are used with different buffers. Thus the use of both approaches together is necessary for the most comprehensive optimization screen. Regardless, the initial screen conditions are constantly referred to when generating the *AED* optimization screen, primarily as a guide to reagent concentrations.

³<https://github.com/ubccr/cockatoo/>.

Table 3.1 Parameters of the proteins ©2016 IEEE

Protein	pI	MW	% α -Helix	% SS-Sheet	% Coil
Tt82	4.85	27,900	34	5.8	24.5
Tt106	5.71	22,500	31.9	7.7	18.8
Tt189	5.8	19,600	24.1	6.5	25.9

Three commercial screens were chosen to have a diverse array of precipitants with some overlap as defined by the C6 webtool [25]. The measured diversities are: *HRHT* to *JCSG+* = 0.527, *HRHT* to *MCSG-3* = 0.489, *JCSG+* to *MCSG-3* = 0.367. Some repetition of conditions is present, and these are used as internal controls for scoring and reproducibility. The fourth, Screen 4a, was devised by examination of the components of the three commercial screens. A number of components are only present once or twice, and Screen4a was devised to increase the overall occurrence of these low-frequency components so that conclusions about their efficacy are not based upon a single result.

3.7 Experiments and Evaluation

This section briefly explains experiments done for evaluation of the *AED* method. Proteins were originally subjected to crystallization screening using a single 96 condition screen as previously reported [29]. Subsequent efforts have used four 96 conditions screens; Hampton Research High Throughput (*HRHT*, cat. #*HR2-130* [1]), Molecular Dynamics *JCSG+* screen (cat. #*MD1-40* [3]), Microlytics *MCSG-3* Screen (cat. #*MCSG-3* [2]), and a 96 condition screen under development in-house identified as Screen4a. All proteins were trace fluorescently labeled with the dye 5- (and 6)-carboxyrhodamine 6G (Molecular Probes cat.#*C-6157*) prior to screening [17, 29]. Crystallization screening plates were set up using 96 well plates having 3 drop positions per well (Corning CrystalEX, cat. #3553), with the protein: precipitant ratio's (v/v) for the drops being 1:1, 2:2, and 4:1. Plates were imaged using the in-house developed Crystal X2 imager [32] (iXpressGenes/Molecular Dimensions), with the first set of images immediately after set up, on days 1, 2, 4, and thence on a weekly basis for the next six weeks. Plates were scored by visual observation, with the scores then adjusted by reference to the fluorescent images [29]. Thus the primary function of the fluorescent images was to remove non-protein objects from the data, the discovery of crystals that were missed by visual examination, and the assignment scores of 4.

3.7.1 *Proteins for Preliminary Experiments*

The proteins were chosen to have a range of scoring outcomes based upon a single crystallization screen. The three proteins employed in collecting preliminary data are: *Tt189*, annotated as a nucleoside diphosphate kinase; *Tt82*, annotated as a HAD superfamily hydrolase, and *Tt106*, annotated as a nucleoside kinase. These proteins were chosen as being facile, moderately difficult, and difficult crystallizers, respectively. Secondary structure predictions were made using NetSurfP [27]. Protein molecular weights and pI's were calculated using the ExPASy server [18]. A cutoff prediction of 0.8 was used to estimate the percent of secondary structural features for each protein. The protein parameters are given in Table 3.1. In the case of *Tt106*, no crystals were obtained in the initial screening experiments, which involved six replicate plates [29].

3.7.2 *Results for Preliminary Data*

Optimization screens were devised based upon the *AED* analysis of the scored screening results, the 96 condition *AED* screens were then prepared and set up. For these preliminary data sets, the *AED* optimization screen conditions covered a broader range, with both precipitants 1 and 2 being varied over a range of conditions. Each grouping represents a family of screen conditions around a common theme, consisting of the same buffer and precipitants 1 and 2. Results analysis, as shown in Table 3.3, count the “families” where crystals were found, not the individual conditions. The results for *Tt189* are shown in Fig. 3.5, with each family of conditions outlined in red. For all three proteins the *AED* derived conditions were judged to be novel relative to the starting screen. When compared to all commercially available screens 7 of the 8 conditions were found to be novel, i.e., not occurring elsewhere. For the protein *Tt106*, the *AED* optimization screen only resulted in crystals after a second optimization round using additives with the *AED*-derived conditions.

Success and Novelty of AED Screens. The crystallization screen components that were determined to have the greatest positive effect were determined by the *AED* software, and a 96 condition optimization screen generated using those components for each protein. Optimization was in 96 well sitting drop plates, with the protein being *TFL*'d to facilitate results analysis. The successful conditions were identified and scored. Those conditions giving 2D and 3D crystals were then used to search the C6 database [25] for similar conditions across all commercially available screens as a determination of their uniqueness. Some sample images are provided in Fig. 3.6. As the optimization screens had different concentration ratios for the same precipitant pairs, each ratio where a hit was obtained was searched and the lowest C6 score was used.

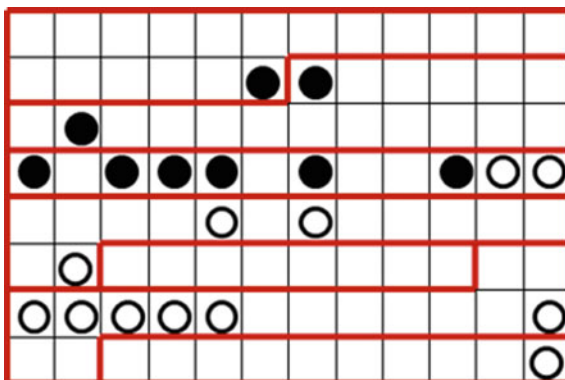
Table 3.2 shows the score distribution of preliminary data versus *AED* results. According to the table, *AED* generated more crystals than the preliminary data.

Table 3.2 Data distribution ©2016 IEEE

Score	Tt189		Tt82		Tt106	
	AED %	HSHT %	AED %	HSHT %	AED %	HSHT %
0	0.00	0.00	10.42	31.25	0.00	18.75
1	3.13	68.75	65.63	47.92	30.21	44.79
2	40.63	0.00	13.54	6.25	32.29	21.88
3	6.25	8.33	5.21	0.00	4.17	0.00
4	3.13	12.50	0.00	4.17	0.00	10.42
5	23.96	5.21	2.08	6.25	12.50	3.13
6	1.04	0.00	2.08	0.00	0.00	1.04
7	12.50	0.00	0.00	4.17	10.42	0.00
8	9.38	5.21	1.04	0.00	10.42	0.00

Although *AED* results generated more crystals, not all cocktails are novel compared to all commercial cocktails. Table 3.3 shows the number of novel conditions generated by *AED*. The numerical values in the first two columns after the protein name refer to the number of conditions with that score in the original screening experiment (numerator) versus those with that score in the optimization screen (denominator). The third column lists the number of optimization conditions that are novel compared to the original screen, while the last column lists those that are novel compared to all available screens. All found conditions were judged to be novel compared to the original screen on the basis of the cutoff score criteria. For *Tt189*, one optimization condition was identical to an existing commercial screen condition, but had no identity with any of the original input screen conditions.

Fig. 3.5 Results for the preliminary data *AED* screen of protein Tt189. The filled black circles represent conditions where 3D crystals were obtained, while the open circles are those where 2D plate crystals were obtained. Each family of conditions is outlined in red
©2016 IEEE



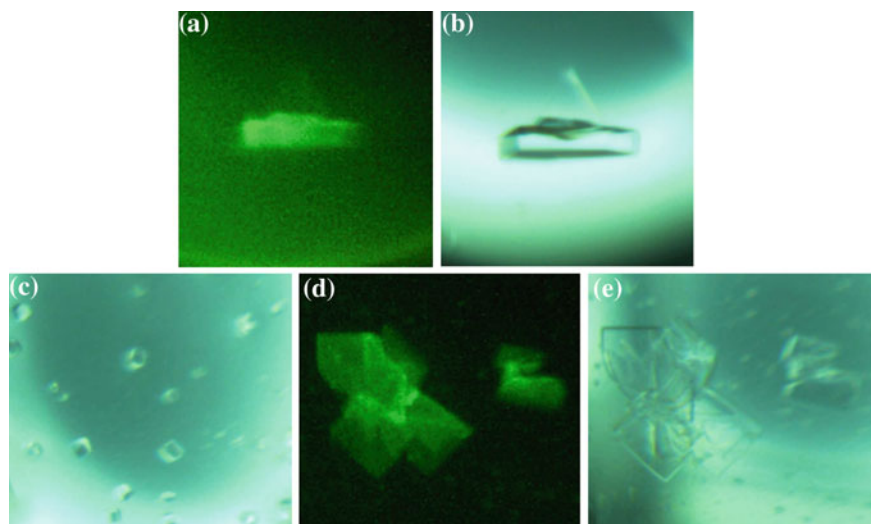


Fig. 3.6 Sample protein images **a, b** Tt82, **c–e** Tt189 © 2016 IEEE

Table 3.3 Summary of Experiments ©2016 IEEE

Protein annotated function	HSHT screen ^a	Optimize screen	Novel Cond. versus	Novel Cond. versus
	Score = 7	Score = 8, 9	HSHT Screen ^b	All Screens ^b
Tt189 (Nucleoside diphosphate kinase)	0 / 2	5 / 3	5	4
Tt82 (HAD superfamily hydrolase)	1 / 1	0 / 1	2	2
Tt106 (Nucleoside kinase)	0 / 0	0 / 1	1	1

^aHSHT: Hampton Screen High-Throughput

^bUsing C6 tool for scores of 7, 8, and 9 threshold value of 0.3

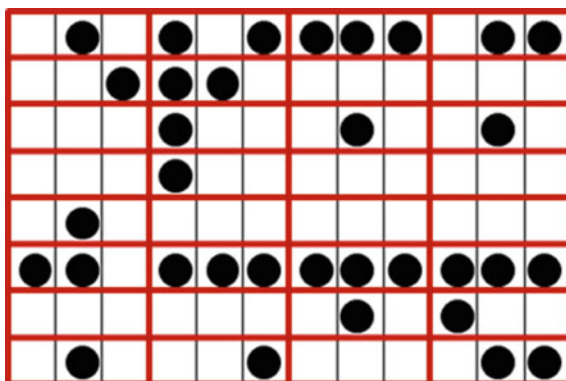
3.7.3 Expanded Screen Analysis

The proteins employed are a protein from the archaeal exosome complex RrP42 plus the three described above from the hyperthermophilic archaeon *Thermococcus thio-reducens* [28], an inorganic pyrophosphatase from *Staphylococcus aureus*, and human holo-transferrin (hTFN, Sigma, cat.# T-4132).

The proteins were subjected to the expanded screen tests and the results obtained are given in Table 3.4. In this case, only outcomes giving faceted 3D crystals are used for an endpoint. For these proteins, the AED optimization screen conditions were in groups of three, and each condition giving a crystal was counted.

Protein Tt189, from the preliminary results, was repeated. The results (Fig. 3.7) indicate that of the 32 families of conditions optimized 20 of them resulted in 3D crystals (63%) compared to the 3 out of 7 (43%) from the preliminary data. The

Fig. 3.7 AED optimization screen for protein Tt189, in this case, is generated using the combined results from four different 96 condition crystallization screens. The individual families of conditions are outlined in red. Only those conditions resulting in 3D crystals are shown ©2016 IEEE



results are shown in Fig. 3.7 also indicate where the more “robust” crystallization conditions are to be found, those where all three concentrations of precipitant 1 resulted in crystals.

For *Staphylococcus aureus* IPPase (*SalPP*), two crystals were obtained in the four screens or the AED optimized screen. However, the AED screen did result in a number of conditions that had a score of 5, non-faceted crystals. The analysis had indicated that low MW polyethylene glycols, divalent cations, and basic pH’s were the lead factors for obtaining crystals. The AED-derived screen results confirmed the high pH and low MW polyethylene glycols and further indicated that Ca⁺⁺, but not Mn⁺⁺ or Mg⁺⁺, was the best divalent cation. Every well containing Ca⁺⁺ resulted in spheroids or rough non-faceted crystals, while none of those containing Mn⁺⁺ or Mg⁺⁺ had any. While these are not suitable for diffraction analysis, they can be used as a source of seed crystals [11]. The optimization conditions were subsequently tested using crystallization by capillary counter diffusion [26], which resulted in the two hits obtained.

Table 3.4 Optimization results ©2016 IEEE

Protein	# of Crystals 4X screens/AED conditions (family's)
Holo Human transferrin	1/5 (4)
RrP42 (archaeal exosome protein)	4/15 (7)
Tt189 (nucleoside diphosphate kinase)	10/33 (20)
Tt106 (nucleoside kinase)	1/9 (6)
Tt82 (HAD superfamily hydrolase)	8/3 (2)
Stapylococcus aureus inorganic pyrophosphatase	0/2 (2)

For three of the four proteins, more crystallization conditions were determined by the *AED* screen than were found using the four “set” screens. In two of these cases, more families of conditions were determined.

3.7.4 Evaluation of Ranked Results

Evaluation of ranked results is needed to compare several methods based on whether they rank crystalline candidate conditions at the top of their list or not. Unfortunately, there are not even handful of successful and actively used prevalent computational methods that analyze previous experiments. However, such comparison of rankings is still needed when there are competitive virtual screening methods.

The traditional ranking methods are sensitive to irrelevant samples appearing before relevant samples. For protein crystallization, if all relevant cocktails are included in a well plate, it is not critical to have screens leading to noncrystals. The list of cocktails is partitioned into bins and then the number of relevant (crystalline) screens in each bin is analyzed. Ideally, the good candidate cocktails should appear in bins that correspond to the top of the ranked list. *Bin – Recall* [15] measures how “close” the cocktails that yield crystals to the top of the ranked list. It generates a normalized value, which is close to 1 (or 100%) when the ranking results are similar to the best case, and it is 0 when the results are far from the best case.

Bin – Recall is computed based on the formulation given in Eq. 3.13:

$$R_{bin} = \frac{\sum_{j=1}^{|B|} \delta_j (\sum_{i=S_{cmin}}^{S_{cmax}} \omega_i n_{i,j}) - \sum_{i=1}^n S_i \delta_{\lfloor \frac{n-i}{binSize} \rfloor} \omega(S_i)}{\sum_{i=1}^n S_i \delta_{\lfloor \frac{i}{binSize} \rfloor} \omega(S_i) - \sum_{i=1}^n S_i \delta_{\lfloor \frac{n-i}{binSize} \rfloor} \omega(S_i)} \quad (3.13)$$

where $|B|$ is the number of bins, δ_j is the weight of the bin j , ω_i is the weight of the score i , and $n_{i,j}$ is the number of score i in bin j . S is the list of ordered scores, where $cmin$ is the minimum crystal score and $cmax$ is the maximum crystal score. The denominator of the expression is used to normalize the measure dividing by the best scenario (i.e., all crystalline conditions appear in the top bin) minus the worst scenario (i.e., all crystalline conditions appear in the lower bins). The numerator computes the value based on the distribution of the scores to the bins and subtracts the worst case. *Bin – Recall* measure allows to give high weights (ω_i) to cocktails or samples having high scores. Similarly, bins can also be assigned weights (δ_j) based on where all crystalline conditions should appear. In this case, the top bin having low distances to crystals is given the highest weight. Ideally, the goal is to obtain *Bin – Recall* value of 100%. It depends on the expert to determine the number of bins for analysis.

The partitioning of scores into 3 bins with respect to distance to crystals is provided in Table 3.5. Protein Tt106 had high number of score 8 cocktails in Bin 1 when considering distance to score 4 and all crystals. Its bin-recall measure with respect to distance to score 4 (and also all crystals) is computed as follows:

$$R_{bin} = \frac{3 * \begin{bmatrix} 8 \\ 7 \\ 6 \\ 5 \end{bmatrix}^T * \begin{bmatrix} 6 \\ 3 \\ 0 \\ 4 \end{bmatrix} + 2 * \begin{bmatrix} 8 \\ 7 \\ 6 \\ 5 \end{bmatrix}^T * \begin{bmatrix} 2 \\ 6 \\ 0 \\ 6 \end{bmatrix} + 1 * \begin{bmatrix} 8 \\ 7 \\ 6 \\ 5 \end{bmatrix}^T * \begin{bmatrix} 2 \\ 1 \\ 0 \\ 2 \end{bmatrix} - 1 * \begin{bmatrix} 8 \\ 7 \\ 6 \\ 5 \end{bmatrix}^T * \begin{bmatrix} 10 \\ 10 \\ 0 \\ 12 \end{bmatrix}}{3 * \begin{bmatrix} 8 \\ 7 \\ 6 \\ 5 \end{bmatrix}^T * \begin{bmatrix} 10 \\ 10 \\ 0 \\ 12 \end{bmatrix} - 1 * \begin{bmatrix} 8 \\ 7 \\ 6 \\ 5 \end{bmatrix}^T * \begin{bmatrix} 10 \\ 10 \\ 0 \\ 12 \end{bmatrix}} = 63.33\% \quad (3.14)$$

In this equation, the weights of bins are assigned as $\delta_1 = 3$, $\delta_2 = 2$, and $\delta_3 = 1$. There are some score 8 cocktails in Bin 2 and Bin 3. Moreover, 6 of score 7 cocktails appear in Bin 2. The presence of high scoring cocktails in lower bins reduced its score.

3.8 Summary

In this chapter, we have provided an overview of protein crystallization screening methods that analyze outputs of previously conducted experiments in the wet labs. Three successful screening methods for analysis was based on neural networks, genetic algorithms, and the new associative experimental design. The computational methods may generate numerous cocktails and the output screens should be optimized for wet lab experiments. The optimizations include the elimination of prohibited conditions, prioritizing remaining conditions, optimizing concentration values and ranking prioritized cocktails. Although obtaining crystalline conditions in the new screens is important, a ranking of cocktails based on their likelihood of crystalline conditions is also critical if a number of screens are designed for wet lab experiments. More computational methods are needed to help crystallographers design successful experiments. Among these methods, the *AED* is actively used and generated a good number of novel conditions that did not exist in commercial screens for a variety of proteins. There are reasons beyond simply obtaining a crystal for using a method such as the *AED* analysis:

- **Finding more robust conditions.** Crystal nucleation is a stochastic process, and it is not uncommon to set up the same condition multiple times with varying outcomes [25, 29]. The *AED* analysis approach not only helps to find new crystallization conditions but also, as implemented herein, finds more “robust” crystallization conditions, i.e., those that are less sensitive to the concentration of one or more of the components present. This is shown in Fig. 3.7, where for each family, there are three different concentrations of precipitant #1. Those conditions that are more sensitive are identified by only one outcome having 3D crystals in a family, and

Table 3.5 3-Bin partition of the proteins based on different ranking schemes ©2016 IEEE

		Score	Bin 1	Bin 2	Bin 3
Protein Tt189	Min distance to score 4	5	9	6	8
		6	1	0	0
		7	4	5	3
		8	2	2	5
	Min distance to score 5	5	6	10	7
		6	0	1	0
		7	2	4	6
		8	6	2	1
	Min distance to score 8	5	7	10	6
		6	1	0	0
		7	0	4	8
		8	2	5	2
	Min distance to all crystals	5	10	7	6
		6	1	0	0
		7	2	4	6
		8	4	3	2
Protein Tt82	Min distance to score 4	5	0	2	0
		6	1	1	0
		7	0	0	0
		8	1	0	0
	Min distance to score 5	5	1	1	0
		6	0	2	0
		7	0	0	0
		8	0	0	1
	Min distance to score 7	5	0	2	0
		6	0	1	1
		7	0	0	0
		8	0	0	1
	Min distance to all crystals	5	0	2	0
		6	1	1	0
		7	0	0	0
		8	1	0	0
Protein Tt106	Min distance to score 4	5	4	6	2
		6	0	0	0
		7	3	6	1
		8	6	2	2
	Min distance to score 5	5	5	4	3
		6	0	0	0
		7	4	4	2
		8	5	4	1
	Min distance to all crystals	5	4	6	2
		6	0	0	0
		7	3	6	1
		8	6	2	2

those that are less sensitive have crystals in all three concentrations of precipitant #1.

- **Improving existing conditions.** The existing found crystallization conditions may not be readily repeatable, or may not give crystals diffracting to a sufficient resolution. *AED* analysis can reveal an expanded range of conditions, some or many of which may resolve these problems.
- **Possibly new space groups (to facilitate binding analysis).** Binding studies where potential ligands are soaked into a crystal to determine their location upon diffraction analysis require that the binding sites be available, not occluded by crystallographic contents. Space groups obtained in initial screening experiments may not be suitable for these studies, prompting a search for new packing arrangements.
- **Improved diffraction resolution.** Having good looking crystals does not automatically translate to good diffraction resolution. However, having crystals where previously one had none, such as with the protein Tt106, does markedly improve one's chances of obtaining a structure. Thus, a primary reason for the *AED* analysis is to find crystallization conditions where there previously were none. Additionally, crystal nucleation is a stochastic process. From Figs. 3.5 and 3.7, we see that there are families having many crystallization conditions, and families only having 1 or none. It is intuitively apparent that those with many conditions are more robust, less sensitive to component concentrations and more likely to result in crystals, than those with few conditions. This is important when carrying out additional screening trials and optimizations for improved diffraction resolution and for studies such as for substrate binding or drug development.
- **Improved crystal size (for neutron diffraction).** Although not shown in the data presented, the *AED* optimization results yielded a range of crystal sizes. Neutron diffraction requires crystals $\leq 1 \text{ mm}^3$ in size. Conditions that favor larger crystals can be determined from these results and are likely a more favorable starting point for growth of large volume crystals.

As shown by comparing Figs. 3.5 and 3.7, using more screens in the initial search gives a larger search space for the *AED* analysis. Commercially available screens have a finite number of precipitants present. Increasing the number of screens results in exposure to an expanded range of conditions, although some are only present in 1 or 2 of the conditions. For this reason, Screen 4a was formulated to increase the occurrence of these occasional precipitants to complement the other three screens.

Not all proteins yielded crystals upon *AED* optimization screening. In the case of Tt106, the crystals were obtained from the *AED*-identified conditions after additional optimization using crystallization additives. In the case of SaIPP, the *AED* analysis indicates those conditions, which should be most likely to result in crystals, and as such is the starting point for subsequent screening experiments. *AED* analysis results in screen conditions, thus screens, that are formulations of the components most likely to yield crystals of that protein.

Acknowledgements The the first and second paragraphs (except the first sentences) of Sect. 3.3 are Reprinted from Progress in Biophysics and Molecular Biology, Volume 88, Issue 3, Lawrence

J. DeLucas, David Hamrick, Larry Cosenza, Lisa Nagy, Debbie McCombs, Terry Bray, Arnon Chait, Brad Stoops, Alexander Belgovskiy, W. William Wilson, Marc Parham, Nikolai Chernov, Protein crystallization: virtual screening and optimization, Pages 285–309, Copyright (2005) with permission from Elsevier.

The second paragraph (except the first two sentences) and the third paragraph of Sect. 3.4 are Reprinted (adapted) with permission from *Crystal Growth and Design* 2011 11 (7), Emmanuel Saridakis, Novel Genetic Algorithm-Inspired Concept for Macromolecular Crystal Optimization, 2993–2998. Copyright (2011) American Chemical Society. ©2016 IEEE. Reprinted, with permission, from I. Dinc, M. L. Pusey, and R. S. Aygün, “Optimizing Associative Experimental Design for Protein Crystallization Screening,” in *IEEE Transactions on NanoBioscience*, vol. 15, no. 2, pp. 101–112, March 2016. doi: <https://doi.org/10.1109/TNB.2016.2536030>.

References

1. Hampton Research Screen HT. https://hamptonresearch.com/documents/product/hr000783_crystal_screen_2.xls. Accessed 1 November 2015.
2. Microlytics MCSG-3 Screen. http://www.microlytic.com/sites/default/files/MCSG3_Formulations_0_0_0.pdf. Accessed 1 November 2015.
3. Molecular Dynamics JCGS+ Screen. <http://www.moleculardimensions.com/applications/upload/Md1-40%20JCGS%20Plus%20HT-96.pdf>. Accessed 1 November 2015.
4. Abergel, C., Moulard, M., Moreau, H., Loret, E., Cambillau, C., & Fontecilla-Camps, J. C. (1991). Systematic use of the incomplete factorial approach in the design of protein crystallization experiments. *Journal of Biological Chemistry*, 266(30), 20131–20138.
5. Asenjo, J. A., & Andrews, B. A. (2011). Aqueous two-phase systems for protein separation: a perspective. *Journal of Chromatography A*, 1218(49), 8826–8835.
6. Asenjo, J. A., & Andrews, B. A. (2012). Aqueous two-phase systems for protein separation: phase separation and applications. *Journal of Chromatography A*, 1238, 1–10.
7. Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, 27(4), 325–349.
8. Brodersen, D. E., Andersen, G. R., & Andersen, C. B. F. (2013). Mimer: an automated spreadsheet-based crystallization screening system. *Acta Crystallographica Section F*, 69(7), 815–820.
9. Bruno, A.E., Ruby, A.M., Luft, J.R., Grant, T.D., Seetharaman, J., Montelione, G.T., Hunt, J.F., and Snell, E.H. Comparing chemistry to outcome: the development of a chemical distance metric, coupled with clustering and hierarchal visualization applied to macromolecular crystallography.
10. Carter, C. W, Jr., & Carter, C. W. (1979). Protein crystallization using incomplete factorial experiments. *The Journal of Biological Chemistry*, 254(23), 12219–12223.
11. D’Arcy, A., Bergfors, T., Cowan-Jacob, S. W., & Marsh, M. (2014). Microseed matrix screening for optimization in protein crystallization: what have we learned? *Acta Crystallographica Section F: Structural Biology Communications*, 70(9), 1117–1126.
12. DeLucas, L. J., Hamrick, D., Cosenza, L., Nagy, L., McCombs, D., Bray, T., et al. (2005). Protein crystallization: virtual screening and optimization. *Progress in Biophysics and Molecular Biology*, 88(3), 285–309.
13. Dinc, I. (2016). *Associative Data Analytics and its Application to Protein Crystallization Analysis*. Ph.D dissertation, University of Alabama in Huntsville.
14. Dinc, İ., Pusey, M.L., and Aygün, R.S. (2015). Protein crystallization screening using associative experimental design. In *Bioinformatics Research and Applications* (pp. 84–95). Springer.
15. Dinc, İ., Pusey, M. L., & Aygün, R. S. (2016). Optimizing Associative Experimental Design for Protein Crystallization Screening. *IEEE Transactions on NanoBioscience*, 15(2), 101–112.

16. Doudna, J. A., Grosshans, C., Gooding, A., & Kundrot, C. E. (1993). Crystallization of ribozymes and small rna motifs by a sparse matrix approach. *Proceedings of the National Academy of Sciences*, 90(16), 7829–7833.
17. Forsythe, E., Achari, A., & Pusey, M. L. (2006). Trace fluorescent labeling for high-throughput crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 62(3), 339–346.
18. Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., and Bairoch, A. (2005). *Protein identification and analysis tools on the ExPASy server*. Springer.
19. Giegé, R. (2013). A historical perspective on protein crystallization from 1840 to the present day. *FEBS Journal*, 280(24), 6456–6497.
20. Jancarik, J., & Kim, S.-H. (1991). Sparse matrix sampling: a screening method for crystallization of proteins. *Journal of Applied Crystallography*, 24(4), 409–411.
21. Kwon, J. S.-I., Nayhouse, M., Christofides, P. D., & Orkoulas, G. (2013). Modeling and control of protein crystal shape and size in batch crystallization. *AIChE Journal*, 59(7), 2317–2327.
22. Luft, J. R., Newman, J., & Snell, E. H. (2014). Crystallization screening: the influence of history on current practice. *Structural Biology and Crystallization Communications*, 70(7), 835–853.
23. McPherson, A., & Cudney, B. (2014). Optimization of crystallization conditions for biological macromolecules. *Structural Biology and Crystallization Communications*, 70(11), 1445–1467.
24. McPherson, A., & Gavira, J. A. (2014). Introduction to protein crystallization. *Acta Crystallographica Section F: Structural Biology Communications*, 70(1), 2–20.
25. Newman, J., Fazio, V. J., Lawson, B., & Peat, T. S. (2010). The c6 web tool: a resource for the rational selection of crystallization conditions. *Crystal Growth and Design*, 10(6), 2785–2792.
26. Ng, J. D., Gavira, J. A., & García-Ruíz, J. M. (2003). Protein crystallization by capillary counterdiffusion for applied crystallographic structure determination. *Journal of structural biology*, 142(1), 218–231.
27. Petersen, B., Petersen, T. N., Andersen, P., Nielsen, M., & Lundegaard, C. (2009). A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC structural biology*, 9(1), 1.
28. Pikuta, E. V., Marsic, D., Itoh, T., Bej, A. K., Tang, J., Whitman, W. B., et al. (2007). *Thermococcus thioreducens* sp. nov., a novel hyperthermophilic, obligately sulfur-reducing archaeon from a deep-sea hydrothermal vent. *International Journal of Systematic and Evolutionary Microbiology*, 57(7), 1612–1618.
29. Pusey, M., Barcena, J., Morris, M., Singhal, A., Yuan, Q., & Ng, J. (2015). Trace fluorescent labeling for protein crystallization. *Structural Biology and Crystallization Communications*, 71, 7.
30. Raja, S., Murty, V. R., Thivaharan, V., Rajasekar, V., & Ramesh, V. (2011). Aqueous two phase systems for the recovery of biomolecules—a review. *Science and Technology*, 1(1), 7–16.
31. Saridakis, E. (2011). Novel Genetic Algorithm-Inspired Concept for Macromolecular Crystal Optimization. *Crystal Growth and Design*, 11(7), 2993–2998.
32. Sigdel, M., Pusey, M. L., & Aygun, R. S. (2013). Real-time protein crystallization image acquisition and classification system. *Crystal Growth and Design*, 13(7), 2728–2736.
33. Snell, E. H., Nagel, R. M., Wojtaszczyk, A., O’Neill, H., Wolfley, J. L., & Luft, J. R. (2008). The application and use of chemical space mapping to interpret crystallization screening results. *Acta Crystallographica Section D: Biological Crystallography*, 64(12), 1240–1249.
34. Stevens, R. C. (2000). High-throughput protein crystallization. *Current Opinion in Structural Biology*, 10(5), 558–563.

Chapter 4

Robotic Image Acquisition

Abstract Protein crystallization is a complex phenomenon requiring thousands of experiments corresponding to different crystallization conditions for successful crystallization. In recent years, high-throughput robotic setups have been developed to automate the protein crystallization experiments, and imaging techniques are used to monitor the crystallization progress. Having an automated system to classify the images according to the crystallization phases can be very useful to crystallographers. This chapter describes the design and implementation of a stand-alone, low-cost, and real-time system for protein crystallization image acquisition and classification with a goal to assist crystallographers in scoring crystallization trials.

4.1 Introduction

Protein crystallization is the core part of protein crystallography studies. Numerous factors such as protein purity, pH, temperature, protein concentration, the type of precipitant, and the crystallization methods play an important role in crystallization [16]. The correct combination of all these factors is essential for the formation of crystals. However, it is difficult to predict exact conditions for protein crystallization [7]. Therefore, thousands of crystallization trials are often required for successful crystallization. Several robotic systems have been developed to automate crystallization process. Berry et al. [2] previously provided a review of the developments in high-throughput robotic setups to automate the crystallization experiments.

Crystallization trials should be observed periodically to assess the evolving progress of crystal growth or crystallization pathway. Knowledge about the crystallization phase helps in making several decisions. For instance, unsuccessful crystallization trials can be discarded. X-ray diffraction can be applied to single optically clear crystal. Likewise, if a protein is in the pathway of crystallization, the conditions can be optimized to get a crystalline outcome [15]. Therefore, high-throughput robotic systems should not only distinguish between crystal and non-crystals but also identify the likely-lead conditions for optimization.

Crystallization process may need to be monitored in different time spans (e.g., daily, weekly or monthly) depending on the protein and crystallization conditions. Since a large number of crystallization trials is required and these trials should be periodically assessed, manual analysis takes significant time. A number of research and commercial systems have been developed to facilitate the imaging of crystallization setups. In the past, commercial systems have been very large systems requiring expensive setup. High-throughput systems can reduce the tedious work to be completed by experts by automating the overall process. There are three main factors essential for practical use of these systems: (1) the results should be delivered fast, (2) the analysis results should be reliable, and (3) the cost should be low. High-throughput systems for protein crystallization analysis have in general the following issues:

1. Existing automated systems are very expensive and not portable.
2. Crystal detection is a complex process and usually requires complex image processing algorithms to extract features related to shapes of objects in an image. This makes it difficult to process and classify images in real time.
3. While achieving a real-time and automated system, a good level of accuracy needs to be maintained.

The scoring methods, image processing & feature extraction methods, classification methods, running time of analysis, and the accuracy of the overall system play critical role in usability of these systems in addition to the hardware components.

Levels of scoring Because of the high-throughput crystallization approach, manual review becomes impractical. Therefore, automated image scoring systems have been developed to collect and classify the crystallization trial images. The fundamental aim is to discard the unsuccessful trials, identify the successful trials, and possibly identify the trials which could be optimized. A significant amount of previous work (for example, Zuk & Ward (1991) [26], Cumba et al. (2003) [5], Cumba et al. (2005) [3], (4) Zhu et al. (2004) [25], Berry et al. (2006) [2], Pan et al. (2006) [13], Po & Laine (2008) [14] has described the classification of crystallization trials into non-crystal or crystal categories. Yang et al. (2006) [24] described classification into three categories (clear, precipitate, and crystal). Bern et al. (2004) [1] classified the images into five categories (empty, clear, precipitate, microcrystal hit, and crystal). Likewise, Saitoh et al. (2006) [18] described classification into five categories (clear drop, creamy precipitate, granulated precipitate, amorphous state precipitate, and crystal). Spraggon et al. (2002) [22] described classification of the crystallization imagery into six categories (experimental mistake, clear drop, homogeneous precipitant, inhomogeneous precipitant, microcrystals, and crystals). Cumba et al. (2010)[4] have developed the most optimistic system which classifies the images into three categories or 10 categories. It should be noted that there is no standard for categorizing the images, and different research studies have proposed different categories in their own way.

Image Processing and Feature Extraction Most of the proposed algorithms start image processing by determining the region of interest (droplet boundary) to define the search region for crystals, a computationally expensive process. The general technique applied here is to first apply an edge detection algorithm such as Sobel edge detection or Canny edge detection which is followed by some curve fitting algorithms such as Hough transform (Berry et al. (2006) [2], Pan et al. (2006) [13], Spraggon et al. (2002) [22], Zhu et al. (2004) [25]). Bern et al. (2004) [1] determined the drop boundary by applying edge detection followed by dynamic programming curve tracking algorithm. Yang et al. (2006) [24] used a dynamic contour method on Canny edge image to locate the droplet boundary. Cumba et al. (2003) [5] applied a probabilistic graphical model with a two-layered grid topology to segment the drop boundary. Po & Laine (2008) [14] used multiple population genetic algorithms for region of interest detection. Saitoh et al. (2004) [19] and Saitoh et al. (2006) [18] simplified this process by defining a fixed 150[pixel] \times 150[pixel] portion inside a well as the region of interest for search of crystals.

For feature extraction, a variety of image processing techniques have been proposed. Zuk & Ward (1991) [26] used the Hough transform to identify straight edges of crystals. Bern et al. (2004) [1] extracted gradient and geometry-related features from the selected drop. Pan et al. (2006) [13] used intensity statistics, blob texture features, and results from Gabor wavelet decomposition to obtain the image features. Research studies of Cumba et al. (2003) [5], Saitoh et al. (2004) [19], (14) Spraggon et al. (2002) [22], and Zhu et al. (2004) [25] used a combination of geometric and texture features as the input to their classifier. Saitoh et al. (2006) [18] used global texture features as well as features from local parts in the image and features from differential images. Yang et al. (2006) [24] derived the features from gray-level co-occurrence matrix (GLCM), Hough transform and discrete Fourier transform (DFT). Liu et al. (2008) [8] extracted features from Gabor filters Gabor wavelet, filter, integral histograms, and gradient images to obtain 466-dimensional feature vector. Po & Laine (2008) [14] applied multiscale Laplacian pyramid filters and histogram analysis techniques for feature extraction. Cumba et al. (2010) [4] presented the most sophisticated feature extraction techniques for the classification of crystallization trial images. Features such as basic statistics, energy, Euler numbers, Radon-Laplacian features, Sobel edge features, microcrystal features, and GLCM features are extracted to obtain a 14,908-dimension feature vector. Although increasing the number of features may help improve accuracy, it may slow down the classification process. In addition, the use of irrelevant features may deteriorate the performance of some classifiers.

Running Time of Experiments Because of the high-throughput rate of image collection, the speed of processing an image becomes an important factor. One of the most time-consuming steps is the determination of a region of interest or the drop boundary. Likewise, extraction of a large number of geometric and texture features increases the time and image processing complexities. The system by Pan et al. (2006) [13] required 30s per image for feature extraction. Po & Laine mentioned that it takes 12.5s per image for the feature extraction in their system [14]. Because

of high computational requirement, they are considering implementation of their approach on the Google computing grid. Feature extraction described by Cumba et al. (2010) [4] is the most sophisticated, which could take 5 h per image on a normal system. To speed up the process, they execute the feature extraction using a web-based distributed computing system. Overall, the image processing and feature extraction have been computationally expensive making it infeasible for real-time processing.

Classification To obtain the decision model for classification, a variety of classification techniques have been used. Zhu et al. (2004) [25] and Liu et al. (2008) [8] applied a decision tree with boosting. Bern et al. (2004) [1] used a decision tree classifier with hand-crafted thresholds. Pan et al. (2006) [13] applied an SVM learning algorithm. Saitoh et al. (2006) [18] applied a combination of decision tree and SVM classifiers. Spraggon et al. (2002) [22] applied self-organizing neural networks. Po et al. (2008) [14] combined genetic algorithms and neural networks to obtain a decision model. Berry et al. (2006) [2] determined scores for each object within a drop using learning vector quantization, self-organizing maps, and Bayesian algorithms. The overall score for the drop is calculated by aggregating the classification scores of the individual objects. Cumba et al. (2003) [5] and Saitoh et al. (2004) [19] applied linear discriminant analysis. Yang et al. (2006) [24] applied hand-tuned rule-based classification followed by linear discriminant analysis. Cumba et al. (2005) [3] used association rule mining, while Cumba et al. (2010) [4] used multiple random forest classifiers generated via bagging and feature subsampling.

Accuracy of the System With regard to correctness of classification, the best reported accuracy for the binary classification, i.e., classification into two categories, is 96.56% (83.6% true positive rate and 99.4% true negative rate, while classifying 8% of crystals into non-crystal categories) using deep convolutional neural network (CNN). Po et al. (2008) [14] achieved 93.5% average true performance (88% true positive and 99% true negative rates). Saitoh et al. have achieved accuracy in the range of 80–98% for different image categories [19]. Likewise, the automated system by Cumba et al. (2010) [4] detects 80% of crystal-bearing images, 89% of precipitate images, and 98% of clear drops accurately. The performance of the various systems, however, cannot be compared directly as they have used different datasets, different class categories, and number of categories. The current systems are not fully reliable, and there is still much room for improvement in terms of performance.

This chapter introduces how to build a low-cost, portable, real-time, and comparatively accurate robotic microscopy system for analysis of crystallization trial images. A low-cost architecture should not require expensive hardware parts. The keyword *real-time* has a semantics beyond *fast* computation. Real-time processes have deadlines. In this domain, the deadline is set as the time to move the microscope from one well to another well. The processing and analysis should be completed within this period. While achieving real-time analysis, the accuracy cannot be sacrificed. Later, this chapter also covers how classifiers could be built to minimize missing detection of crystals. The feature set used in this chapter has 45 features obtained from binarized images using three thresholding techniques. The use of small number of

features helps achieve real-time analysis without sacrificing accuracy. The detailed analysis of features is provided in Chap. 5.

4.2 Components of a Robotic Setup

Several robotic systems have been developed to automate crystallization process. Berry et al. (2006) [2] provide a review of the developments in high-throughput robotic setups to automate the crystallization experiments. Many robotic setups collect the images under white light. Processing and analysis of white-light images might be challenging since crystal regions are not easily separable from the solution. On the other hand, fluorescence microscopy may simplify image analysis significantly by adding a few stages to screen preparation. This chapter discusses the use of fluorescence microscopy for evaluating protein crystallization trials.

4.2.1 Well Plates

Crystallization trials setups use well plates that allow experimenting different combinations of crystallization conditions. A typical well plate consists of wells arranged in rows and columns. The experimental protein solution is placed in the wells. Figure 4.1 shows a typical well plate.¹ Some plates have several protein drop positions per precipitant reservoir. The structure of a well plate determines the scanning process for robotic setup.

4.2.2 Fluorescence Microscopy

It has been shown that trace fluorescent labeling can be a powerful tool for visually finding protein crystals [7, 15]. The design of a low-cost assembled fluorescence microscopy system that utilizes trace fluorescent labeling of protein solution that results in higher image intensity for the solution containing crystals, thereby simplifying the feature extraction process, has been described in [21]. The commercial model of the assembled microscopy system *Crystal X2* from iXpressGenes, Inc. has been used to collect the images shown in this treatise.

Studies on trace fluorescent labeled proteins have shown image intensity to be proportional to the structure or packing density of the proteins solid state [7, 15]. The fluorescence approach considerably simplifies finding crystals in a droplet, reducing the problem to one of finding the high-intensity regions, as opposed to finding the straight lines or particular shapes of objects that are often of low contrast.

¹https://commons.wikimedia.org/wiki/Main_Page

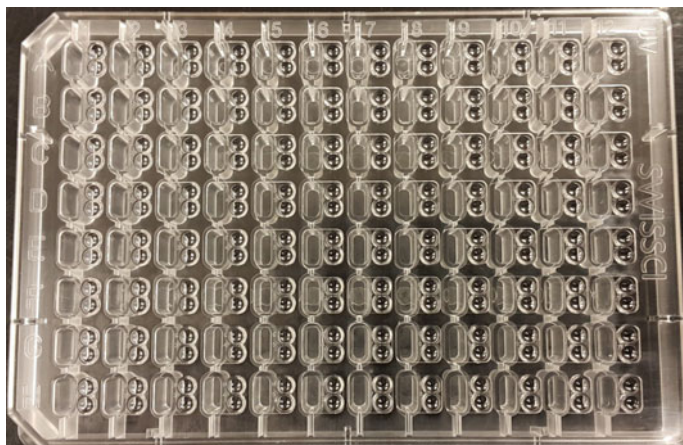


Fig. 4.1 Sample well plate

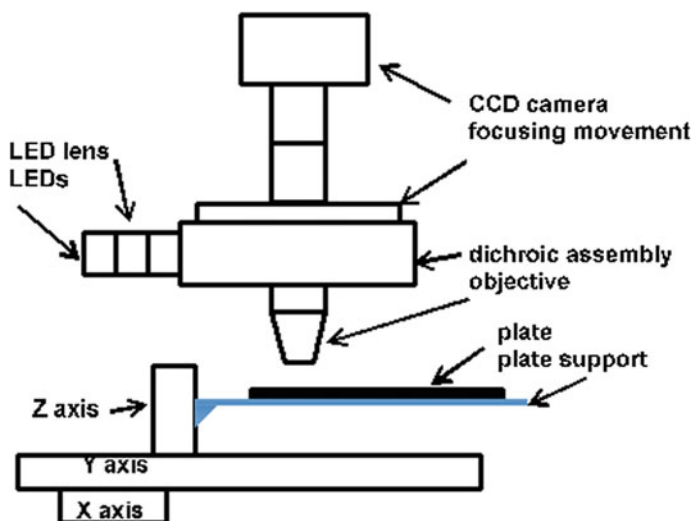


Fig. 4.2 Microscope structure. Reprinted (adapted) with permission from *Crystal Growth & Design* 2013 13 (7), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 2728–2736. Copyright (2013) American Chemical Society

Morphological analysis can be carried on sufficiently intense regions to determine if they can be formally classified as a crystal (presence of straight lines) or as a “bright spot” lead condition. This makes the feature extraction phase simple and faster than traditional pure image processing systems using white-light images for protein crystal detection.

The layout of an in-house assembled fluorescence microscopy system is shown in Fig. 4.2. Excitation light is supplied by an ultrabright light-emitting diode (LED) and emission filters. The light is focused and imaged by a 35 mm imaging lens. Image acquisition is done by a color camera connected to the computer through an ethernet cable. Stepper motors are interfaced through serial port of the connected PC. The motors control the position of the stages in the X- and Y-axes and of the camera (focus) in the Z-axis. The X, Y, and Z movements of the stepper motor-driven stages allow the camera to be positioned the exact drop positions for each precipitant condition. The crystallization plate is placed manually in the plate holder. The system supports standard plates and additional plate designs can be readily accommodated.

In Crystal X2, the standard “scope” is equipped with two high-intensity light-emitting diodes (LEDs), having peak emission of 530 and 590 nm. The light source and filter set are optimized for use with carboxyrhodamine (CR, Molecular Probes/Invitrogen cat. # C-6157) and Texas Red (TR, Molecular Probes/Invitrogen cat. # T-10244). Other fluorescent probes can be used if the light source and filters are changed accordingly. Other available high-intensity LED-based excitation wavelengths are 365, 385, 400, 420, 455, 470, 505, 530, 590, 617, 625, and 656 nm. Wavelengths below 400 nm will require different optical elements and objectives. Emission wavelengths above 650 nm will require a camera without an IR filter.

When selecting alternative probes, care should be taken that they are not sensitive to, e.g., pH, specific ions, etc. Many crystallization screens cover a broad range of chemical conditions, and many of the available fluorescent probes are sensitive to components or conditions present in the screening cocktails. However, while this has not (yet) been rigorously shown, being buried within a crystal may also serve to shield the probes from conditions that may negatively affect their fluorescence.

CRSE (succinimidyl ester) is available from Molecular Probes/Invitrogen in 5 mg bottles. Other sized containers may be available from other vendors. It is useful to directly add 1 mL of anhydrous solvent to the contents of the bottle and use the resulting solution. Between uses, the bottle can be stored at -20°C . However, experimenter should be careful to prewarm the bottle prior to opening and to avoid introduction of water to the bottle.

PCR tubes containing pre-aliquoted reactive dye can be set up to avoid having to subject the stock dye solution to repeated warming–cooling cycles, and the aliquots should be stored in a freezer immediately after preparation. Before opening, the tubes should be warmed to room temperature. The reactive dye is moisture sensitive, and repeated warming–cooling cycles increase the likelihood that moisture will get into the tubes, condense, and hydrolyze the dye. When removing a PCR tube of dye for use, take care to not keep the remaining tubes out to where they warm up. The advantage is that they are pre-measured, and one does not have to worry about the long-term stability of a larger bottle over repeated trips to and from the freezer.

To derivatize the protein, the scheme described in Pusey et al. (2015) [17] is used. This prepares the canonical 1 mL of protein at concentrations around 10 mg/mL which is routinely prepared in our laboratory. If smaller quantities of protein are to be derivatized, reduce the volumes that are buffer exchanged accordingly. Also, to avoid over labeling the protein, dilute the aliquot of dye proportionately. Note that

according to the manufacturer's instructions, when desalting or buffer exchanging solution volumes < 70 μL a chaser solution should be added on top to ensure that all the protein is properly eluted (<http://www.piercenet.com/instructions/2161729.pdf>). This may result in a final solution volume greater than desired, and thus an added concentration step. A smaller desalting column is available from the manufacturer, able to handle a maximum volume of 12 μL , about one-tenth the capacity of the 0.5 mL volume columns. The proteins used in these studies are trace fluorescently labeled. Carboxyrhodamine (Invitrogen) is favored as the covalent labeling probe of choice due to its high absorptivity, quantum yield, and relative lack of pH sensitivity.

4.3 Image Acquisition

The basic flow of the image acquisition is shown in Fig. 4.3. The protein crystallization screening plate is manually loaded into the assembled microscopy system. First, the probe (light) configuration is loaded, and the camera is initialized with proper settings. The plate configuration is then loaded to seek the coordinate of each well in the plate. At the start, the camera is positioned to the top-left corner of the well plate. For each well, the camera is positioned above the well, and the image is captured and saved in the repository. This process is repeated until all the wells are scanned. The commercial version of the microscopy system, Crystal X2, is developed by iXPress-Genes, Inc. It takes around 12 min to collect images from a 3-celled 96-well plate. The Crystal X2 system also comes with a classification framework for categorizing images automatically.

4.4 Image Processing and Segmentation

Expert-classified images are used to obtain a decision model for the classification of new images. It is important to focus on fast and effective image processing techniques so that the time for processing an image is less than the time between collecting two images. The steps of the image processing and feature extraction are explained below. Image feature extraction involves preprocessing steps such as color conversion, image thresholding, edge detection, region segmentation, etc. The main goal here is to extract useful features considering the feasibility for real-time analysis for the target system. Deeper analysis of using various features for analysis of crystallization trial images is provided in Chap. 5.

Consider an image I of size $H \times W$. Let $I(x, y)$ represent the pixel at location (x, y) where $1 \leq x \leq H$ and $1 \leq y \leq W$. In a color image, each pixel consists of red (R), green (G), and blue (B) components and can be described as a 3-tuple (R, G, B). The red, green, and blue intensity values of a pixel at $I(x, y)$ are represented as $I_R(x, y)$, $I_G(x, y)$, and $I_B(x, y)$, respectively.

Fig. 4.3 Image acquisition flow. Reprinted (adapted) with permission from Crystal Growth & Design 2013 13 (7), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 2728–2736. Copyright (2013) American Chemical Society

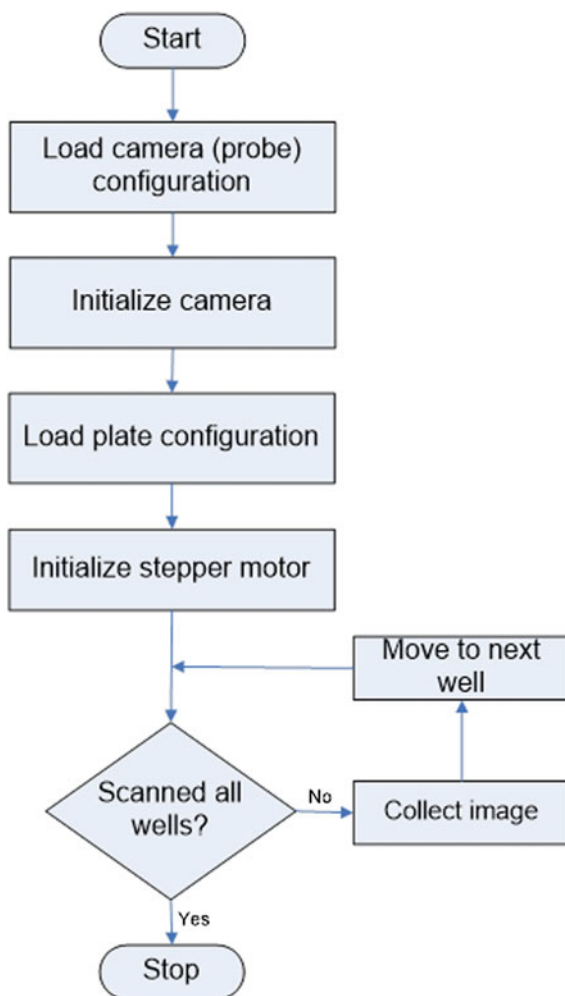


Figure 4.4 shows the components of Crystal X2. First, images are down-sampled and then median filter is applied for noise removal. Next, binary images are generated by three thresholding techniques. Then, image intensity features are extracted by combining the binary image and median-filtered image. Likewise, blobs are generated from the binary images and features are extracted related to the shape or size of the individual objects. Details on the feature extraction process are explained below.

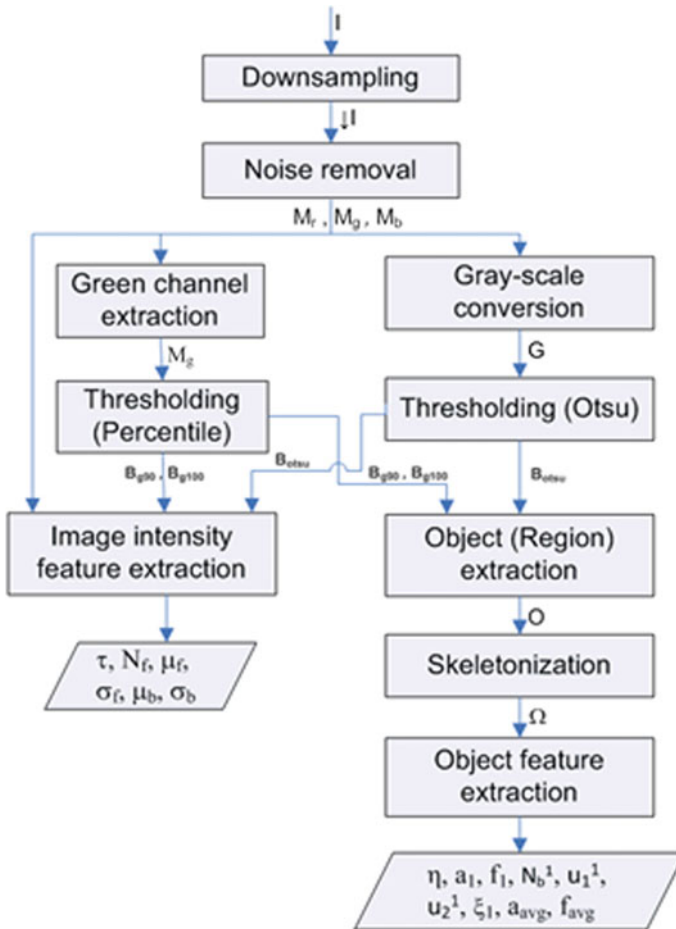


Fig. 4.4 Flowchart for image processing and feature extraction. Reprinted (adapted) with permission from Crystal Growth & Design 2013 13 (7), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 2728–2736. Copyright (2013) American Chemical Society

4.4.1 Image Preprocessing

Image down-sampling A high-resolution image may keep unnecessary details for image classification, especially, if the image has significant noise. In addition, processing a high-resolution image increases the computation time significantly. Therefore, the images are down-sampled before further processing. Suppose image $I(H \times W)$ is to be down-sampled by k times. Then, the resulting image $\downarrow I$ is of size $h \times w$ where $h = H/k$ and $w = W/k$. In these experiments, the original size of the images is 2560×1920 . By down-sampling it by eightfold, image size is reduced to

320 × 240. For this dataset, the down-sampled images contain sufficient detail for feature extraction.

Noise removal The down-sampled image $\downarrow I$ is passed through a median filter to remove random scattered noise pixels. Among different filters, median filter provided the best results for noise removal. To apply the median filter, a neighborhood window of size $(2p + 1) \times (2q + 1)$ around a point (x, y) is selected. Suppose ${}^x_p R_q^y$ represents a region in the original image centered around (x, y) with top-left coordinate $(x - p, y - q)$. F maps a 2D data into 1D set and median $(F({}^x_p R_q^y))$ provides the median value in the selected neighborhood around (x, y) . The red component in the resulting region (image) is denoted by $M_r(x, y)$ and is given by (4.1)

$$M_r(x, y) = \text{median}(F({}^x_p \downarrow I_q^y)) \quad (4.1)$$

Similarly, the components for green, $M_g(x, y)$, and blue, $M_b(x, y)$, are calculated.

Grayscale conversion The result from the median image M is a color image with RGB values for each pixel. From this image, a grayscale image G is derived which consists of a single intensity value for each pixel. The gray-level intensity at each pixel is calculated as the average of the color values for red, green, and blue components in M . The conversion can be expressed in the form of 4.2 [10].

$$G(x, y) = 0.2989 * M_r(x, y) + 0.5870 * M_g(x, y) + 0.1140 * M_b(x, y) \quad (4.2)$$

4.4.2 Segmentation

Thresholding is applied to create a binary (black and white) image from a color or grayscale image. Essentially, the objective is to classify all the image pixels as a foreground (object) or a background pixel. In basic thresholding, a threshold value is selected. The set of pixels with gray-level intensity below the threshold τ are considered as background pixels and the remaining are considered as foreground pixels. A pixel in the binary image, $B(x, y) \in 0, 1$ is defined as in (4.3).

$$B \xrightarrow{\tau} (G) = \begin{cases} B(x, y) = 0, & \text{if } G(x, y) < \tau \\ B(x, y) = 1 & \text{otherwise} \end{cases} \quad (4.3)$$

If the threshold changes based on the content of an image, such thresholding is called as dynamic thresholding. Images vary depending on crystallization techniques and imaging devices. This makes it difficult to use a fixed threshold for binarization. Therefore, dynamic thresholding methods are preferred. A single technique does not produce desired results for all images. Therefore, it is helpful to investigate several thresholding methods perhaps by varying the thresholding parameters. Comparison of Otsu's thresholding and green percentile thresholding is provided below. Later in

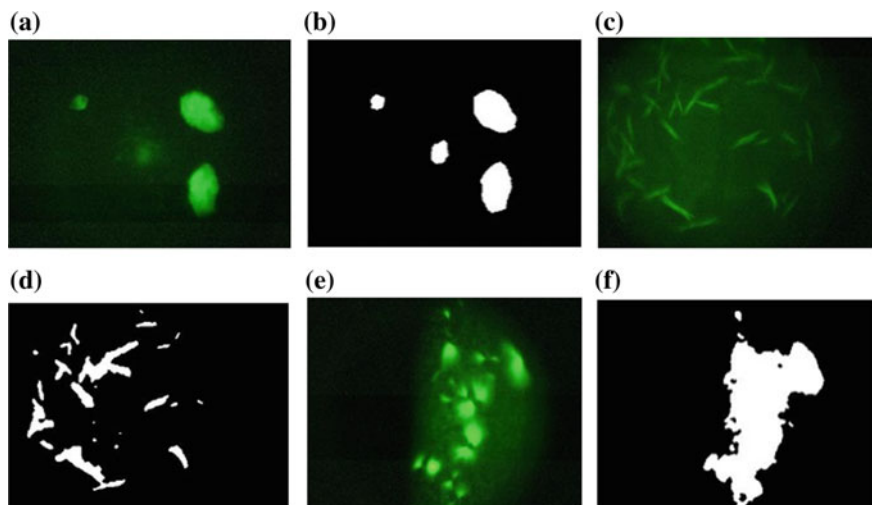


Fig. 4.5 a, c, and f: median-filtered images; b, d, and f: the Otsu thresholded images for a, c, and f, respectively. Reprinted (adapted) with permission from *Crystal Growth & Design* 2013 13 (7), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 2728–2736. Copyright (2013) American Chemical Society

Chap. 8, we introduce super-thresholding method on how to choose the best thresholding method for a crystal image.

Otsu's thresholding Otsu's method [12] iterates through all possible threshold values and calculates a measure of spread of the pixel levels in foreground or background region. The threshold value (τ_o) for which the sum of foreground and background spreads is minimal is selected, and binary image ($B_{otsu} = \xrightarrow{\tau_o} (G)$) is constructed applying this threshold. Down-sampled images and corresponding Otsu thresholded images followed by the median filter are given in Fig. 4.5.

From the original and binary images in Fig. 4.5, it can be observed that the same technique may not yield good results for all images. In the binary images shown in Fig. 4.5b and d, the objects and background are distinguished well. However, in the binary image shown in Fig. 4.5f for the original image in Fig. 4.5e, objects and the background are not well separated. Hence, the result is not as desired. If the protein solution drop is also illuminated, crystals are not distinguishable in the thresholded image. This causes difficulty in extracting correct features from the image.

Green percentile image thresholding When green light is used as the excitation source for fluorescence-based acquisition, the intensity of the green pixel component is observed to be higher than the red and blue components in the crystal regions [21]. This feature can be utilized for green percentile image binarization. Let τ_p be the intensity of green component such that the number of pixels in the image with green component below τ_p constitutes p% of the pixels. For example, if $p = 90\%$, τ_{90} is the intensity of green such that 90% of the green component pixels will be less than

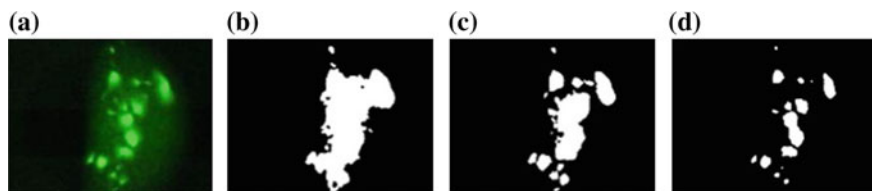


Fig. 4.6 Results showing the application of the three thresholding techniques for a sample image. **a** Original, **b** Otsu's threshold, **c** Green percentile threshold ($p = 90$), **d** Green percentile threshold ($p = 99$). Reprinted (adapted) with permission from *Crystal Growth & Design* 2013 13 (7), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 2728–2736. Copyright (2013) American Chemical Society

τ_{90} . Image binarization is then done using the value of τ_p and a minimum gray-level intensity condition $\tau_{min} = 40$. All pixels with gray-level intensity greater than τ_{min} and having green pixel component greater than τ_p constitute the foreground region, while the remaining pixels constitute the background region. As the value of p goes higher, the foreground (object) region in the binary image usually becomes smaller. In the original Crystal X2, $p = 90$ and $p = 99$ are used as two green percentile thresholding techniques. These are represented as G_{90} and G_{99} . G_{99} threshold only maintains the pixels of the highest intensity.

Figure 4.6 shows the results of applying three thresholding techniques for a crystallization trial image. In the binary image using Otsu's method (Fig. 4.6b), the crystal region information is lost. Hence, the result is not as desired. The G_{90} method (Fig. 4.6c) performs slightly better. For this particular image, Fig. 4.6d provides the best result as the crystal regions are well separated from the background. If the protein solution drop is also illuminated, crystals are not distinguishable in the thresholded image. This causes difficulty in extracting correct features from the image. Therefore, instead of relying on a single thresholding method, it is better to apply multiple thresholding techniques and extract features.

These experiments show that different thresholding techniques provide good thresholding for different images. Otsu's method [12] produced results that were useful to identify the large regions such as precipitates and solution droplet. However, the method performed poorly to separate the crystal objects. Note that the objective here is to identify the objects in the images and then be able to extract features related to those objects that help in classifying it to a particular category and differentiate from others. In this regard, the features from Otsu's method as well as features from the percentile methods are useful. Therefore, combining the features from multiple thresholding techniques can be helpful.

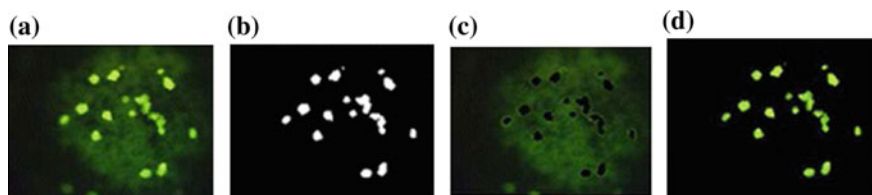


Fig. 4.7 Separating background and foreground regions. **a** M (image after noise removal), **b** binary image, **c** background pixels, **d** foreground pixels. Reprinted (adapted) with permission from Crystal Growth & Design 2013 13 (7), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 2728–2736. Copyright (2013) American Chemical Society

4.5 Feature Extraction

Rather than extracting myriad features from images, features actually useful for analysis should be analyzed. To recognize crystal categories, features related to intensity and shapes of regions after applying thresholding help to complete analysis in a short time. Chapter 5 provides in-depth comparison and analysis of features for protein crystallization. This section covers how a small number of features could be helpful to classify crystallization trial images if trace fluorescent labeling is used.

4.5.1 Intensity Features

The variation in intensity is a useful image feature. This is because crystals have the highest illumination compared to precipitates. Once a binary image is obtained, it is used as a mask to differentiate foreground and background region. After that, the features related to intensity statistics in the background and foreground region are extracted. The image in Fig. 4.7b is the binary image of the image in Fig. 4.7a obtained by applying a thresholding technique. Figure 4.7c shows the image with background pixels from the original (median-filtered) image and foreground pixels in black. Similarly, Fig. 4.7d shows the foreground pixels in the original image with background pixels in black.

Using the original image and the binary image, the following image features are extracted. Note that these features are dependent on a binary image. For feature extraction, Otsu's method, G_{90} method and G_{99} methods are applied to obtain the binary image. Using each of these binary and the median-filtered images, binary image features are obtained. The following is a set of features extracted for classifying images.

- (i) Threshold intensity (τ) for the corresponding thresholding technique.
- (ii) The number of white pixels in the binary image (N_f)

$$N_f = \sum_{i=1}^h \sum_{j=1}^w B(i, j) \quad (4.4)$$

(iii) Average image intensity in the foreground region (μ_f):

$$\mu_f = \frac{1}{N_f} \sum_{i=1}^h \sum_{j=1}^w G(i, j) \cdot B(i, j) \quad (4.5)$$

(iv) Standard deviation of intensity in the foreground region (σ_f):

$$\sigma_f = \sqrt{\frac{1}{N_f} \sum_{i=1}^h \sum_{j=1}^w ((\mu_f - G(i, j)) \cdot B(i, j))^2} \quad (4.6)$$

(v) Average image intensity in the background region (μ_b):

$$\mu_b = \frac{1}{h * w - N_f} \sum_{i=1}^h \sum_{j=1}^w G(i, j) (1 - B(i, j)) \quad (4.7)$$

(vi) Standard deviation of intensity in the background region (σ_b):

$$\sigma_b = \sqrt{\frac{1}{h * w - N_f} \sum_{i=1}^h \sum_{j=1, B(i, j)=0}^w ((\mu_b - G(i, j)) \cdot (1 - B(i, j)))^2} \quad (4.8)$$

4.5.2 Region Features

Thresholding should distinguish crystals as objects. However, other non-crystal objects might appear in the foreground. The shape and sizes of these foreground objects are important features to cluster the images into different categories.

Region Identification Connected component labeling [20] is applied on binary images to extract high-intensity regions or blobs. The binary image could be obtained from any of the thresholding methods. Let O be the set of the blobs in a binary image B , and B consists of n number of blobs. The i th largest blob is represented by O_i where $0 \leq i \leq n$ and $area(O_i) \geq area(O_{i+1})$. Each blob O_i is enclosed by a minimum bounding rectangle (MBR) centered at $m^i(x, y)$ having width w_i and height h_i . Figure 4.8b shows a binary image of the original image in Fig. 4.8a which consists of four blobs. Extraction of the individual blobs is given in Fig. 4.8c. The minimum size of the blob could be defined as 3×3 pixels. The MBR of O_i is represented as

$$\begin{matrix} m_x^i & m_y^i \\ w_i/2 & R_{h_i/2} \end{matrix}$$

For each O_i , skeletonization ($\Omega_i = \text{skel}(O_i)$) is applied to get the boundaries of the blob. The skeletonization is a hit and miss morphological operation with the structuring element S given in (4.9). Each point in a binary image of O_i where the pixel's neighborhood matches the structuring element is a hit and the corresponding pixel in the output is zero; otherwise, it remains the same. The resulting image consists of objects converted to single pixel thickness. Figure 4.8d shows the skeletonization of the blobs in Fig. 4.8c.

$$S = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \quad (4.9)$$

Region (blob) features More information about the crystals is obtained by extracting shape features like uniformity, symmetry, size, etc. of the regions. Using the original image and extracted blobs, the following blob features are extracted. There might be many number of blobs in the binary image. The large-sized blobs are more interesting than other blobs due to likelihood of being crystals. If the number of blobs is less than the maximum number of blobs, the feature values for missing blobs are set to 0.

Area of the blob (a): The area or the number of white pixels in blob O_i is represented as a_i and is calculated as in (4.10):

$$a_i = \sum_{x=1}^{h_i} \sum_{y=1}^{w_i} O_i(x, y) \quad (4.10)$$

Measure of fullness (f): Measure of fullness indicates whether the blob completely covers its MBR or not. It is calculated as the ratio of area of the blob to the area of its MBR as defined by (4.11):

$$f_i = \frac{a_i}{(w_i * h_i)} \quad (4.11)$$

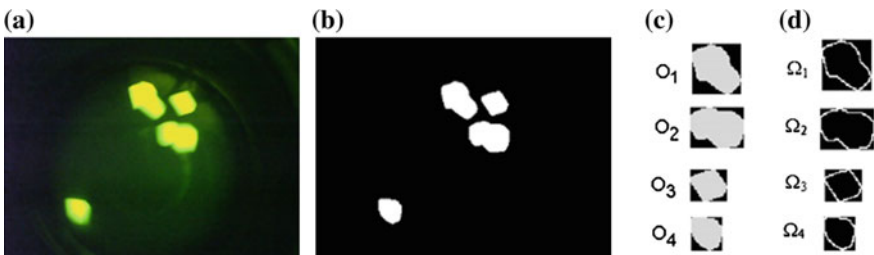


Fig. 4.8 **a** I (image after noise removal) **b** Binary image, **c** O : Objects (regions) using connected component labeling **d** Ω : the skeletonization of object. Reprinted (adapted) with permission from Crystal Growth & Design 2013 13 (7), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 2728–2736. Copyright (2013) American Chemical Society

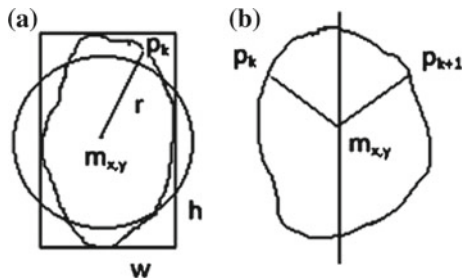


Fig. 4.9 Blob uniformity and symmetry **a** Blob uniformity, **b** Blob symmetry. Reprinted (adapted) with permission from *Crystal Growth & Design* 2013 13 (7), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 2728–2736. Copyright (2013) American Chemical Society

Boundary pixel count (N_b): The skeleton image (Ω_i) is used to compute the number of pixels in the boundary of the blob. This is calculated using (4.12):

$$N_b^i = \sum_{i=1}^{h_i} \sum_{j=1}^{w_i} \Omega_i(x, y) \quad (4.12)$$

Measure of boundary uniformity (u_1, u_2): A measure of boundary smoothness is calculated by comparing the distance of each boundary pixel from the center of the MBR to the assumed radius ($r = (w_i + h_i)/2$) as shown in Fig. 4.9a. Let P_i be the set of points on the perimeter of the blob O_i , i.e., the skeleton Ω_i . Two measures u_1^i and u_2^i are defined related to boundary uniformity defined by (4.13) and (4.14), respectively,

$$u_1^i = \frac{1}{N_b^i} \sum_{p \in P_i} |dist(p, m_{x,y}^i) - r| \quad (4.13)$$

$$u_2^i = 1 - \frac{1}{N_b^i} \sum_{p \in P_i} z_{x,y}^i \quad (4.14)$$

where $z_{x,y}^i$ is defined as in (4.15).

$$z_{x,y}^i = \begin{cases} 0 & \text{if } |dist(p, m_{x,y}^i) - r| \leq \epsilon \\ 1 & \text{otherwise} \end{cases} \quad (4.15)$$

Here, ϵ is the allowable difference and set as $(\epsilon) = 0.15 * w_i/2 * h_i/2$.

Measure of symmetry: Symmetry can be a useful measure especially in distinguishing irregular objects. The measure of left-right symmetry (symmetry along Y-axis) is calculated as shown in Fig. 4.9b. Each blob is scanned row-wise. Let p_k

be the k th boundary pixel. Then, the measure of symmetry (ζ_i) corresponding to the blob O_i is calculated as in (4.16):

$$\zeta_i = 1 - \frac{1}{N_b^i} \sum_{p \in P_i} z_{x,y}^i \quad (4.16)$$

where $z_{x,y}^i$ is defined as in (4.17)

$$z_{x,y}^i = \begin{cases} 0 & \text{if } |dist(p, m_{x,y}^i) - dist(p_{k+1}, m_{x,y}^i)| \leq \epsilon \\ 1 & \text{otherwise} \end{cases} \quad (4.17)$$






Using the above measures, nine blob-related features are computed as follows:

1. Number of blobs (η).
2. Area of the largest blob ($a_1 = \sum_{i=1}^m \sum_{j=1}^n Bx, y^1$).
3. The largest blob fullness ($f_1 = (a_1 / (w_1 * h_1))$).
4. The largest blob boundary pixel count ($N_b^1 = \sum_{i=1}^m \sum_{j=1}^n \Omega_{x,y}^1$).
5. The largest blob boundary uniformity measure (u_1^1) as defined in Eq. 4.13.
6. The largest blob uniformity measure (u_2^1) as defined in Eq. 4.14.
7. The largest blob measure of symmetry (ξ_1^1) as defined in Eq. 4.15.
8. Average area of the top five largest blobs excluding largest blob ($a_{avg} = (1/k) \sum_{i=2}^k a_i$ and $k = \min(\eta, 6)$ where is the number of blobs).
9. Average fullness of the top five largest blobs excluding largest blob ($f_{avg} = (1/k) \sum_{i=2}^k f_i$ and $k = \min(\eta, 6)$).

Table 4.1 provides the nine blob-related features for the image in Fig. 4.8a with thresholding method G_{99} . Six of these features are related to the largest blob. It should be noted that the blobs may not necessarily represent crystals in an image. In some binary images, the whole drop can appear as a large white region. Thus, the features from the largest area are important not only to identify crystals but also to identify falsely thresholded images. Besides the features from the largest blob, the average area and average fullness from top five large-sized blobs excluding the largest blob are extracted. These features provide aggregated information for the other large blobs and are especially useful to distinguish precipitates where the binary image consists of many blobs with nonuniform shapes.

For each image, three thresholding techniques are applied to obtain three binary images. From each binary image, six intensity-related features and nine blob-related features are extracted. Therefore, a total of $3 * (6 + 9) = 45$ features per image are extracted.

Table 4.1 The blobs and blob features for the image in Fig. 4.7a. Reprinted (adapted) with permission from Crystal Growth & Design 2013 13 (7), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 2728–2736. Copyright (2013) American Chemical Society

#	Features	Feature values			
					
	Regions	O_1	O_2	O_3	O_4
1	No of blobs (η)			4	
2	Area (a)	1065	1197	553	598
3	Fullness (f)	0.65	0.78	0.71	0.79
4	Boundary pixel count (N_b)	158	156	108	106
5	Boundary uniformity (u_1)	0.49	0.39	0.87	0.65
6	Boundary uniformity (u_2)	3.58	3.37	1.27	2.12
7	Measure of symmetry (ξ)	0.62	0.80	0.71	0.74
8	Average area (a_{avg})		583.33		
9	Average fullness (f_{avg})		0.50		

4.6 Accuracy and Timing Analysis

The simplest classification of the crystallization trials distinguishes between the non-crystals (trial images not containing crystals) and crystals (images having crystals). However, misclassification of crystals as non-crystals leads to a critical miss. There is no perfect classification system, and each classification system is susceptible to missing crystals. If two classes (crystals and non-crystals) are defined, the expert needs to go over images classified as (a) crystals to verify them and (b) non-crystals to detect missing crystals. This would require the expert to check all images, and this type of classification is not helpful for the expert. Instead, rather than using two categories (i.e., crystals and non-crystals) a third category, likely-leads, between crystals and non-crystals, could be added. The expert may need to scan “likely-leads” class to detect missing crystals but not the images in non-crystals. This would save significant effort in terms of manual scanning all images.

Hampton’s research defines a scoring system having a range of nine outcomes for a crystallization trial. In this study, images are first categorized into three basic categories: non-crystals, likely-leads, and crystals. The mapping of these phases into image categories with respect to Hampton’s research is provided in Table 2.1 in Chap. 2. The dataset used in experiments consists of 2250 images that are manually classified by an expert. Most images belong to the non-crystal category. Additional crystal images are added into the dataset to include all kinds of crystals and to reduce

the class imbalance in the training. The distribution of the image categories is given in Table 4.2.

Testing is done by applying tenfold cross-validation. In this process, the entire training set is first split randomly into 10 equal-sized subsamples. Each subsample is used for testing, while remaining subsamples are used for training. This process is repeated 10 times with each subsample being used exactly once for testing. The results are combined to get a single estimation for the complete training set.

Ensemble classifiers can help reduce the risk of missing crystals. To show the advantage of using an ensemble classifier, results using Multilayer Perceptron Neural Network (MLP) are compared with max-ensemble classifier [21].

4.6.1 Multilayer Perceptron Neural Network (MLP)

MLP is a widely used classification algorithm in pattern recognition problems [6]. The model consists of one or more hidden layers between input and output layers and weights are associated with connecting nodes. Training is done using back-propagation learning algorithm. MLP classifier is applied over a 45-dimensional vector obtained using the features from all three thresholding methods (Otsu, G_{90} , G_{99}).

The classification results using MLP with a single hidden layer, 24 nodes in the hidden layer, and 0.3 learning rate are provided in the form of a contingency table in Table 4.3. The overall accuracy is 90% [(1469 + 299 + 262)/2250]. Table 4.3 also provides the precision–recall using one-vs-all for each category. The non-crystals are fairly well detected (97%). This category corresponds to the crystallization conditions which are discarded from further experiments. Since most of the images belong to this category, the effort for manual review of the classification results is greatly reduced. Likewise, the recall for the likely-leads and crystals categories are 0.74 and 0.79, respectively. The system misses around 2% (6 out of 332) actual crystals. The images classified as likely-leads and crystals are to be reviewed by an expert. The misclassification of the non-crystal images to the higher categories leads to 6% [(42 + 3)/(405 + 324)] unnecessary checks for the images that surely do not contain

Table 4.2 Distribution of images into different categories. Reprinted (adapted) with permission from *Crystal Growth & Design* 2013 13 (7), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 2728–2736. Copyright (2013) American Chemical Society

Category	No of images	Percentage %
Non-crystals	1514	67.3
Likely-leads	404	18.0
Crystals	332	14.8
Total images	2250	

crystals. The precision and recall for non-crystal category are very high compared to the measures for the other two categories.

4.6.2 Max-Class Ensemble Method

Ensemble methods provide a model for combining predictions from multiple classifiers. Essentially, the goal is to reduce the risk of misclassification. Bagging and boosting are two popular methods of selecting samples for ensemble methods [23]. Most often, majority voting or class-averaging is used to determine the result score from an ensemble classifier. Protein crystallization has a class imbalance problem. Not necessarily all classes are represented by the same amount. All precipitates start with the first state and only successful crystallization process will lead to the last state (crystalline outcome). The number of precipitates that lead to crystals is minority. Typical classifiers are biased toward the crowded classes and try to predict them with high sensitivity. Although overall accuracy is improved, the crystals might be missed. The cost of missing a crystal is significantly high. A majority voting approach that is used by traditional ensemble techniques might fail for these cases. The max-class ensemble method that can minimize the risk of missing crystals works as follows. Let $M_k^t(p_m)$ denote the class of the precipitate p_m using classifier M_k at time instant t . Then, the max-class ensemble method is defined as $\max_{1 \leq k \leq w} (\max_{1 \leq t \leq T} M_k^t(p_m))$, where $1 \leq k \leq w$ and $1 \leq t \leq T$ assuming w classifiers and T observations.

Feature extraction depends on the quality and correctness of the binary (or thresholded) images. As mentioned earlier, the comparative performance of thresholding techniques may vary for different images. Therefore, in this max-ensemble method, three MLP classifiers are executed using the features from each thresholding method (Otsu, G_{90} , G_{99}). Another MLP classifier is executed with all features combined. Each image has now four predicted classes which could be the same or different. The resulting class (or score) is the maximum class (or score) from all these classifiers.

Table 4.3 Classification results using MLP classifier. Reprinted (adapted) with permission from Crystal Growth & Design 2013 13 (7), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 2728–2736. Copyright (2013) American Chemical Society

Actual/Observed	0	1	2	Actual total	Recall
0	1469	42	3	1514	0.97
1	46	299	59	404	0.74
2	6	64	262	332	0.79
Observed total	1414	383	453	2250	0.83
Precision	0.97	0.74	0.81	0.84	

Table 4.4 Classification results using max-class ensemble classifier. Reprinted (adapted) with permission from Crystal Growth & Design 2013 13 (7), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 2728–2736. Copyright (2013) American Chemical Society

Actual/Observed	0	1	2	Actual total	Recall
0	1402	97	15	1514	0.93
1	8	270	126	404	0.67
2	4	16	312	332	0.94
Observed total	1414	383	453	2250	0.84
Precision	0.99	0.70	0.69	0.80	

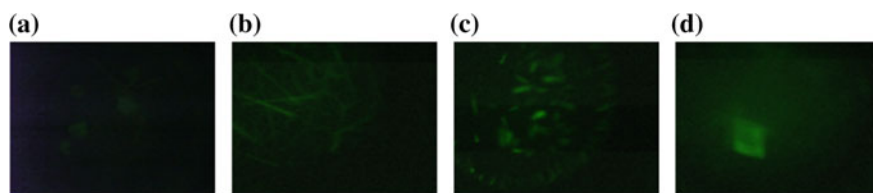


Fig. 4.10 a–d Crystals classified as non-crystal using max-class ensemble classifier. Reprinted (adapted) with permission from Crystal Growth & Design 2013 13 (7), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 2728–2736. Copyright (2013) American Chemical Society

Classifying a crystal as a non-crystal is a more critical problem than classifying a non-crystal as a crystal. Because of the cost of missing a crystal, the critical performance measure for protein crystallization dataset is the recall of the crystal category. The recall using MLP classifiers is 0.79, which is not high. The max-class ensemble classifier is biased toward high classes (or scores). The classification results using a max-class ensemble classifier over four MLP classifiers (three MLP classifiers with 15-dimensional feature inputs from Otsu, G_{90} and G_{99} thresholding methods, and another MLP classifier with 45-dimensional feature input) is given in Table 4.4. The max-ensemble class chooses the highest score (or class) among these classifiers (Table 4.4).

The overall accuracy with max-class ensemble method is 88% [(1402 + 270 + 312)/2250]. In comparison to the first approach, the number of false negatives (i.e., classifying an image in a high class to a lower class) is decreased at the cost of increase in false positives (i.e., classifying an image in a low class to a higher class). Both the precision and recall for class 0 (non-crystals) are very high. The precision for class 0 (non-crystals) is close to 99%. The recall for class 2 is increased from 0.79 (first method) to 0.94 (max-class ensemble method). Only 1.2% (4 out of 332) actual crystals are predicted as non-crystals. Figure 4.10 shows crystal images that are predicted as non-crystal. The image intensity in the images in Fig. 4.10a, b is very low. The crystal images missed have either very low image intensity or are blurred. Higher intensity excitation lighting, camera gain settings, and adequate focusing can help eliminate such errors.

4.6.3 Computation Time

System Information A sample application is developed as a windows form-based application in C#. Image feature extraction is implemented using AForge Imaging library (<http://code.google.com/p/aforge/>). Training and classification is implemented using Weka data mining library (<http://www.cs.waikato.ac.nz/ml/weka/>). Visual Studio 2010 is used as the IDE. The user interface consists of a tabbed layout for image acquisition, image scoring, and system settings. The repository of images is maintained inside a directory and the records are maintained in MySQL database. On a Windows 7 Intel Core i7 CPU @2.8 GHz system with 4 GB memory, it takes around 1 h 40 min to process and classify 2250 images. This amounts to less than 3 s to process and classify an image, and fits well into the average sample to sample translation time of 6 s for the described system.

4.7 Summary

This chapter provided the design and implementation of a stand-alone system for protein crystallization image acquisition and crystallization. The image acquisition utilizes a low-cost in-house assembled fluorescence microscopy system. Image analysis is carried out on the fluorescence images. The main advantage of this approach is the ability to rapidly identify crystals and potential lead crystallization conditions by analyzing image intensity and high-intensity regions. The implementation of an efficient (fast) and effective (with good accuracy) image classification system to automatically classify the images into one of non-crystal, likely-leads, or crystal categories is also explained.

The max-class ensemble method is able to reduce the risk of error, and the percentage of missing crystals as non-crystals is around 1.2%. This means that the system only misses 1.2% of the crystals. Since the described system exhibits high accuracy for the non-crystal category, this minimizes unnecessary reviews (images in non-crystal category) by the expert. Therefore, the effort for manual review is greatly reduced. Image processing for a 96-well plate with three cells in each well takes less than 15 min. This time is less than the time for acquisition which takes around 30 min to scan through the whole plate. This allows the image acquisition and classification to be executed in parallel. Even though the correct classification of non-crystal images is very high, the system does not distinguish between the likely-leads and crystal categories very well. This makes the manual review essential for the two categories.

Acknowledgements The majority of this chapter is Reprinted (adapted) with permission from *Crystal Growth & Design* 2013 13 (7), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 2728–2736. Copyright (2013) American Chemical Society. Some modifications have been made to fit into this book.

References

1. Bern, M., Goldberg, D., Stevens, R. C., & Kuhn, P. (2004). Automatic classification of protein crystallization images using a curve-tracking algorithm. *Journal of Applied Crystallography*, 37(2), 279–287.
2. Berry, I. M., Dym, O., Esnouf, R., Harlos, K., Meged, R., Perrakis, A., et al. (2006). Spine high-throughput crystallization, crystal imaging and recognition techniques: current state, performance analysis, new technologies and future aspects. *Acta Crystallographica Section D: Biological Crystallography*, 62(10), 1137–1149.
3. Cumbaa, C., & Jurisica, I. (2005). Automatic classification and pattern discovery in high-throughput protein crystallization trials. *Journal of Structural and Functional Genomics*, 6(2–3), 195–202.
4. Cumbaa, C. A., & Jurisica, I. (2010). Protein crystallization analysis on the world community grid. *Journal of Structural and Functional Genomics*, 11(1), 61–69.
5. Cumbaa, C. A., Lauricella, A., Fehrman, N., Veatch, C., Collins, R., Luft, J., et al. (2003). Automatic classification of sub-microlitre protein-crystallization trials in 1536-well plates. *Acta Crystallographica Section D: Biological Crystallography*, 59(9), 1619–1627.
6. Duda, R. O., & Hart, P. E. (1972). Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1), 11–15.
7. Forsythe, E., Achari, A., & Pusey, M. L. (2006). Trace fluorescent labeling for high-throughput crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 62(3), 339–346.
8. Liu, R., Freund, Y., & Spraggon, G. (2008). Image-based crystal detection: a machine-learning approach. *Acta Crystallographica Section D: Biological Crystallography*, 64(12), 1187–1195.
9. Luft, J. R., Newman, J., & Snell, E. H. (2014). Crystallization screening: the influence of history on current practice. *Structural Biology and Crystallization Communications*, 70(7), 835–853.
10. MATLAB. (2013). *version 7.10.0 (R2013a)*. The MathWorks Inc., Natick.
11. Onzalez, R., & Woods, R. (2008). *Digital image processing*. Prentice Hall.
12. Otsu, N. (1975). A threshold selection method from gray-level histograms. *Automatica*, 11(285–296), 23–27.
13. Pan, S., Shavit, G., Penas-Centeno, M., Xu, D.-H., Shapiro, L., Ladner, R., et al. (2006). Automated classification of protein crystallization images using support vector machines with scale-invariant texture and gabor features. *Acta Crystallographica Section D: Biological Crystallography*, 62(3), 271–279.
14. Po, M. J., & Laine, A. F. (2008). Leveraging genetic algorithm and neural network in automated protein crystal recognition. In *30th annual international conference of the IEEE engineering in medicine and biology society, 2008. EMBS 2008* (pp. 1926–1929): IEEE.
15. Pusey, M., Forsythe, E., & Achari, A. (2008). Fluorescence approaches to growing macromolecule crystals. In *Structural proteomics* (pp. 377–385): Springer.
16. Pusey, M. L., Liu, Z.-J., Tempel, W., Praissman, J., Lin, D., Wang, B.-C., et al. (2005). Life in the fast lane for protein crystallization and X-ray crystallography. *Progress in Biophysics and Molecular Biology*, 88(3), 359–386.
17. Pusey, M., Barcena, J., Morris, M., Singhal, A., Yuan, Q., & Ng, J. (2015). Trace fluorescent labeling for protein crystallization. *Acta Crystallographica Section F: Structural Biology Communications*, 71(7), 806–814.
18. Saitoh, K., Kawabata, K., & Asama, H. (2006). Design of classifier to automate the evaluation of protein crystallization states. In *Proceedings 2006 IEEE international conference on Robotics and automation, 2006. ICRA 2006* (pp. 1800–1805): IEEE.
19. Saitoh, K., Kawabata, K., Kunimitsu, S., Asama, H., & Mishima, T. (2004). Evaluation of protein crystallization states based on texture information. In *Proceedings. 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004 (IROS 2004)* (Vol. 3, pp. 2725–2730): IEEE.
20. Shapiro, L., & Stockman, G. C. (2001). *Computer vision*. ed: Prentice Hall.

21. Sigdel, M., Pusey, M. L., & Aygun, R. S. (2013). Real-time protein crystallization image acquisition and classification system. *Crystal Growth and Design*, 13(7), 2728–2736.
22. Spraggon, G., Lesley, S. A., Kreusch, A., & Priestle, J. P. (2002). Computational analysis of crystallization trials. *Acta Crystallographica Section D: Biological Crystallography*, 58(11), 1915–1923.
23. Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining* (1st ed.). Boston: Addison-Wesley Longman Publishing Co., Inc.
24. Yang, X., Chen, W., Zheng, Y. F., & Jiang, T. (2006). Image-based classification for automating protein crystal identification. In *Intelligent computing in signal processing and pattern recognition* (pp. 932–937): Springer.
25. Zhu, X., Sun, S., & Bern, M. (2004). Classification of protein crystallization imagery. In *26th annual international conference of the IEEE engineering in medicine and biology society, 2004. IEMBS'04* (Vol. 1, pp. 1628–1631): IEEE.
26. Zuk, W. M., & Ward, K. B. (1991). Methods of analysis of protein crystal images. *Journal of Crystal Growth*, 110(1), 148–155.

Chapter 5

Classification of Crystallization Trial Images

Abstract Large number of features are extracted from protein crystallization trial images to improve the accuracy of classifiers for predicting the presence of crystals or phases of the crystallization process. The excessive number of features and computationally intensive image processing methods to extract these features make utilization of automated classification tools on stand-alone computing systems inconvenient due to the required time to complete the classification tasks. In this chapter, we provide an analysis of combinations of image feature sets, feature reduction, and classification techniques for crystallization images benefiting from trace fluorescent labeling. Features are categorized into intensity, graph, histogram, texture, shape-adaptive, and region features (using binarized images generated by Otsu's, green percentile, and morphological thresholding). The effects of normalization, feature reduction with principal components analysis (PCA), and feature selection using random forest classifier are also investigated. Moreover, the time required to extract feature categories is computed and an estimated time of extraction is provided for feature category combinations. The analysis in this chapter shows that research groups can select features according to their hardware setups for real-time analysis.

5.1 Introduction

Protein crystallization is a highly empirical process that depends on numerous factors such as pH and temperature of the environment, protein concentration, the type of precipitant, ionic strength of the solution, gravity, the crystallization methods, etc. [24] A combination of all these factors suitable for the protein being crystallized is critical for the formation of crystals, and the prediction of these parameters is quite challenging since there is no prior information about the protein solubility [12, 14]. Therefore, thousands of experimental trials may be required for successful crystallization. Today, high-throughput robotic systems are routinely used to increase the chance of successfully obtaining crystals. Because of the high-throughput crystallization trials, manual review of crystallization trials becomes practically discouraging in terms of time and resources. Therefore, automated image scoring systems have

been developed to collect and classify crystallization trial images. The fundamental aim is to discard the unsuccessful trials, identify the successful trials, and possibly identify those trials which could be optimized.

5.1.1 Challenges of Protein Crystallization Classification

Imaging techniques are used to capture the state change or the possibility of forming crystals [25]. Building a reliable system to classify and analyze a crystallization trial can be very helpful to the crystallographers by reducing the number of tedious manual reviews of unsuccessful outcomes or providing the phase of the crystallization process. Such a system requires extracting features from images. After these features are used to train a classifier, the classifier model is used to classify new trial images. However, building a classifier model with high accuracy is challenging due to the following reasons.

1. **Many Phases of Crystallization Process.** The instruction sheets with crystallization screens from Hampton Research describe 9 possible protein crystallization trial outcomes or phases¹ [15] (Clear drop, Phase separation, Granular precipitate, Microcrystals, Posettes/spherulites, Needles, 2D Plates, Small 3D crystals, Large 3D crystals). Figure 5.1 shows sample protein crystallization trial images obtained using trace fluorescent labeling [32], where each image corresponds to a specific phase of crystallization. In an analysis of the screening images, it is important to predict/detect the current phase of the experiment. Phases that yield crystalline outcomes or likely-leads are more valuable than other categories. Misclassification of the images in a higher category (e.g., crystal category) into a lower category (e.g., non-crystal category) is a serious problem as it results in a lead condition being missed. The misclassification of a lower category result to a higher is not as serious, and can be considered as a cost of capturing all possible leads.
2. **Unbalanced Distribution of Data.** The distribution of data in different categories (or phases) is unbalanced. Frequency of higher (crystalline) categories are less than the frequency of lower categories. The classification models can be affected adversely by the unbalanced distribution. They may classify in favor of more frequent but less important categories.
3. **Complexity of Image Analysis.** Nonuniform shapes and varying orientation of crystals impose complexity in image analysis. Intra-class diversity of a single crystal subcategory is significantly high. It is difficult to build a classifier with high accuracy that can model all variations.
4. **Multiple Types of Crystals in a Single Image.** A single image can consist of objects (crystals) in different morphologies, such as dendrites and 3D crystals. In

¹<http://hamptonresearch.com>.

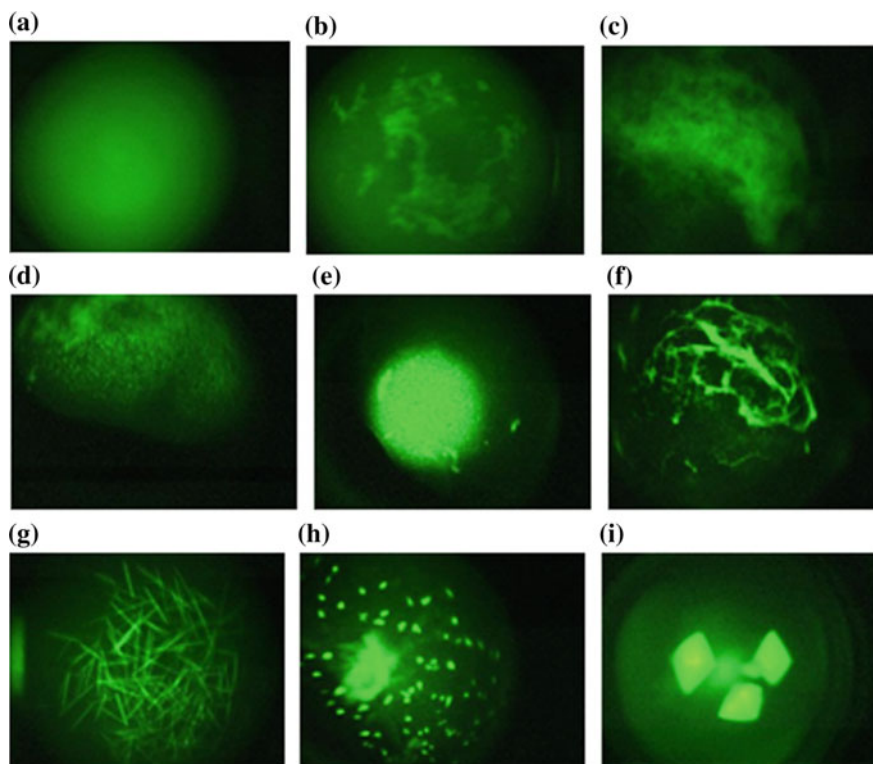


Fig. 5.1 Sample protein crystallization trial images. Reprinted (adapted) with permission from *Crystal Growth & Design* 2013 13 (7), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 2728-2736. Copyright (2013) American Chemical Society

such cases, the expected class for the image would be the class corresponding to the highest class among all crystal objects.

- 5. Low and Varying Image Quality.** Since crystals are floating in a 3D well, not all crystals may be captured in focus. To observe the phases of crystallization, images are captured a number of times during the process. The lighting conditions may vary each time the images are collected. Varying illumination and focusing affect the preprocessing of images and features used for classification.
- 6. Ambiguity in Labeling Trial Images.** Protein crystallization is an evolving process. In some scenarios, there is a semantic transition between categories, meaning the images cannot be clearly assigned to one category. Similarly, ambiguities and subjectivity of the viewer or an expert can affect the labeling process or expert scoring.

5.1.2 Factors for Classification

In general, protein crystallization trial image analysis work is compared with respect to the accuracy of classification. The accuracy depends on the number of categories, features, and the ability of classifiers to model the data. Moreover, the hardware resources, training time, and real-time analysis of new images are important factors that affect the usability of these methods.

The Number of Categories A significant amount of previous work (for example, Zuk and Ward [52], Cumba et al. [8], Cumba et al. [5], Zhu et al. [51], Berry et al. [2], Pan et al. [29], Po and Laine [30]) classified crystallization trials into non-crystal or crystal categories. Yang et al. [48] classified the trials into three categories (clear, precipitate, and crystal). Bern et al. [1] classified the images into five categories (empty, clear, precipitate, microcrystal hit, and crystal). Likewise, Saitoh et al. [34] classified into five categories (clear drop, creamy precipitate, granulated precipitate, amorphous state precipitate, and crystal). Spraggon et al. [43] proposed classification of the crystallization images into six categories (experimental mistake, clear drop, homogeneous precipitant, inhomogeneous precipitant, microcrystals, and crystals). Cumba et al. [7] developed a system that classifies the images into three or six categories (phase separation, precipitate, skin effect, crystal, junk, and unsure). Yann et al. [49] classified into 10 categories (clear, precipitate, crystal, phase, precipitate and crystal, precipitate and skin, phase and crystal, phase and precipitate, skin, and junk). It should be noted that there is no standard for categorizing the images, and different research studies proposed different categories in their own way. Hampton's scheme specifies 9 possible outcomes of crystallization trials. The classification in this chapter is based on Hampton's scale in Table 2.1 in Chap. 2.

Features for Classification. For feature extraction, a variety of image processing techniques have been proposed. Zuk and Ward [52] used the Hough transform to identify straight edges of crystals. Bern et al. [1] extract gradient and geometry-related features from the selected drop. Pan et al. [29] used intensity statistics, blob texture features, and results from Gabor wavelet decomposition to obtain the image features. Research studies by Cumba et al. [8], Saitoh et al. [35], Spraggon et al. [43], and Zhu et al. [51] used a combination of geometric and texture features as the input to their classifier. Saitoh et al. [34] used global texture features as well as features from local parts in the image and features from differential images. Yang et al. [48] derived the features from gray-level co-occurrence matrix, Hough transform, and discrete Fourier transform (DFT). Liu et al. [22] extracted features from Gabor filters Gabor wavelet, filter, integral histograms, and gradient images to obtain 466-dimensional feature vector. Po and Laine [30] applied multiscale Laplacian pyramid filters and histogram analysis techniques for feature extraction. Similarly, other extracted image features included Hough transform features [30], discrete Fourier transform features [45], features from multiscale Laplacian pyramid filters [47], histogram analysis features [5], Sobel-edge features [46], etc. Cumba et al. [7] presented the most sophisticated feature extraction techniques for the classification of crystallization trial images. Features such as basic statistics, energy, Euler numbers, Radon–Laplacian features,

Sobel-edge features, microcrystal features, and gray-level co-occurrence matrix features were extracted to obtain a 14, 908-dimensional feature vector. They utilized a web-based distributed system and extracted as many features as possible hoping that the huge set of features could improve the accuracy of the classification [7].

Time Analysis of Classification. Because of the high-throughput rate of image collection, the speed of processing an image becomes an important factor. The system by Pan et al. [29] required 30 s per image for feature extraction. Po and Laine mentioned that it took 12.5 s per image for the feature extraction in their system [30]. Because of high computational requirement, they considered implementation of their approach on the Google computing grid. Feature extraction described by Cumba et al. [7] is the most sophisticated, which could take 5 h per image on a normal system. To speedup the process, they executed the feature extraction using a web-based distributed computing system. Yann et al. [49] utilized deep convolutional neural network (CNN), where training took 1.5 days for 150,000 weights and around 300 passes and classification takes 86 ms for 128x128 image on their GPU-based system.

Classifiers for Protein Crystallization. To obtain the decision model for classification, various classification techniques have been used. Zhu et al. [51] and Liu et al. [22] applied a decision tree with boosting. Bern et al. [1] used a decision tree classifier with handcrafted thresholds. Pan et al. [29] applied a support vector machines (SVM) learning algorithm. Saitoh et al. [34] applied a combination of decision tree and SVM classifiers. Spraggon et al. [43] applied self-organizing neural networks. Po et al. [30] combined genetic algorithms and neural networks to obtain a decision model. Berry et al. [2] determined scores for each object within a drop using self-organizing maps, learning vector quantization, and Bayesian algorithms. The overall score for the drop was calculated by aggregating the classification scores of individual objects. Cumba et al. [8] and Saitoh et al. [35] applied linear discriminant analysis. Yang et al. [48] applied hand-tuned rules-based classification followed by linear discriminant analysis. Cumba et al. [5] used association rule mining, while Cumba et al. [7] used multiple random forest classifiers generated via bagging and feature subsampling. In [38], classification performance using semi-supervised approaches was investigated. The recent study by Hung et al. [19] proposed protein crystallization image classification using elastic net. Dinc et al. [13] evaluated the classification performance using 5 different classifiers, and feature reduction using principal components analysis (PCA) and normalization methods for the non-crystal and likely lead datasets. Yann et al. [49] utilized deep convolutional neural networks (CNN) with 13 layers: (0) 128x128 image, (1) contrast normalization, (2) horizontal mirroring, (3) transformation, (4) convolution (5x5 filter), (5) max pooling (2x2 filter), (6) convolution (5x5 filter), (7) max pooling (2x2 filter), (8) convolution (5x5 filter), (9) max pooling (2x2 filter), (10) convolution (3x3 filter), (11) 2048 node fully connected layer, (12) 2048 fully connected layer for rectified linear activation, and (13) output layer using softmax.

Accuracy of Classification. With regard to the correctness of a classification, the best reported accuracy for the binary classification (i.e., classification into two categories)

is 96.56% (83.6% true positive rate and 99.4% true negative rate) using deep CNN [49]. Despite high accuracy rate, around 16% of crystals are missed. Using genetic algorithms and neural networks [30], an accuracy of 93.5% average true performance (88% true positive and 99% true negative rates) is achieved for binary classification. Saitoh et al. achieved accuracy in the range of 80–98% for different image categories [35]. Likewise, the automated system by Cumba et al. [7] detected 80% of crystal-bearing images, 89% of precipitate images, and 98% of clear drops accurately. The accuracy also depends on the number of categories. As the number of categories increases, the accuracy goes down since there are more misclassifications possible. For 10-way classification using deep CNN, Yann et al. [49] achieved 91% accuracy with around 76.85% true positive rate for crystals and 8% of crystals categorized into classes not related to crystals. While overall accuracy is important, true positive rate (recall or sensitivity) for crystals may carry more value. As crystallographers would like to trust these automated classification systems, it is not desirable to see successful crystalline cases are missed by these systems.

Table 5.1 provides an overview of factors that affect classification in the literature. The analysis in this chapter will show whether it is possible to achieve high accuracy with a small set of feature set using a proper classifier considering as many as 10 categories for real-time analysis. An exhaustive set of experiments using all feature combinations and representative classifiers are conducted to achieve real-time analysis.

5.1.3 Feature Analysis for Building Real-Time Classifiers

The task of building classifier models with high accuracy in the presence of aforementioned issues is challenging. To improve the classification performance, there has been a trend to increase the number of image features and size of datasets. Since it is not known which features may be helpful, all possible features that can be extracted are used to train classifiers hoping that irrelevant features are automatically eliminated or given low weights by the classifiers. For example, Cumba et al. [7] extracted 14,908 dimensional feature vector per image for classifying protein crystallization images. Overall, the image processing and feature extraction have been computationally expensive for huge number of features making it unfeasible for real-time processing. Such systems employ high-performance, grid, distributed, or cloud computing systems for manipulating large feature sets. Acquisition of high-end, high-performance, and expensive computing systems becomes a barrier for small research labs with limited resources and budget to develop and experiment new promising ideas in a timely manner.

Since extracting numerous features puts a significant computational burden on a typical stand-alone computing system, experts may need to wait for hours before seeing the classification results. Reduction of features is inevitable for building real-time classifiers. A wide number of techniques used white light imaging for extracting features. The feature extraction and image processing are cumbersome for white

Table 5.1 Factors affecting classification

Work	Image categories	Feature extraction	Classification method	Classification accuracy
Zuk and Ward [52]	NA	Edge features	Detection of lines using Hough transform and line tracking	not provided
Walker et al. [45]	7	Radial and angular descriptors from Fourier Transform	Learning vector quantization	14–97% for different categories
Xu et al. [47]	2	Features from multiscale Laplacian pyramid filters	Neural network	95% accuracy
Wilson [46]	3	Intensity and geometric features	Naive Bayes	Recall 86% for crystals, 77% for unfavorable objects
Hung et al. [19]	3	Shape context, Gabor filters, Gabor wavelet, filter, and Fourier transforms	Cascade classifier on naive Bayes and random forest	74% accuracy
Spraggon et al. [43]	6	Geometric and texture features	Self-organizing neural networks	47–82% for different categories
Cumba et al. [8]	2	Radon transform line features and texture features	Linear discriminant analysis	85% accuracy with ROC 0.84
Saitoh et al. [35]	5	Geometric and texture features	Linear discriminant analysis	80–98% for different categories
Bern et al. [1]	5	Gradient and geometric features	Decision tree with handcrafted thresholds	12% FN and 14% FP
Cumba et al. [5]	2	Texture features, line measures, and energy measures	Association rule mining	85% accuracy with ROC 0.87
Zhu et al. [51]	2	Geometric and texture features	Decision tree with boosting	14.6% FP and 9.6% FN
Berry et al. [2]	2	NA	Learning vector quantization, self-organizing maps, and bayesian algorithm	NA

(continued)

Table 5.1 (continued)

Work	Image categories	Feature extraction	Classification method	Classification accuracy
Pan et al. [29]	2	Intensity stats, texture features, Gabor wavelet decomposition	Support vector machine	2.94% FN and 37.68% FP
Yang et al. [48]	3	Hough transform, DFT, GLCM features	Hand-tuned thresholds	85% accuracy
Saitoh et al. [34]	5	Texture features, differential image features	Decision tree and SVM	90% for 3-class problem
Po and Laine [30]	2	Multiscale Laplacian pyramid filters and histogram analysis	Genetic algorithm and neural network	Accuracy: 93.5% with 88% TP and 99% TN
Liu et al. [22]	Crystal likelihood	Features from Gabor filters, integral histograms, and gradient images	Decision tree with boosting	ROC 0.92
Cumba et al. [7]	3 and 6	Basic stats, energy, Euler numbers, Radon-Laplacian, Sobel-edge, GLCM	Multiple random forest with bagging and feature subsampling	Recall 80% crystals, 89% precipitate, 98% clear drops
Sigdel et al. [40]	3	Intensity and blob features	Multilayer perceptron neural network	1.2% crystal misses with 88% accuracy
Sigdel et al. [38]	3	Intensity and blob features	Semi-supervised	75–85% overall accuracy
Dinc et al. [13]	3 and 2	Intensity and blob features	5 classifiers, feature reduction using PCA	96% on non-crystals, 95% on likely-leads
Yann et al. [49]	10	Deep learning on grayscale image	Deep CNN with 13 layers	90.8% accuracy

light images. For experiments in this chapter, Crystal X2 [40] system developed at iXpressGenes, Inc. is used, and then captured images of trace fluorescent-labeled crystallization trials [32] are analyzed. The crystal regions have high intensity in images when trace fluorescent labeling is used. The high contrast between the background and crystals alleviates the image processing and feature extraction. Hence, the number of features can be reduced significantly. Another reason for feature reduction is that the use of irrelevant features may deteriorate the performance of some classifiers. Therefore, it is very important to determine the minimal set of image features that can be used to obtain a reliable classification performance.

Herein, the image features, feature reduction techniques and classification techniques for the images captured using trace fluorescent labeling are investigated. A number of feature set combinations are experimented, some new features are introduced, and a combination of feature sets is proposed for a real-time classification system while maintaining comparatively high accuracy. To identify the relevant set of features for this problem domain, trying all combinations of features is not feasible. Hence, features are categorized into intensity, region, graph, histogram, texture, and shape-adaptive features. Region features are extracted using binarized images generated by Otsu's [28], green percentile thresholding, and morphological thresholding. The effects of normalization, feature reduction with principle components analysis (PCA) [20], and feature selection using random forest classifier are also evaluated. The time required to extract feature categories is computed and an estimated time of feature extraction is provided for feature category combinations. In this way, research groups may ignore some feature groups since they may not have significant effect on the accuracy. This also enables research groups to select features according to their hardware setups for real-time analysis.

In this chapter, a 9-point scoring system (Hampton's scores) is used to classify protein crystallization trial images using hierarchical classification. The first-level of classification categorizes into non-crystals, likely-leads, and crystals. The total number of subcategories is 10 (one more than Hampton's scale to include a category for unclear bright images as shown in Fig. 2.5). The complete feature set contains around 300 features. Feature sets are categorized into 10 groups, and classifiers are evaluated exhaustively on all combinations of these feature groups. Random forest (RF), naïve Bayesian (BYS), support vector machine (SVM), decision tree (DT), and artificial neural network (ANN) classifiers are utilized in these experiments. Moreover, the performance of feature selection and normalization is investigated. The goal is to identify a minimal set of feature sets that will achieve good accuracy for real-time applications. Around 8,624 experiments (different combinations of feature categories, binarization methods, feature reduction/selection, normalization, and crystal categories) are conducted and a summary of the experimental results is provided. Based on the analysis of experiments, it is possible to answer the question: "what set of features satisfies a minimum accuracy measure m within time t ?"

5.2 Data Preprocessing

Data preprocessing may help to improve the performance of knowledge discovery from the dataset. Data preprocessing may involve application of data reduction and data transformation methods. To evaluate data reduction, a random forest feature selection with mean decrease in accuracy (*MDA – RF*) [3] method was applied. Normalization of feature vectors was also considered as some classifiers are sensitive to the ranges of features. Individual effects of z-score normalization, PCA feature reduction, and random forest feature selection methods were examined. Then, various state-of-the-art classification methods are employed in order to get benefit from different types of classifiers in the literature such as probabilistic, categorical, and ensemble classifiers.

5.2.1 Feature Normalization

Data values are measured in different scales or ranges since they have different meanings. Some classification techniques suffer from range differences because the distance metrics are highly sensitive to data range. In order to eliminate this negative effect, normalization maintains a similar range for all data by mapping the data to a predefined range or utilizing the mean and standard deviation of the data. Some classifiers benefit from normalization significantly (such as neural networks), while some of them are not affected by range differences (such as naïve Bayesian and decision trees). Z-score normalization was employed to evaluate the effects of normalization. For this, the data is normalized with respect to its mean (μ_v) and standard deviation (σ_v). The new value (v') of original data (v) is calculated as in Eq. (5.1).

$$v' = \frac{v - \mu_v}{\sigma_v} \quad (5.1)$$

5.2.2 Dimensionality Reduction and Feature Selection

It is possible to have a high number of features to represent a sample in classification problems. However, some of these features may not be informative enough and can be eliminated without any (or with minor) loss of accuracy. Some of them may be highly correlated or some of them might be measured with high noise. In such cases, data reduction techniques are offered to eliminate these useless features. PCA is one of the widely accepted techniques to reduce dimensionality [20]. In simple terms, PCA transforms complete dataset to a new subspace where every dimension is connected to an eigenvalue. The new feature corresponding to the largest eigenvalue represents the most informative feature. Using this idea, a subset of the most descriptive eigenvectors (or principal components) can be selected and rest of them

can be eliminated. The original dataset is transformed into a lower dimensional space using this subset of eigenvectors where a smaller size feature vector represents the same sample.

Another common way to reduce the size of data is feature selection. To evaluate feature selection, in this study, mean decrease in accuracy (MDA) algorithm [3] in random forest classifier is used. MDA assigns rankings to the features by randomly permuting the values of each feature and measuring the change in mean error.

5.2.3 Image Processing

Automatically determining the phase of crystallization trial images is a complex process and requires sophisticated algorithms to extract features related to the shape and size of objects in an image. Different image processing techniques are applied to the original images and then image features are extracted from several stages of these steps.

For the notations in the subsequent subsections, assume that (1) I represents an image of size $h \times w$, (2) $I(x, y)$ represents the pixel at location (x, y) where $1 \leq x \leq h$ and $1 \leq y \leq w$, (3) I_G is the green component of image I , (4) I_{gray} is the graylevel image of image I , (5) B_m represents the binary image of image I using method m , and (6) E represents edge image using edge detection methods such as Sobel or Canny.

Image Thresholding The objective of image thresholding is to simplify the image analysis by separating the foreground pixels from the background. Thresholding is often the first step in image analysis. Obtaining a good binary image is very critical in image analysis because any error in the binary image will get propagated to further processing steps. Numerous image binarization techniques have been proposed in the literature. However, as described in [10, 11] and Chap. 8 there is not a single technique which works well in all image domains. In this chapter, 3 different image binarization techniques are investigated: Otsu's threshold [28], green percentile image binarization [40] with two percentiles, and morphological thresholding [9].

Otsu's thresholding. Otsu's method [28] iterates through all possible threshold values and calculates a measure of spread of the pixel levels in foreground or background region. The threshold value (τ_o) for which the sum of foreground and background spreads is minimal is selected. The binary image ($B_{otsu} = \xrightarrow{\tau_o} (I_{gray})$) is constructed by applying this threshold to the image.

Green percentile thresholding. This method utilizes green color component of image pixels for thresholding. Let τ_p be the intensity of green component such that the number of pixels in the image with green component below τ_p constitute $p\%$ of the pixels. For example, if $p = 90\%$, τ_{90} is the intensity of green such that 90% of the green component pixels will be less than τ_{90} . Image binarization is then done using the value of τ_p and a minimum gray-level intensity condition $\tau_{min} = 40$. All pixels with gray-level intensity greater than τ_{min} and having green pixel component greater

than τ_p constitute the foreground region while the remaining pixels constitute the background region. As the value of p goes higher, the foreground (object) region in the binary image usually becomes smaller. For the given value of p , the method is represented as G_p . For example, G_{90} is the green percentile thresholding method with $p = 90\%$. G_{90} and G_{99} are applied for binarization of images in the experiments.

Morphological Thresholding. In this method, the images are binarized based on mathematical morphological operations along with some preprocessing methods. The method can be summarized as follows:

1. Apply image-opening function to get background surface: This is one of the basic mathematical morphological operations as in Eq. (5.2):

$$A \cdot B = (A \ominus B) \oplus B \quad (5.2)$$

where \ominus and \oplus denote erosion and dilation, respectively. The basic effect of the erosion operator on a binary image is to erode away the boundaries of regions of foreground pixels. In other words, after this operation the foreground regions generally shrink based on a structure element. On the other hand, after dilation operation the foreground regions generally expand.

2. Subtract background image from grayscale image.
3. Adjust pixel intensities to enhance the images: Contrast stretching is applied to increase the contrast between foreground and background.
4. Binarize the grayscale image using Otsu's thresholding method.
5. Apply image-opening function to generate the final binary image.

Region Segmentation Connected component labeling [36] is applied on binary images to extract high intensity regions or blobs. The binary image can be obtained from any of the thresholding methods. Let O be the set of the blobs in a binary image B , and B consists of n number of blobs. The i th largest blob is represented by O_i , where $1 \leq i \leq n$ and $area(O_i) \geq area(O_{i+1}), \forall i$. Each blob O_i is enclosed by a minimum bounding rectangle (MBR) centered at (m_x^i, m_y^i) having width w_i and height h_i . Ω_i represents the skeleton of blob O_i . Features related to the shape and size of the top largest blobs are extracted.

5.3 Classifiers

Classification results are highly dependent on several factors such as data type or distribution. In the literature, different classifiers are offered for different factors. In this study, 5 different classifiers were examined to determine the best classifier for this particular problem domain. The selected classifiers are described below.

1. *Decision Tree (DT)*: Decision tree is a rule-based classifier that utilizes a tree-based graph of features to decide the class of a sample. In the training stage, a tree structure is constructed where internal nodes represent features and leaf nodes

have class labels. In the testing stage, the test sample is classified by reaching the leaf node from the feature hierarchy of the tree. The decision trees are effective on categorical data types. It requires relatively less time to construct a training model (tree) and testing is also quite fast once the tree is induced [44].

2. *Random Forest (RF)*: Random forest is an ensemble-type classifier that comprises many decision tree classifiers (weak classifier). In the training stage, every decision tree is constructed based on randomly selected samples (bootstrap). Remaining samples (out-of-bag) are used in the testing stage. While constructing a decision tree, not all features are used. A feature subset is also selected randomly. For the final decision, results of all decision trees are combined based on a voting mechanism [44]. MATLAB code was used for RF which is based on algorithm by Leo Breiman et al.² [33]. The number of trees for the random forest classifier is set as 500. The square root of the total number of features is selected as the number of candidate features at one node of a decision tree [6].
3. *Support Vector Machines (SVM)*: Support Vector Machine is a binary supervised classification method. In the training stage, a decision surface (hyperplane) is determined based on boundary samples called *support vectors*. SVM tries to find the optimal hyperplane that maximizes the margin between the two classes. If the data is not linearly separable, SVM can be applied by transforming the input data to high-dimensional feature spaces using kernel functions [44].
4. *Naïve Bayesian Classifier (BYS)*: BYS is a probabilistic classifier technique that decides the class of a sample by providing the probability of its membership to the classes. The class with the highest probability is predicted as the result class. In BYS, the features of the data samples are assumed to be independent of other features. This assumption simplifies building a training model. The training stage is fast and classification is independent of the range of the feature values [44]. Also, BYS is considered to be robust to the noisy samples.
5. *Artificial Neural Networks (ANN)*: Artificial Neural Networks is a supervised classification technique that is composed of interconnected nodes (neurons). Neurons can be organized in layers depending on the complexity of the problem. It tries to learn the weights of the connections between input and output neurons to minimize the error of classification as new data are evaluated in the training stage. ANN is a commonly used technique for various classification problems such as autonomous vehicle driving, speech recognition, face recognition, etc. [26, 44]. In this study, the MATLAB built-in neural network toolbox with two layers is used. The hidden layer has $n - 1$ nodes where n is the number of features in the dataset.

²<https://code.google.com/p/randomforest-matlab/>.

5.4 Feature Sets

To analyze the classification performance for different features, the image features are grouped into different groups such as intensity features, histogram, texture, region, graph, and shape-adaptive features. Feature extraction stage was done mostly using MATLAB programming language. However, in a small portion of the implementation, C# was also used.

5.4.1 Intensity Features

Features related to intensity distribution in an image can provide a basic feature set to categorize images into different categories. In general, the images consisting of crystals have high illumination compared to the images without crystals. Using the grayscale image I_{gray} , the 6 image intensity features (average image intensity, minimum image intensity, maximum image intensity, standard deviation of intensity, Otsu's threshold intensity, and threshold effectiveness metric) listed in Table 5.2 are extracted.

5.4.2 Histogram Features

The intensity histogram of an image provides a graphical representation of the image intensity distribution. The histogram provides information about the distribution of all pixel values or group of values in the image. For the fluorescence-based images, the green color channel carries the most information. Therefore, the intensity values in this channel are used to compute the histogram features. The number of bins was determined as 256 (between 0 and 255) for each green channel level. Histogram for the green level is defined as

Table 5.2 List of intensity features

Symbol	Description	Formulation
i_μ	Average image intensity	$\frac{1}{w*h} \sum_{i=1}^h \sum_{j=1}^w I_{gray}(i, j)$
i_{min}	Minimum image intensity	$min_{1 \leq i \leq h, 1 \leq j \leq w} I_{gray}(i, j)$
i_{max}	Maximum image intensity	$max_{1 \leq i \leq h, 1 \leq j \leq w} I_{gray}(i, j)$
σ	Standard deviation of intensity	$\sigma = \sqrt{\frac{1}{h*w} \sum_{i=1}^h \sum_{j=1}^w (i_\mu - I_{gray}(i, j))^2}$
τ_o	Otsu's threshold intensity	[28]
e_o	Threshold effectiveness metric	[23]

Table 5.3 List of histogram features

Symbol	Description	Formulation
μ	Average image intensity	$\frac{1}{w*h} \sum_{k=0}^{k=255} k * H[k]$
σ	Std devn of intensity	$\sqrt{\frac{1}{w*h} \sum_{k=0}^{k=255} (k - \mu)^2 * H[k]}$
s	Skewness	$\frac{1}{(w*h)*\sigma^{1.0}} \sum_{k=0}^{k=255} (k - \mu)^3 * H[k]$
k	Kurtosis	$\frac{1}{(w*h)*\sigma^2} \sum_{k=0}^{k=255} (k - \mu)^4 * H[k]$
vE	Entropy	$-\sum_{k=0}^{255} N[k] \log(N[k])$, where $N[k] = H[k]/(w * h)$
$g_1^1, g_1^2, g_1^3, \dots, g_3^3$	GLCM autocorrelation	Equation 5.5
ia_1, ia_2, ia_3	Image autocorrelation	Equation 5.6
mg_1, mg_2, mg_3	GLCM power spectrum magnitude	$mg_i = \text{mean2}(\text{fftshift}(\text{fft2}(P_i)))$, $1 \leq i \leq 3$
mi	Image power spectrum magnitude	$mi = \text{mean2}(\text{fftshift}(\text{fft2}(I)) ^2)$

$$H[k] = \sum_{p=1}^w \sum_{q=1}^h \begin{cases} 1 & \text{if } I_G(p, q) = k \\ 0 & \text{otherwise} \end{cases} \quad (5.3)$$

Green Level Co-occurrence Matrix (GLCM) is a matrix of distribution of co-occurring values of green level intensity at a given offset Δ_x, Δ_y [16]. GLCM matrix P using the green color channel is defined as in Eq. (5.4).

$$P_{\Delta_x, \Delta_y}(i, j) = \sum_{p=1}^{w-\Delta_x} \sum_{q=1}^{h-\Delta_y} \begin{cases} 1 & \text{if } I_G(p, q) = i \text{ and } I_G(p + \Delta_x, q + \Delta_y) = j \\ 0 & \text{otherwise} \end{cases} \quad (5.4)$$

With (Δ_x, Δ_y) as $(1, 0)$, $(0, 1)$, and $(1, 1)$, 3 GLCMs are obtained and represented as P_1, P_2 , and P_3 , respectively. Using green channel image I_G , intensity histogram H and GLCMs P_1, P_2 , and P_3 , the 21 image features listed in Table 5.3 are extracted. The average intensity, standard deviation, skewness, kurtosis, and entropy measure are the image features related to intensity distribution. GLCM autocorrelation is a measure of linear dependence between the elements of co-occurrence matrix with offset of Δ_m and Δ_n . The GLCM autocorrelation g_k with offset (Δ_m, Δ_n) using GLCM P_k is defined as in Eq. (5.5).

$$g_{k, \Delta_m, \Delta_n} = \frac{\sum_{i=\Delta_m}^{255} \sum_{j=\Delta_n}^{255} P_k(i, j) * P_k(i - \Delta_m, j - \Delta_n)}{\sum_{i=\Delta_m}^{255} \sum_{j=\Delta_n}^{255} \max(P_k(i, j), P_k(i - \Delta_m, j - \Delta_n))^2} \quad (5.5)$$

Using P_1, P_2 , and P_3 GLCMs, and (Δ_m, Δ_n) as $(1, 0)$, $(0, 1)$, and $(1, 1)$, $3*3 = 9$ GLCM autocorrelation features are obtained.

Image autocorrelation is defined as the measure of linear dependence between pixels of the image with offset of Δm and Δn and computed as in Eq. (5.6).

$$ac_{\Delta m, \Delta n} = \frac{\sum_{i=\Delta m}^{255} \sum_{j=\Delta n}^{255} I_G(i, j) * I_G(i - \Delta m, j - \Delta n)}{\sum_{i=\Delta m}^{255} \sum_{j=\Delta n}^{255} (I_G(i, j))^2} \quad (5.6)$$

Three image aut-correlation features using $(\Delta m, \Delta n)$ as $(1, 0)$, $(0, 1)$, and $(1, 1)$ are obtained. The green color channel of the image is used as the input. Similarly, the power spectrum is calculated using P_1, P_2, P_3 , and I , and the magnitude is used as the image feature.

5.4.3 Texture Features

A texture is a set of texture elements or texels occurring in some regular pattern. In this study, a total of 23 texture features are employed, collected from 3 different studies ([16], [42], [4]), and MATLAB built-in functions [23]. The list of features is provided in Table 5.4. Since 4 angular GLCM matrices are generated for texture analysis, 4 values are computed for each of 23 features in Table 5.4 leading to $4 * 23 = 92$ values. By taking the mean and the range of the 4 values per feature, the number of features is reduced to 46.

Let N_g denote the number of distinct green levels in the quantized image; $p(i, j)$ represent the (i, j) th entry in the normalized GLCM, $p_x(k)$ denote the k th entry of the matrix obtained by summing rows of $p(i, j)$, and $p_y(k)$ represent the k th entry of the matrix obtained by summing columns of $p(i, j)$. The following notation is used in the formulation of the features provided in Table 5.4.

- $p_{x+y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j) \mid i + j = k$
- $p_{x-y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j) \mid |i - j| = k$
- $\mu_x = \sum_i \sum_j i \cdot p(i, j)$
- $\mu_y = \sum_i \sum_j j \cdot p(i, j)$
- $\sigma_x = \sum_i \sum_j (i - \mu_x)^2 \cdot p(i, j)$
- $\sigma_y = \sum_i \sum_j (j - \mu_y)^2 \cdot p(i, j)$
- $HXY = - \sum_i \sum_j p(i, j) \log(p(i, j))$
- HX and HY are entropies of p_x and p_y
- $HXY1 = - \sum_i \sum_j p(i, j) \log\{p_x(i)p_y(i)\}$
- $HXY2 = - \sum_i \sum_j p_x(i)p_y(i) \log\{p_x(i)p_y(i)\}$

In Table 5.4, the MATLAB homogeneity feature (f_{10}) and inverse difference feature (f_{21}) are actually two different labels and implementations of the same feature. Although both features were extracted for the experiments, one of these features can be eliminated based on the programming environment.

Table 5.4 List of texture features

	Feature	Formulation
f_1	Autocorrelation [16]	$\sum_i \sum_j (ij)p(i, j)$
f_2	Contrast [16]	$\sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j) \mid i - j = n \right\}$
f_3	Correlation (MATLAB) [23]	$\sum_i \sum_j \frac{(i-\mu_x)(j-\mu_y)p(i, j)}{\sigma_x \sigma_y}$
f_4	Correlation [16]	$\sum_i \sum_j \frac{(ij)p(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y}$
f_5	Cluster prominence [42]	$\sum_i \sum_j (i + j - \mu_x - \mu_y)^4 p(i, j)$
f_6	Cluster shade [42]	$\sum_i \sum_j (i + j - \mu_x - \mu_y)^3 p(i, j)$
f_7	Dissimilarity [42]	$\sum_i \sum_j i - j \cdot p(i, j)$
f_8	Energy [16]	$\sum_i \sum_j p(i, j)^2$
f_9	Entropy [42]	$-\sum_i \sum_j p(i, j) \log(p(i, j))$
f_{10}	Homogeneity (MATLAB) [23]	$\sum_i \sum_j \frac{p(i, j)}{1 + i - j }$
f_{11}	Homogeneity [42]	$\sum_i \sum_j \frac{1}{1 + (i - j)^2} p(i, j)$
f_{12}	Maximum probability [42]	$MAX_{i, j} p(i, j)$
f_{13}	Sum of squares: Variance [16]	$\sum_i \sum_j (i - \mu)^2 p(i, j)$
f_{14}	Sum average [16]	$\sum_{i=2}^{2N_g} i p_{x+y}(i)$
f_{15}	Sum entropy [16]	$-\sum_{i=2}^{2N_g} p_{x+y}(i) \log\{p_{x+y}(i)\}$
f_{16}	Sum variance [16]	$\sum_{i=2}^{2N_g} (i - f_{15})^2 p_{x+y}(i)$
f_{17}	Difference variance [16]	$var(p_{x-y})$
f_{18}	Difference entropy [16]	$-\sum_{i=0}^{N_g-1} p_{x-y}(i) \log\{p_{x-y}(i)\}$
f_{19}	Information measure of correlation 1 [16]	$\frac{HXY - HXY1}{\max\{HX, HY\}}$
f_{20}	Information measure of correlation 2 [16]	$(1 - \exp[-2(HXY2 - HXY)])^{1/2}$
f_{21}	Inverse difference (INV) [4]	$\sum_i \sum_j \frac{p(i, j)}{1 + i - j }$
f_{22}	Inverse difference normalized [4]	$\sum_i \sum_j \frac{p(i, j)}{1 + i - j /N_g}$
f_{23}	Inverse difference moment [4]	$\sum_i \sum_j \frac{p(i, j)}{1 + ((i - j)/N_g)^2}$

5.4.4 Region Features

Image thresholding separates the foreground and background in the image. By thresholding the protein crystal images, crystals are expected to be distinguished as foreground objects. Although other non-crystal objects might also appear as the foreground, features from the binary images can provide important information about

Table 5.5 List of global binary image features

Symbol	Description	Formulation
N_f	Number of white pixels in B	$\sum_{x=1}^h \sum_{y=1}^w B(x, y)$
μ_f	Foreground average intensity	$\frac{1}{N_f} \sum_{i=1}^h \sum_{j=1}^w I_{gray}(i, j) \cdot B(i, j)$
σ_f	Foreground standard deviation intensity	$\sqrt{\frac{1}{N_f} \sum_{i=1}^h \sum_{j=1}^w ((\mu_f - I_{gray}(i, j)) \cdot B(i, j))^2}$
μ_b	Background average intensity	$\frac{1}{h*w-N_f} \sum_{i=1}^h \sum_{j=1}^w I_{gray}(i, j)(1 - B(i, j))$
σ_b	Background standard deviation intensity	$\sqrt{\frac{1}{h*w-N_f} \sum_{i=1}^h \sum_{j=1, B(i, j)=0}^w ((\mu_b - I_{gray}(i, j)) \cdot (1 - B(i, j)))^2}$
N	Number of blobs	No. of connected components
r_c	Image fullness	$convexHullArea(B)/(h * w)$

the content of an image. Similarly, features related to the shape and size of individual objects are useful to categorize the images into different categories.

Using the gray-level image I_{gray} and binary image B , the 7 global binary image features (the number of white pixels in B , foreground average intensity, standard deviation of foreground intensity, background average intensity, standard deviation of background intensity, number of blobs, and image fullness) listed in Table 5.5 are extracted. More information about the objects is obtained by extracting features related to intensity statistics and shapes of the individual blobs. Nine blob features (average intensity, standard deviation of intensity, number of pixels, number of white pixels, perimeter, convex hull area, blob eccentricity, blob extent, and equivalent circular diameter) are extracted for each of the top k largest blobs. Table 5.6 provides the list of 9 blob features. If the number of blobs n is less than k , the value 0 is used as the feature value for the blobs $O_{n+1}..O_k$. Since a single technique may not always provide correct binary image, 4 different image binarization methods (Otsu, G_{90} , G_{99} , and morphological thresholding) are applied. These images are used to extract region based image features. From each binary image, 52 ($7 + 5*9 = 52$) image features are obtained for the 5 largest blobs (i.e., $k = 5$). Region *Otsu*, Region G_{90} , Region G_{99} , and Region *Morph* represent the features obtained using Otsu, G_{90} , G_{99} , and morphological thresholding methods, respectively.

Table 5.6 List of blob features

Symbol	Description	Formulation
μ_o^i	Average intensity of O_i	$\frac{1}{w_i \times h_i} \sum_{j=m_x^i-w^i/2}^{m_x^i+w^i/2} \sum_{k=m_y^i-h^i/2}^{m_y^i+h^i/2} I_{gray}(j, k)$
σ_o^i	Standard deviation of intensity of O_i	$\sqrt{\frac{1}{w_i \times h_i} \sum_{j=m_x^i-w^i/2}^{m_x^i+w^i/2} \sum_{k=m_y^i-h^i/2}^{m_y^i+h^i/2} (\mu_o^i - I_{gray}(j, k))^2}$
N_o^i	Number of pixels in O_i	$h_i * w_i$
Nf_o^i	Number of white pixels in O_i	$\sum_{x=1}^{h_i} \sum_{y=1}^{w_i} O_i(x, y)$
p_o^i	Perimeter of O_i	$\sum_{i=1}^{h_i} \sum_{j=1}^{w_i} \Omega_i(x, y)$
ch_o^i	Convex hull area of O_i	[23]
e_o^i	Blob eccentricity of O_i	[23]
be_o^i	Blob extent of O_i	[23]
bd_o^i	Equivalent circular diameter of O_i	[23]

5.4.5 Graph Features

The structure of an object as a graph has a significant importance in image analysis since it defines the boundaries of an object in the image. Edge detection followed by some post-processing steps to extract features that are useful to define the shapes of objects [39] is applied. In addition, Hough line transform is applied to extract line features. Table 5.7 provides graph-related features.

In Table 5.7, S is the set of graphs in I , S_i the i th graph in S , L is the set of edges in I , $|L(S_i)|$ is the number of edges in graph S_i , and $\alpha(l_i, l_j)$ represents the angle between l_i and l_j .

5.4.6 Shape-Adaptive Features

Shape-adaptive Discrete Cosine Transform (SA-DCT) is a 2D Discrete Cosine Transform (DCT) method for coding arbitrarily shaped image segments [50]. Image coding can be applied either to region of interest (blobs) or the background region. In this study, SA-DCT is applied on the top largest blobs. Table 5.8 provides the list of image features extracted from each blob after applying the SA-DCT. Otsu's thresholding is applied to obtain the binary image. SA-DCT is then applied on top 5 largest blobs. Thus, 15 DCT features are obtained from an image. If a binary image contains less than 5 blobs, 0 is assigned to all feature values of missing blobs.

Table 5.7 Graph features

Feature	Symbol	Description	Formulation
Edge [39]	η	Number of graphs (connected edges)	$\eta = S $
	η_1	Number of graphs with a single edge	$\eta_1 = S_i $, where $ L(S_i) = 1$
	η_2	Number of graphs with 2 edges	$\eta_2 = S_i $, where $ L(S_i) = 2$
	η_c	Number of graphs whose edges form a cycle	$\eta_c = S_i $, where S_i is a cyclic graph
	η_p	Number of line normals	$\eta_p = \sum \perp(S_k), \perp(S_k) = \begin{cases} 1 & \exists l_i \in L_k \text{ and } \exists l_j \in L_k \text{ and} \\ & 70 \leq \alpha(l_i, l_j) \leq 90 \\ 0 & \text{otherwise} \end{cases}$
	μ_l	Average length of edges in all segments	$\mu_l = \frac{\sum_{i \in \mathbf{L}} l_i}{ \mathbf{L} }$
	S_l	Sum of lengths of all edges	$S_l = \sum_{i \in \mathbf{L}} l_i$
	l_{max}	Maximum length of an edge	$l_{max} = \max_{1 \leq i \leq \mathbf{L} } (l_i)$
	c_o	1 if $\eta_c > 0$, 0 otherwise	$c_o = \exists S, S$ is a cyclic graph
	l_o	1 if $\eta_p > 0$, 0 otherwise	$l_o = (\exists l_i \in L_k \text{ and } \exists l_j \in L_k \text{ and } 70 \leq \alpha(l_i, l_j) \leq 90)$
	η_{hc}	Number of Harris corners	[17]
Hough	η_{hl}	Number of Hough lines	[18]
	μ_{hl}	Average length of Hough lines	[18]

Table 5.8 Shape-adaptive DCT features

Symbol	Description
C_m^i	Maximum of nonzero coefficients of SA-DCT of O_i
C_μ^i	Average of nonzero coefficients of SA-DCT of O_i
C_N^i	Number of nonzero coefficients of SA-DCT of O_i

5.5 Analysis of Feature Sets

There are a number of difficulties for classifying crystallization trial images as mentioned in the introduction. First, there are many categories (9 categories according to Hampton's scale) to classify with high intra-class diversity. As the number of categories increases, developing a reliable classification model becomes more difficult. Second, labeling the data is difficult due to the temporal transition between categories and the presence of multiple types of crystals in images. Third, the low percentage of representation of critical categories gives bias to more populated but less impor-

tant categories. To overcome these problems, a 2-stage classification was considered that divides the classification problem into 3-class classification (non-crystals, likely-leads, and crystals) at the first level, and classification into subcategories in the second level as shown in Fig. 2.5. To balance the data distribution, all available data from critical categories was used while reducing the images from frequently occurring image categories. For time analysis, the time to extract each feature set was computed. The classification results based on overall accuracy and sensitivity of critical categories were ranked. Fivefold and tenfold cross-validation was used for measuring the accuracy in different tests. Accuracy measures along with time analysis for classification help to select the best feature sets for real-time stand-alone computing system.

In this study, the experiments are designed in an exhaustive manner to be able to evaluate effectiveness of different factors for classification of protein crystal images. Different feature sets, classifiers, normalization, and feature reduction techniques are considered. Experiments are carried out for all possible cases, and the performance is calculated for each case. The goal is to determine the best condition (feature set/classifier/transformation tuple) that can yield the highest accuracy on protein crystallization images. The selection of features for hierarchical classification is provided in Sect. 5.5.2. The results with respect to the time complexity and real-time processing are evaluated. A total of **8624** experiments are carried out to test 9 major objectives, listed in Table 5.9. According to the table, Exp. IDs from “1” through “4” represent the first-level experiments that are described in Sect. 5.5.3, and Exp. IDs from “5” to “7” describe the second-level experiments explained in Sect. 5.5.4. In addition, Exp. IDs “8” and “9” correspond to timing calculation of the experiments explained in Sect. 5.6.

5.5.1 Data

The images are collected using the Crystal X2 by iXpressGenes, Inc. This is a fluorescence-based microscopy system for scanning protein crystallization screening trial plates. All the images are hand scored by an expert according to Hampton’s scale. Table 5.10 provides the distribution of the dataset into different categories. The dataset includes a total of 2,756 images composed of 1,600 non-crystal images, 675 likely-lead images, and 481 crystal images. The image resolution is 320×240 reduced from the camera resolution of $2,560 \times 1,920$. Some images were difficult to assign a subcategory due to blurriness, illumination problems, significant high intensity in the image, and presence of crystals at different phases. Because of this, *doubtful* subcategory is added to each category, and the images with ambiguous subcategory were assigned to these *doubtful* subcategories. Doubtful images are used for training at the first level, but these images are discarded while building a training model for subcategory classification.

Table 5.9 List of classification experiments

Exp ID	Tasks	No. of experiments ^a
1	Run all classifiers for 511 feature set (5 classifiers with/without normalization)	$2 * 5 * 511 * 1 = 5110$
2	Run the best classifier 5 times and take the average for the best 70-feature set (RF)	$1 * 1 * 64 * 5 = 320$
3	Run classifiers PCA for 10,20, ...,50 features	$1 * 5 * 5 * 2 = 50$
4	Run classifiers using RF feature selection (10,20,...,50)	$1 * 5 * 5 * 2 = 50$
5	Run BYS, DT, and RF (with and without normalization, with graph features) for crystal subcategories	$2 * 3 * 511 * 1 = 3066$
6	Run RF, DT, and BYS classifiers with and without normalization for likely-lead subcategories	$2 * 3 * 1 * 1 = 6$
7	Run RF, DT and BYS classifiers with and without normalization for non-crystal subcategories	$2 * 3 * 1 * 1 = 6$
8	Calculate training and testing time of the random forest for the largest feature	$1 * 1 * 1 * 5 = 5$
9	Calculate timings for feature extraction of an image	$1 * 1 * 11 * 1 = 11$
	Total number of experiments	8624

^a In the table, the notation to calculate the number of experiments for a task is $\eta_n * \eta_c * \eta_f * \eta_r$. In this notation, η_n refers to the number of normalizations that are applied to feature set, η_c refers to the number of classifiers used, η_f refers to the number of feature sets that are used for the corresponding experiments, and η_r is the number of repetition of the experiments

Table 5.10 Dataset image distribution

Category	Total images	Subcategory	No. of images	Percentage (%)
Non-crystals	1600	Clear drop	1273	46.19
		Phase separation	1	0.04
		Precipitate	204	7.4
		Doubtful	122	4.43
Likely-leads	675	Microcrystals	122	4.43
		Unclear bright images	369	13.39
		Doubtful	184	6.68
Crystals	481	Dendrites/Spherulites	63	2.29
		Needles	153	5.55
		2D Plates	8	0.29
		Small 3D crystals	129	4.68
		Large 3D crystals	35	1.27
		Doubtful	93	3.37
		Total	2756	

5.5.2 *Evaluating Features for Hierarchical Classification*

Initially, the main goal of the experiments was to classify protein crystallization trial images into the categories of the first level. Analyzing pixel intensities was generally enough for the first-level classification. Once good results with the first level are obtained, subcategory classification is applied for each category of the first level. Ideally, it would be good if the feature set that works great for the first level also works best for the second level. The experiments for the second level are not restricted by the optimal feature set of the first level. For further subcategory classification, the same feature set has been tested first. If the same feature set provides reasonable performance, there would be no need to extract any more features. However, if the accuracy of subcategory classification is not satisfactory, all combinations of feature sets are run for the subcategory as well.

Intensity (Table 5.2), histogram (Table 5.3), texture (Table 5.4), region (Tables 5.5 and 5.6), Hough (Table 5.7), and shape-adaptive (Table 5.8) features are used for the first-level classification. The boundaries of crystal regions may actually be critical to identify crystals. Using "Hough" features did not provide satisfactory results for crystal subclassification. Adding edge features (Table 5.7) in addition to Hough features may improve the classification accuracy. The main factor for adding this additional set is the diverse set of images in crystal categories (Panels C, D, and E in Fig. 2.2 and Panels A–E in Fig. 2.3): dendrites/spherulites, needles, plates, small 3D, and large 3D crystals. Later it was observed that graph features (Table 5.7) turned out to be important for crystal subclassification.

5.5.3 *First-Level (3-Class) Classification*

For the first level of classification, 5110 experiments are run for all possible feature sets with and without normalization on 5 different classifiers (Exp. ID 1 in Table 5.9). There are 9 different feature sets as mentioned above. Based on those features, $2^9 - 1 = 511$ different combinations of feature sets were generated for the first-level classification. For the first-level classification, only Hough features of the graph feature set in Table 5.7 were utilized rather than the complete graph feature set. After analyzing the results of Exp. ID 1, the best 64 feature sets were selected that provided the highest accuracy. Using the selected feature sets, the experiments were rerun 5 times and the average was taken to ensure that the results are consistent (Exp. ID 2 in Table 5.9). In addition to these experiments, the effects of feature reduction and selection methods on the classification performance were investigated. PCA was applied to the complete feature set (excluding 11 edge features which are added later for crystal subcategories in Table 5.7) by reducing from 298 features to 5 feature subsets (10, 20, 30, 40, and 50 features). Later, the 50 experiments (Exp. ID 3 in Table 5.9) are run. Similarly, random forest feature selection algorithm was applied to reduce the features (10, 20, 30, 40, and 50 features) (Exp. ID 4 in Table 5.9) similar

to the PCA experiments. Then 50 new experiments were run for new feature sets. Totally 5530 experiments were carried out for the first level of classification.

Accuracy Measures. To evaluate the correctness of the classification four measures: accuracy, probabilistic accuracy (Pacc) [37], sensitivity, and adjusted sensitivity were evaluated. Let matrix C represent the $N \times N$ confusion matrix for an N -class problem. The value C_{ij} refers to the number of items of class i predicted as class j . For the first-level (3-class) classification, adjusted sensitivity is calculated as in (5.7).

$$\text{adjusted sensitivity} = \frac{\sum_{i=2}^{i=3} C_{2i} + C_{3i}}{\sum_{i=1}^{i=3} C_{2i} + C_{3i}} \quad (5.7)$$

Here, classes 1, 2, and 3 represent non-crystals, likely-leads, and crystal categories, respectively. The adjusted sensitivity does not penalize if crystals are classified as likely-leads since experts analyze the likely-lead category as well.

Best Performing Feature Sets. Table 5.11 shows the best 10 results of 5110 experiments in Exp. ID 1 in descending order with respect to the accuracy measure. Here, the highest accuracy result (96.3%) is achieved by applying random forest classifier on the following normalized feature sets: intensity features, region features using Otsu, region features using G_{99} , and histogram features. As can be seen in the table, the other results are also satisfactory as much as the first one. Note that the DCT features require significant extraction time and provide very little or no contribution to the overall classification performance. Therefore, in the second level of classification, DCT features are excluded from the experiments.

Re-evaluating the Best Results. After conducting 5110 experiments, the best 64 feature sets were selected to validate the consistency of their high performance. Then, these particular experiments were repeated for these 64 feature sets 5 times and their average performance was calculated. In Table 5.12, the feature sets along with the accuracies of the best 8 (out of 64) experiments are provided. The set of intensity features, region features using Otsu, region features using G_{90} , region features using G_{99} , and histogram features gave the best accuracy (96.1%) using random forest classifier. According to the time analysis, the best feature set can be extracted in 1.080 s. This is not the lowest time in the table, but it is a reasonable time for real-time applications.

Feature Reduction using PCA. Feature reduction was also considered to determine its effect on the classification performance. First, the size of complete feature set is reduced using PCA. Five new feature sets (10, 20, 30, 40, and 50 features) were generated that include the most representative ones in the new feature space. For each feature set, the experiments were evaluated using all classifiers with and without normalization (Exp. ID 3 in Table 5.9). The accuracy measures were calculated and the results are provided in Table 5.13. The highest accuracy can be reached using 30 or 20 features (with PCA transformation) using random forest classifier after applying normalization. The change in principal component variances with respect to the number of features is shown in Fig. 5.2. By analyzing Table 5.13, it can be inferred that the number of features can be reduced to 20 with a small loss of accuracy (around

Table 5.11 Classification results for preliminary experiment using random forest classifier (Experiment ID 1)

Feature set	Norm.	Acc	Pacc	Sensitivity	Adjusted sensitivity
Intensity, Region Otsu, Region G_{99} , Histogram	Yes	0.963	0.942	0.867	1
Intensity, Region Otsu, Region G_{99} , Region Morph, Histogram, DCT	No	0.963	0.942	0.871	1
Intensity, Region Otsu, Region G_{99} , Hough, Texture, Histogram, DCT	Yes	0.963	0.941	0.863	1
Intensity, Region Otsu, Region G_{99} , Histogram	No	0.962	0.94	0.881	1
Intensity, Region Otsu, Region G_{90} , Region G_{99} , Region Morph, Hough, Histogram, DCT	No	0.962	0.94	0.867	1
Intensity, Region Otsu, Region G_{90} , Region G_{99} , Region Morph, Texture, Histogram	Yes	0.962	0.939	0.865	1
Intensity, Region Otsu, Region G_{99} , Hough, Histogram, DCT	Yes	0.962	0.939	0.871	1
Intensity, Region Otsu, Region G_{99} , Hough, Histogram, DCT	No	0.962	0.939	0.869	1
Intensity, Region G_{99} , Hough, Texture, Histogram	Yes	0.962	0.938	0.861	1
Intensity, Region Otsu, Region G_{90} , Region G_{99} , Region Morph, Histogram	No	0.962	0.938	0.861	1

3% lower than the best case in Table 5.11). However, the sensitivity is almost 0.13 lower than the best sensitivity.

Feature Selection using Random Forest MDA. Similar to the feature reduction, the effects of feature selection was also considered in the experiments. To select more reliable features, MDA (mean decrease in accuracy) algorithm in random forest was preferred. Five feature sets (having 10, 20, 30, 40, and 50 representative features) were generated. For each feature set, the experiments were evaluated using all classifiers with and without normalization (Exp. ID 4 in Table 5.9). Similar to the PCA reduction results in Table 5.13, four accuracy measures were calculated. The results were reported in Table 5.14. Best results were achieved using 30 features with random forest classifier after normalizing the dataset. The comparison of the best results in Tables 5.13 and 5.14 shows that feature selection provides better accuracy than feature reduction in the experiments.

Performance of Individual Feature Sets. Finally, the power of the individual feature sets was investigated. The performance of each feature set was evaluated using all classifiers with and without normalization. Table 5.15 shows the best results for each

Table 5.12 Classification results for the best 8 of 64 experiments using random forest classifier

Feature set	Norm.	Acc	Pacc	Sensitivity	Adjusted sensitivity	Time per image (sec)
Intensity, Region Otsu, Region G_{90} , Region G_{99} , Histogram	No	0.961	0.938	0.87	1	1.08
Intensity, Region Otsu, Region G_{90} , Region G_{99} , Region Morph, Texture, Histogram	No	0.96	0.935	0.857	1	1.31
Region Otsu, Region G_{90} , Region G_{99} , Histogram	Yes	0.959	0.935	0.861	1	1.028
Region Otsu, Region G_{90} , Region G_{99} , Histogram, DCT	No	0.959	0.934	0.852	1	26.668
Region Otsu, Region G_{99} , Histogram	Yes	0.959	0.934	0.858	1	0.77
Region Otsu, Region G_{90} , Region G_{99} , Histogram	No	0.959	0.934	0.859	1	1.028
Intensity, Region Otsu, Region G_{90} , Region G_{99} , Texture, Histogram, DCT	No	0.958	0.934	0.854	1	26.756
Region Otsu, Region G_{99} , Histogram, DCT	No	0.957	0.931	0.853	1	26.409

Table 5.13 Classification results after feature reduction by PCA

Classifier	# Features	Norm	Acc	Pacc	Sensitivity	Adjusted sensitivity
RF	30	Yes	0.934	0.901	0.740	0.954
RF	20	Yes	0.934	0.905	0.744	0.944
RF	40	Yes	0.931	0.897	0.728	0.948
RF	50	Yes	0.930	0.896	0.719	0.950
RF	50	No	0.928	0.893	0.715	0.940
SVM	50	Yes	0.918	0.870	0.761	0.990
SVM	40	Yes	0.916	0.869	0.763	0.983
SVM	30	Yes	0.910	0.858	0.726	0.985
RF	40	No	0.909	0.880	0.688	0.861
SVM	50	No	0.909	0.858	0.765	0.983

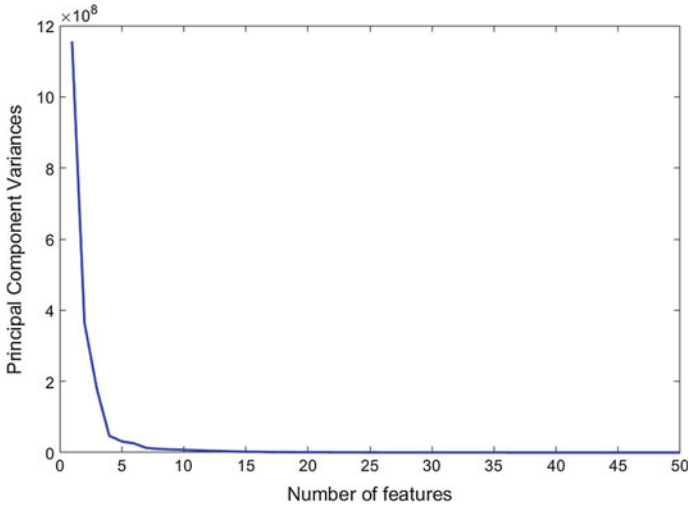


Fig. 5.2 Principal component variances of the best 50 features

Table 5.14 Classification results after feature selection by Random Forest

Classifier	# Features	Norm	Acc	Pacc	Sensitivity	Adjusted sensitivity
RF	30	Yes	0.960	0.936	0.863	0.998
RF	40	No	0.958	0.933	0.852	0.994
RF	50	Yes	0.957	0.932	0.859	0.996
RF	50	No	0.956	0.930	0.859	0.996
RF	30	No	0.954	0.926	0.834	0.994
RF	30	Yes	0.952	0.925	0.817	0.994
RF	20	No	0.950	0.920	0.832	0.992
RF	20	Yes	0.946	0.915	0.817	0.996
SVM	30	Yes	0.938	0.901	0.854	0.996
SVM	50	Yes	0.934	0.895	0.844	0.996

feature set. Additional experiments for these results were not performed since Exp. ID 1 already includes these cases. The best results are obtained using the histogram feature sets with accuracy of 90.8%.

5.5.4 Second-Level Classification

Evaluating the second-level classification independently helps analyze and improve the subcategory classification by ignoring the misclassification from the first level.

Table 5.15 Classification performance with individual feature sets

Feature set	Classifier	Norm	Acc	Pacc	Sensitivity	Adjusted sensitivity
Intensity	ID3	No	0.877	0.836	0.701	0.950
Region Otsu	BYS	Yes	0.751	0.702	0.622	0.915
Region G_{90}	SVM	Yes	0.864	0.818	0.676	0.944
Region G_{99}	SVM	Yes	0.882	0.838	0.723	0.944
Region Morph	BYS	Yes	0.738	0.717	0.580	0.994
Hough	SVM	Yes	0.841	0.737	0.235	0.906
Texture	ID3	Yes	0.822	0.778	0.605	0.877
DCT	BYS	Yes	0.691	0.647	0.480	0.775
Histogram	SVM	Yes	0.908	0.852	0.705	0.996

If the classification accuracy of the first level was low, this could have been risky. However, the first-level accuracy is 96%, which is reasonably high. In the first-level classification, protein crystallization trial images are classified into 3 categories: non-crystals, likely-leads, and crystals. In the second level, each of these categories are further classified into subcategories as shown in Fig. 2.5. For the first level, the feature set composed of intensity features, region features using Otsu, region features using G_{90} , region features using G_{99} , and histogram features (First row in Table 5.12) provided the best result was determined. The sensitivity for the highest ranked category in each subcategory is provided. The highest ranked category is precipitates for non-crystals, microcrystals for likely-leads, and large 3D crystals for the crystals. For two-class classification, if both accuracy and sensitivity are available along with the number of samples in each category, the other sensitivity value could be computed easily.

Non-crystal classification Non-crystals are classified into 3 subcategories: clear drops, phase separation, and precipitates. Phase separation is a relatively rare occurrence. Table 5.16 provides the classification performance for 3 classifiers (Exp. ID 7 in Table 5.9) with and without normalization. These experiments are conducted on the best feature set combination for the first-level classification. Normalization is done using z-score normalization. The sensitivity column refers to the sensitivity for precipitates. Random forest provided the best classification performance and normalization did not make any major difference. The classification accuracy is 98% and the sensitivity for precipitates category is 0.91.

Likely-lead classification In the likely-lead category, there are two subcategories: unclear bright images and microcrystals. The classification performance with 3 classifiers (Naïve Bayes, decision tree, and random forest) is provided in Table 5.17 (Exp. ID 6 in Table 5.9). These experiments are again conducted on the best feature set combination for the first-level classification. The sensitivity column refers to the sensitivity for microcrystals. The best performance (92% accuracy) is obtained

Table 5.16 Non-crystal subclassification

Classifier	Normalization	Accuracy	Pacc	Sensitivity
Naïve Bayes	No	0.88	0.71	0.59
Naïve Bayes	Yes	0.88	0.72	0.68
Decision Tree	No	0.96	0.79	0.85
Decision Tree	Yes	0.96	0.79	0.85
Random Forest	No	0.98	0.81	0.91
Random Forest	Yes	0.98	0.81	0.91

Table 5.17 Likely-lead subclassification

Classifier	Normalization	Accuracy	Pacc	Sensitivity
Naïve Bayes	No	0.59	0.62	0.86
Naïve Bayes	Yes	0.58	0.63	0.93
Decision Tree	No	0.87	0.85	0.74
Decision Tree	Yes	0.88	0.86	0.76
Random Forest	No	0.92	0.91	0.80
Random Forest	Yes	0.91	0.89	0.78

using random forest classifier without normalization. The corresponding sensitivity for microcrystals is 0.80.

Crystal subclassification In the crystal category, there are 5 subcategories: dendrites/spherulites, needles, 2D plates, small 3D crystals, and large 3D crystals. Crystals have geometric shapes that can be defined by edges. Therefore, edge-related features are quite useful to distinguish the crystal subcategories. For crystal subclassification, rather than using only Hough features of the graph feature set, the edge features in Table 5.7 were also included in the experiments to consider the diverse crystal categories. In addition to the selected features useful for the first-level classification and non-crystal and likely-lead classification, classification experiments were performed (Exp. ID 5 in Table 5.9) including graph features described in Sect. 5.4. Table 5.18 shows the top 7 classification performances based on the accuracy using random forest classifier. The sensitivity column refers to the sensitivity of large 3D crystals. The feature set of intensity, region features using Otsu's thresholding, region features using G_{90} , graph and histograms gave the highest accuracy of 74.2%. This feature set can be extracted in 1.267 s. Alternatively, with slightly lower accuracy (74%), the feature set of region using Otsu's thresholding, region using G_{99} , graph and histogram features can be generated in less than a second. The fastest feature set (region features using G_{90} and graph) with accuracy of 73.5% can be generated in 0.779 s.

Table 5.18 Crystal subclassification

Feature set	Norm	Accuracy	Pacc	Sensitivity	Time (s)
Intensity, Region Otsu, Region G_{90} , Graph, Histogram	Yes	0.742	0.667	0.909	1.267
Region Otsu, Region G_{99} , Graph, Texture, Histogram	Yes	0.74	0.684	0.896	0.949
Region Otsu, Region G_{90} , Region G_{99} , Graph, Histogram	Yes	0.737	0.658	0.896	1.408
Region G_{90} , Graph	No	0.735	0.659	0.902	0.779
Intensity, R_ G_{90} , R_ G_{99} , Graph, Histogram	No	0.735	0.667	0.896	1.201
Intensity, R_Otsu, R_ G_{90} , Graph, Histogram	No	0.735	0.657	0.89	1.267
Intensity, Region Otsu, Region G_{99} , Graph, Histogram	No	0.735	0.682	0.878	0.964

5.6 Timing Analysis for Classification

Feature extraction was run on a system with Intel Core i7 2.4 GHz CPU, and 12 GB RAM memory. The image feature extraction routines are implemented using MATLAB 2013b. Some feature extraction modules were implemented using C# on Visual Studio 2012. Classification of data was accomplished using MATLAB. Table 5.19 provides a summary of feature extraction timings for different feature sets. Most of the features can be extracted in less than half a second. The set of DCT features is the most computationally expensive feature set, since it took around 25.5 s to extract DCT features on an average per image. This may be due to inefficient shape-adaptive DCT implementation. However, it is still used in the experiments to observe its benefit to the accuracy of the classification. Texture and intensity features can be extracted quite fast in about 0.037s and 0.052s, respectively.

In the timing analysis, the total time is calculated using individual extraction times in Table 5.19, when a combination of feature sets is selected. For example, if the feature set combination involves intensity, region G_{90} , and texture features, the total time to extract these feature sets combination is computed as $0.052 + 0.258 + 0.037 = 0.347s$.

The computation time must include the time to classify the provided feature sets. The time to classify is based on the random forest classifier as it provided better accuracy than other classifiers (to be explained later in the following subsection). The random forest classifier also provides an upper-bound for classification time as it is more complicated than other compared classifiers in terms of evaluation due to the number of decision trees involved. Random forest takes roughly 0.361 s to test all features, which is less than a half second for the complete set. If the feature set composed of intensity, region G_{90} , and texture features is classified using random forest classifier, the time to extract features and classify is computed as

Table 5.19 Computation time for feature extraction

Feature group	Description	No of features	Avg time per feature	Avg time per image
Intensity	Intensity features	6	0.009	0.052
Region Otsu	Region features using Otsu	52	0.005	0.258
Region G_{90}	Region features using G_{90}	52	0.010	0.495
Region G_{99}	Region features using G_{99}	52	0.004	0.193
Region Morph	Region features using morph thresh	52	0.006	0.311
Graph	Hough features and edge features	13	0.022	0.284
Hough	Hough features only	2	0.049	0.097
Texture	Texture features	46	0.001	0.037
Histogram	Histogram features	21	0.009	0.178
DCT	DCT features	15	1.709	25.639

$0.347 + 0.361 = 0.708s$. For the hierarchical classification, new features may need to be extracted for the other levels, and again a classifier needs to be applied for these levels. Hence, the timings for other levels should be added as well.

Real-time applications have deadlines to complete specific tasks. Reduction of features is essential for building real-time computing systems. The Crystal X2 microscopy system was used to collect the images of protein crystallization experiments benefiting from trace fluorescent labeling. Trace fluorescent labeling [31] helps reduce the number of features significantly with respect to systems using white light. Moreover, since trace fluorescent labeling yields high contrast between crystal regions and the background in trial images, image processing can be done in a simple and fast manner. The time to extract features from images and classify them can be reduced significantly. The time between capturing two images of a crystallization well plate using Crystal X2 is around 3 s. To be able to execute image acquisition and classification in parallel, the feature extraction and classification should be less than the transition time. However, there is a trade-off to consider between the best classification performance and minimum time for feature extraction. While extracting less features may be desirable, it may reduce the classification performance. In the discussions below, the first-level classification and crystal subcategory classification for the second level of the classification are provided since the accuracy of crystal classification is more important than other subcategories.

The Best Feature Sets. Using all features provided almost the same accuracy for the first level as the best feature sets. The best classification performance for the first-level (3-class) classification had 96% accuracy and 0.87 sensitivity using region features from Otsu's, G_{90} , and G_{99} thresholding, intensity, and histogram features. The feature extraction can be completed in 1.08 s for this feature set. Deep CNN [49] achieved 96.56% accuracy for binary classification by missing around 16% of crystals for their dataset. Since the accuracy of the first-level classification is high (around 96%), the misclassification at the first level should not have a significant effect on the second level. The system does not misclassify a crystal as non-crystal at the first level (i.e., the adjusted sensitivity is 1). The best classification performance for crystal subcategories at the second level had 74.2% accuracy and 0.909 sensitivity using normalized intensity, histogram, graph features and region features from Otsu's and G_{90} thresholding. This set of features can be extracted in 1.267s. On the other hand, by using all features, 69.6% accuracy with 0.618 sensitivity for crystal subcategory classification is obtained. Using all features reduced the accuracy and (more importantly) sensitivity significantly for the second level. The sensitivity of classification using all features for crystal subcategories is unacceptably low.

Fast Feature Sets. The fastest feature extraction with the same accuracy for the first level uses normalized histogram features and region features from Otsu's and G_{99} thresholding. This feature set can be extracted in 0.77s. The sensitivity of this feature (0.86) is slightly less than the sensitivity of the best feature set (0.87). Since the classification performance of the fast feature set is close to the performance of the best feature set, this set of features can be preferred to reduce the time for classification. For the crystal subcategory classification, the fastest feature set that can be extracted with high accuracy has only region features from G_{90} and graph features. This smaller feature set has provided 73.5% accuracy and 0.902 sensitivity compared to 74.2% accuracy and 0.909 sensitivity of the best feature set.

Comparison of Feature Sets for Hierarchical Classification. If two levels of classification are run in a hierarchical way, the union of the best feature sets includes intensity, graph, histogram features, and region features from Otsu's, G_{90} , and G_{99} thresholding. In other words, only graph features are added for the second level of classification. The total time for feature extraction increases slightly from 1.08 s to 1.373 s. Note that the time to extract the best feature set was 1.267 s for the second-level classification. If the fast feature sets from both levels are included, the union of feature sets includes histogram, graph features, and region features from Otsu's, G_{90} and G_{99} thresholding. For the fast feature sets, the intersection for the first and second levels is empty. The total time to extract features becomes 1.549s. Using fast feature sets for each level did not improve the overall time at all. The union of the best feature sets can be executed faster for the combination of two levels. If the classifier model is run in a hierarchical way, the overall performance in terms of time should be analyzed with respect to the common features between levels.

5.7 Deep Learning for Protein Crystallization Images

In recent years, deep learning had many successful applications from image recognition to natural language understanding. In Chap. 3, the application of neural networks on protein crystallization screening is explained. Unlike traditional neural networks, convolutional neural networks (CNNs) supports local receptive fields, sparse connectivity, and shared weights. These properties of CNNs allow to learn particular segments of the input, spatial local correlations, and detect same features at different locations of the input. Moreover, layers such as pooling and normalization may be added to support translation invariance and identify local features. In a deep CNN, there are a number of convolution layers that could be followed by pooling and normalization layers. At the end, these nodes are connected to one or more fully connected layers.

CrystalNet is designed for detecting protein crystals and composed of 13 layers [49]. Deep learning methods require a large training dataset. After applying contrast normalization in the first layer, the training set is increased by horizontal mirroring in the second layer and then applying small 2D similarity transformations in the third layer. These stages are followed by a sequence of three convolution and pooling layers. Then these outputs are fed into another convolution layer and fed into two fully connected layers. Their architecture is composed of a sequence of the following layers: image (128×128), contrast normalization (128×128) horizontal mirroring (128×128), transformation (112×112), convolution1 ($64, 112 \times 112, 5$), pooling1 ($64, 56 \times 56, 2, \max$), convolution2 ($64, 56 \times 56, 5$), pooling2 ($64, 28 \times 28, 2, \max$), convolution3 ($128, 28 \times 28, 5$), pooling2 ($128, 14 \times 14, 2, \max$), convolution4 ($128, 6 \times 6, 3$), fully connected layer1 (2048), fully connected layer2 (2048), and output (10). For image, contrast normalization, horizontal mirroring, and transformation layers the numbers in paranthesis indicate the size of output of that layer. For example, transformation layer reduces the size from 128×128 to 112×112 . *Convolution*($m, w * h, f$) indicates that m number of filters or maps generate output of $w \times h$ with a filter size of $f \times f$. For example, *convolution1*($64, 112 \times 112, 5$) generates the same size as the input. This also means that it uses a stride of 1. *Pooling*($m, w * h, f, \max$) indicates that there are m input maps and it generates output of $w \times h$ by applying $f \times f$ max pooling filter. 2×2 max pooling will reduce the size by 2 in each dimension. Each fully connected layer has 2,048 inputs from the previous layers. The output is designed for 10-way classification in their architecture. In the proposed design, element-wise nonlinearity is applied after convolutional and matrix multiplication of fully connected layers to satisfy nonlinearity of the classifier.

In CrystalNet, 150,000 parameters are learned with a learning rate parameter as 0.01, momentum as 0.9, and L_2 regularization constant as 0.0001. Initially, all weights are assigned randomly by using Gaussian distribution with 0 mean and 0.01 standard deviation. Their training set had 68,155 images whereas the validating set had 17,033 images using 80–20 split. Ten classes and the number images in each class (in paranthesis) in the validating set are clear (5877), precipitate (5732), crystal (1391), phase (1121), precipitate and crystal (1339), precipitate and skin (873), phase

and crystal (393), phase and precipitate (83), skin (136), and junk (88). CrystalNet achieved 90.8% accuracy with 0.7685 recall for the crystal category. Despite high accuracy, relatively low recall value for crystal category indicates enhancements needed for this deep CNN architecture.

5.8 Discussion

Accuracy for Hierarchical Classification using the Best Feature Sets. The accuracy of hierarchical classification is computed using the best feature set by applying the random forest classifier. Since fivefold cross-validation is used for evaluation, the training samples used for the second level are also used in the training set of the first level. Similarly, the same case applies for the test set. Such selection limits the selection of training set for the first level. Doubtful images for subcategories are used in training of the first level but not used for the second level. These new experiments in a retrospective way and there could be some slight differences in datasets and their categorization. Hence, the confusion matrices for these cascaded classifications are provided to avoid confusion. Based on the experiments, the accuracies of the first level and second level are 95.46% and 92.79%, respectively. The overall accuracy of the hierarchical classification is 89.22%. The confusion matrix of both levels is provided in Table 5.20. The confusion matrix for the first level is provided in Table 5.21. The confusion matrices for non-crystals, likely-leads, and crystals are provided in Tables 5.22, 5.23, and 5.24, respectively. In the confusion matrices of the second level, “*” indicates incorrect classification samples in the first level.

Time to Classify Images. In these experiments, random forest classifier consistently yielded good accuracy for classifying images at both levels. It took around 0.361 s to evaluate the largest feature set using random forest classifier. If the time to classify using random forest classifier is included, the following timings provided in

Table 5.20 Confusion matrix of hierarchical classification (FL: the first level, SL: the second level)

	SL = True	SL = False
FL = True	2103	147
FL = False	84	23

Table 5.21 Confusion matrix for the first level

	Class	Actual		
		0	1	2
Prediction	0	1474	1	1
	1	2	461	73
	2	2	29	314

Table 5.22 Confusion matrix for non-crystal classification (*: first-level misclassification)

	Non-crystals	Actual		
		Clear drop	Phase separation	Precipitate
Prediction	Clear drop	1265	0	20
	Phase separation	0	0	0
	Precipitate	8	0	181
	*	0	1	3

Table 5.23 Confusion matrix for likely-leads classification (*: first-level misclassification)

	Likely-Leads	Actual	
		Microcrystals	Unclear bright images
Prediction	Microcrystals	97	14
	Unclear bright images	16	334
	*	9	21

Table 5.24 Confusion matrix for crystal classification (*: first-level misclassification)

	Crystals	Actual				
		Dendrites/Spherulites	Needles	2D plates	Small 3D	Large 3D
Prediction	Dendrites/Spherulites	11	1	0	4	0
	Needles	11	99	1	13	0
	2D plates	0	0	0	0	0
	Small 3D	32	7	2	95	12
	Large 3D	0	0	1	5	21
	*	9	46	4	12	2

parentheses for the following feature sets are obtained: the best feature set for the first level (1.441 s), the best feature set for the second level (1.628 s), the fast feature set for the first level (1.131 s), the fast feature set for the second level (1.263 s), the union of the best feature sets (2.094 s), and the union of the fastest feature sets (2.271 s). Note that for the union of feature sets, the random forest classifier is applied twice (one for each level). These timings are promising for incorporating into real-time stand-alone computing systems. Since Crystal X2 takes around 3 s to move from one well to another well (including the time to move the plate and switching the light source), an option for real-time scoring has been implemented into the Crystal X2 system.

The Number of Features. The total number of features used in the experiments is 309. The union of best feature sets had 196 features, which is approximately 36% less than the total number of features. The fast feature set for the first level included 125 features, while the crystal subclassification had 65 features. If classifiers for the

first-level and crystal subcategory classification are used independently, this leads to around 60% and 80% reduction of features for the first-level and crystal subcategory classification using fast feature sets, respectively.

Individual Feature Sets. The individual feature sets were evaluated for the first level. The best classification performance was obtained by applying random forest classifier to normalized histogram features. This yielded 90.8% accuracy with 0.705 sensitivity. Intensity features using decision tree provided 87.7% accuracy with 0.701 sensitivity. DCT features provided the lowest accuracy of 69.1% with 0.48 sensitivity. The performance of histogram features is notable as it uses only 21 features, which can be extracted in 0.178 s. However, its relative low sensitivity (0.705) with respect to the sensitivity of the best feature set (0.87) makes using histogram features alone less desirable.

Use of Multiple Thresholding Methods. In the preliminary experiments, none of the thresholding methods produced good binarization consistently for all images in the dataset due to challenges mentioned in the introduction. Rather than choosing the best thresholding method among these, region features from all thresholded images were extracted and fed to classifiers. Among thresholding techniques, morphological thresholding did not improve accuracy much and it did not appear in feature sets leading to high accuracy. In other cases, good classifiers generally used region features from the two of the thresholding methods. This shows that classifiers can benefit from a set of thresholding methods if at least one of them provides good separation of the background and foreground.

Feature Selection and Reduction. Random forest classifier was used to rank features and PCA for feature reduction. The best accuracy for PCA and feature selection was obtained using 30 features by applying random forest classifier. PCA yielded 93.4% accuracy, while feature selection provided 96% accuracy. The sensitivity of PCA is low (0.74) with respect to the sensitivity of feature selection (0.863). The performance of feature selection is remarkable and slightly less with respect to the performance of the best classifier.

Performance of Classifiers and Generalizability. Random forest classifier consistently performed better than other classifiers. After observing that random forest is more reliable than other classifiers in Exp. ID 1, the best experimental conditions were repeated in Exp. ID 2 using random forest to validate the consistency of their high performance. Normalization barely affected the performance of random forest classifier. There were cases where normalization slightly lowered the performance. A small set of experiments has been performed to measure generalizability over 5 different test sets of 100 samples. SVM had the best generalizability followed by the decision tree and then by the random forest classifier. However, the generalizability could still be an issue for diverse datasets. The experiments provide the best set of feature sets for each classifier. The best model may need to be retrained for a larger new dataset. If the best model cannot generalize well, the next best model that could generalize could be selected for actual experiments. Overfitting is possible

with random forest classifier if many features are used or too many terminal nodes are allowed while building weak classifiers and the dataset does not cover all possible cases. To avoid overfitting, the number of features or the number of terminal nodes may be reduced for the random forest classifier.

5.9 Summary

In this chapter, feature analysis was performed for protein crystallization trial images benefiting from trace fluorescent labeling. Trace fluorescent labeling along with feature analysis method helps enable real-time scoring for the Crystal X2 system. Feature extraction and classification can be completed in around 2 s. For hierarchical classification, it may be reasonable to maximize the common feature sets between levels of classification hierarchy. The best experimental results were obtained using combinations of intensity features, region features using Otsu's thresholding, region features using green percentile G_{90} thresholding, region features using green percentile G_{99} thresholding, graph features, and histogram features. Using this feature set combination, 96% accuracy was achieved for the first level of classification to determine the presence of crystals and 74.2% accuracy for (5-class) crystal subcategory classification using random forest classifier. The correctness of the first-level classification should be given more weight since misclassification at the first level affects the second level. The choice of the fastest feature set for each level does not improve overall time if the set of common features is small or empty.

The use of all features may not only increase the processing time but may also lower the accuracy. Using all features had an adverse effect on the crystal subcategory classification. It reduced the accuracy from 74.2 to 69.6% and sensitivity from 0.909 to 0.618. The experiments show that protein crystallization classification would benefit from feature reduction in terms of time and accuracy. The histogram autocorrelation features ranked high when a feature selection method was applied. Graph features were included in the best feature sets for crystal subcategory classification. DCT features did not have significant positive impact on the accuracy despite its high computational time. Intensity and region features were generally involved in high accuracy feature sets and ranked high in the results of feature selection method. The random forest classifier provided the best results among classifiers in most cases.

If there is no single thresholding method that works well for all images in the dataset, classifiers may benefit from the outcomes of multiple thresholding methods assuming at least one of them produces a good result for an image. The feature sets that yielded high accuracy generally included region features from at least two of the thresholding methods. It was also interesting to observe that the region features from morphological thresholding were not included in the best feature sets.

The exhaustive method of trying different combinations of feature sets, classifiers, crystallization categories, feature selection/reduction methods and normalization helped observe overall performance about feature sets with different classifiers.

Timing analysis for feature sets helps identify the best feature set to achieve a specific accuracy within specific time.

The experiments have been conducted rigorously and improvements or updates have been made as needed throughout the course of experiments. Such updates include ignoring some unnecessary features, updating some existing features, and adding new features as needed. The work mentioned in this chapter can be further enhanced on two dimensions: (1) reduce time to classify and (2) improve accuracy/sensitivity. When feature extraction time per feature set was computed, the timings were computed individually. The feature extraction has common intermediate steps among feature sets. For example, if the foreground and background intensities are computed, the overall intensity of the image can be computed from these features without processing the complete image again. The intermediate steps do not need to be executed again if the outputs of intermediate results are stored. Moreover, each feature set may have irrelevant features that may not improve the accuracy. If irrelevant features are eliminated, the time to extract features is reduced as well. To improve the accuracy/sensitivity, images that were not classified correctly should be identified. A new set of features may need to be extracted and analyzed for those images to improve the accuracy. It would be interesting to evaluate textons [21] that were used to rank crystallization droplets for presence of crystals [27] as another feature set. No significant advantage of using simpler approaches such as linear discriminant analysis or other ensemble methods has been observed, however, they could be tried by identifying best parameter combinations and determined if they improve the overall performance.

Acknowledgements The original version of this chapter appeared as M. Sigdel, I. Dinc, M. S. Sigdel, S. Dinc, M. L. Pusey, and R. S. Aygun, “Feature analysis for classification of trace fluorescent labeled protein crystallization images,” *BioData Mining*, vol. 10, p. 14, 2017 [41]. Some modifications have been made to fit into this book.

References

1. Bern, M., Goldberg, D., Stevens, R. C., & Kuhn, P. (2004). Automatic classification of protein crystallization images using a curve-tracking algorithm. *Journal of applied crystallography*, 37(2), 279–287.
2. Berry, I. M., Dym, O., Esnouf, R., Harlos, K., Meged, R., Perrakis, A., et al. (2006). Spine high-throughput crystallization, crystal imaging and recognition techniques: current state, performance analysis, new technologies and future aspects. *Acta Crystallographica Section D: Biological Crystallography*, 62(10), 1137–1149.
3. Calle, M. L., & Urrea, V. (2011). Letter to the editor: stability of random forest importance measures. *Briefings in Bioinformatics*, 12(1), 86–89.
4. Clausi, D. A. (2002). An analysis of co-occurrence texture statistics as a function of grey level quantization. *Canadian Journal of Remote Sensing*, 28(1), 45–62.
5. Cumbaa, C., & Jurisica, I. (2005). Automatic classification and pattern discovery in high-throughput protein crystallization trials. *Journal of Structural and Functional Genomics*, 6(2–3), 195–202.

6. Cumbaa, C. A., & Jurisica, I. (2010). Protein crystallization analysis on the world community grid. *Journal of Structural Functional Genomics*, 11(1), 61–9.
7. Cumbaa, C. A., & Jurisica, I. (2010). Protein crystallization analysis on the world community grid. *Journal of Structural and Functional Genomics*, 11(1), 61–69.
8. Cumbaa, C. A., Lauricella, A., Fehrman, N., Veatch, C., Collins, R., Luft, J., et al. (2003). Automatic classification of sub-microlitre protein-crystallization trials in 1536-well plates. *Acta Crystallographica Section D: Biological Crystallography*, 59(9), 1619–1627.
9. Dinç, I., Dinç, S., Sigdel, M., Sigdel, M. S., Aygün, R. S., Pusey, M. L. (2015). Chapter 12–dt-binarize: A decision tree based binarization for protein crystal images. In Morgan Kaufmann *In Emerging Trends in Image Processing, Computer Vision and Pattern Recognition*. (pp. 183–199).
10. Dinç, I., Dinç, S., Sigdel, M., Sigdel, M. S., Pusey, M. L., Aygün, R. S. (2014). Dt-binarize: A hybrid binarization method using decision tree for protein crystallization images. In *Proceedings of The 2014 International Conference on Image Processing, Computer Vision & Pattern Recognition, ser. IPCV*, vol. 14, (pp. 304–311).
11. Dinç, I., Dinç, S., Sigdel, M., Sigdel, M., Pusey, M. L., Aygun, R. S. (2016). Super-thresholding: Supervised thresholding of protein crystal images. In *IEEE/ACM transactions on computational biology and bioinformatics*.
12. Dinç, I., Pusey, M. L., Aygün, R. S. (2015). Protein crystallization screening using associative experimental design. In *International Symposium on Bioinformatics Research and Applications*, (pp. 84–95). Berlin: Springer,
13. Dinç, I., Sigdel, M., Dinç, S., Sigdel, M. S., Pusey, M. L., Aygun, R. S. (2014). Evaluation of normalization and pca on the performance of classifiers for protein crystallization images. In *South east conference 2014, IEEE* (pp. 1–6).
14. Dinç, İ., Pusey, M. L., & Aygün, R. S. (2016). Optimizing associative experimental design for protein crystallization screening. *IEEE Transactions on Nanobioscience*, 15(2), 101–112.
15. Hampton research. Accessed 7 June 2016.
16. Haralick, R. M., Shanmugam, K., & Dinstein, I. H. (1973). Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, 6, 610–621.
17. Harris, C., Stephens, M. (1988). A combined corner and edge detector. In *Alvey vision conference* vol. 15, Citeseer, p. 50.
18. Hough, P. (1962) Method and means for recognizing complex patterns, US Patent 3,069,654.
19. Hung, J., Collins, J., Weldetsion, M., Newland, O., Chiang, E., Guerrero, S., Okada, K. (2014). Protein crystallization image classification with elastic net. In *SPIE Medical Imaging*, International Society for Optics and Photonics.
20. Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.
21. Leung, T., & Malik, J. (2001). Representing and Recognizing the Visual Appearance of Materials using Three-dimensional Textons. *International Journal of Computer Vision*, 43(1), 29–44.
22. Liu, R., Freund, Y., & Spraggon, G. (2008). Image-based crystal detection: a machine-learning approach. *Acta Crystallographica Section D: Biological Crystallography*, 64(12), 1187–1195.
23. MATLAB. (2013). *version 7.10.0 (R2013a)*. The MathWorks Inc., Natick, Massachusetts.
24. McPherson, A., Gavira, J. A. Introduction to protein crystallization. *Acta crystallographica. Section F, Structural biology communications* 70, Pt 1 (Jan. 2014), 2–20.
25. Mele, K., Lekamge, B. M. T., Fazio, V. J., & Newman, J. (2014). Using Time Courses To Enrich the Information Obtained from Images of Crystallization Trials. *Crystal Growth & Design*, 14(1), 261–269.
26. Mitchell, T.M. (1997). et al. Machine learning. wcb.
27. Ng, J. T., Dekker, C., Kroemer, M., Osborne, M., & von Delft, F. (2014). Using textons to rank crystallization droplets by the likely presence of crystals. *Acta Crystallographica Section D: Biological Crystallography*, 70(10), 2702–2718.
28. Otsu, N. (1975). A threshold selection method from gray-level histograms. *Automatica*, 11(285–296), 23–27.
29. Pan, S., Shavit, G., Penas-Centeno, M., Xu, D.-H., Shapiro, L., Ladner, R., et al. (2006). Automated classification of protein crystallization images using support vector machines with

- scale-invariant texture and gabor features. *Acta Crystallographica Section D: Biological Crystallography*, 62(3), 271–279.
30. Po, M. J., Laine, A. F. (2008). Leveraging genetic algorithm and neural network in automated protein crystal recognition. In *Proceedings of the 30th Annual International Conference of the IEEE(2008), Engineering in Medicine and Biology Society, EMBS 2008*, (pp. 1926–1929).
 31. Pusey, M. L., Liu, Z.-J., Tempel, W., Praissman, J., Lin, D., Wang, B.-C., et al. (2005). Life in the fast lane for protein crystallization and x-ray crystallography. *Progress in Biophysics and Molecular Biology*, 88(3), 359–386.
 32. Pusey, M., Barcena, J., Morris, M., Singhal, A., Yuan, Q., & Ng, J. (2015). Trace fluorescent labeling for protein crystallization. *Acta Crystallographica Section F*, 71(7), 806–814.
 33. Randomforest-matlab. Accessed 7 June 2016.
 34. Saitoh, K., Kawabata, K., Asama, H. (2006). Design of classifier to automate the evaluation of protein crystallization states. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, ICRA 2006* (pp. 1800–1805).
 35. Saitoh, K., Kawabata, K., Kunimitsu, S., Asama, H., Mishima, T. (2004). Evaluation of protein crystallization states based on texture information. In *Proceedings of the International Conference on 2004 IEEE/RSJ, Intelligent Robots and Systems, 2004.(IROS 2004)* vol. 3, IEEE, (pp. 2725–2730).
 36. Shapiro, L., Stockman, G. C. (2001). Computer vision. 2001. ed: Prentice Hall.
 37. Sigdel, M., Aygün, R. S. (2013) Pacc-a discriminative and accuracy correlated measure for assessment of classification results. In *Machine Learning and Data Mining in Pattern Recognition*. (pp. 281–295). Springer.
 38. Sigdel, M., Dinc, I., Dinc, S., Sigdel, M. S., Pusey, M. L., Aygiun, R. S. (2014). Evaluation of semi- supervised learning for classification of protein crystallization imagery. In *Proceedings of South East Conference, IEEE*, (pp. XX).
 39. Sigdel, M., Sigdel, M. S., Dinc, I., Dinc, S., Aygün, R. S., Pusey, M. L. (2015). Chapter 27 - automatic classification of protein crystal images. In Morgan Kaufmann, *In Emerging Trends in Image Processing, Computer Vision and Pattern Recognition*. (pp. 421–432).
 40. Sigdel, M., Pusey, M. L., & Aygun, R. S. (2013). Real-time protein crystallization image acquisition and classification system. *Crystal Growth & Design*, 13(7), 2728–2736.
 41. Sigdel, M., Dinc, I., Sigdel, M. S., Dinc, S., Pusey, M. L., & Aygun, R. S. (2017). Feature analysis for classification of trace fluorescent labeled protein crystallization images. *BioData Mining*, 10, 14.
 42. Soh, L.-K., & Tsatsoulis, C. (1999). Texture analysis of sar sea ice imagery using gray level co-occurrence matrices. *Geoscience and Remote Sensing, IEEE Transactions on*, 37(2), 780–795.
 43. Spraggon, G., Lesley, S. A., Kreusch, A., & Priestle, J. P. (2002). Computational analysis of crystallization trials. *Acta Crystallographica Section D: Biological Crystallography*, 58(11), 1915–1923.
 44. Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining* (1st ed.). Inc, Boston, MA, USA: Addison-Wesley Longman Publishing Co.
 45. Walker, C. G., Foadi, J., & Wilson, J. (2007). Classification of protein crystallization images using fourier descriptors. *Journal of Applied Crystallography*, 40(3), 418–426.
 46. Wilson, J. (2002). Towards the automated evaluation of crystallization trials. *Acta Crystallographica Section D: Biological Crystallography*, 58(11), 1907–1914.
 47. Xu, G., Chiu, C., Angelini, E. D., Laine, A. F. (2006). An incremental and optimized learning method for the automatic classification of protein crystal images. In *2006 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2006* (pp. 6526–6529). New York.
 48. Yang, X., Chen, W., Zheng, Y. F., Jiang, T. (2006). Image-based classification for automating protein crystal identification. In *Intelligent Computing in Signal Processing and Pattern Recognition*, (pp. 932–937). Berlin: Springer.
 49. Yann, M. L.-J., Tang, Y. (2016). Learning deep convolutional neural networks for x-ray protein crystallization image analysis. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*.

50. Zheng, Y., Wang, X., & Wang, C. (2014). Shape-adaptive dct and its application in region-based image coding. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 7(1), 99–108.
51. Zhu, X., Sun, S., Bern, M. Classification of protein crystallization imagery. (2004). In *Engineering in Medicine and Biology Society, 2004. IEMBS 04. 26th Annual International Conference of the IEEE* vol. 1, IEEE, (pp. 1628–1631).
52. Zuk, W. M., & Ward, K. B. (1991). Methods of analysis of protein crystal images. *Journal of Crystal Growth*, 110(1), 148–155.

Chapter 6

Crystal Growth Analysis

Abstract In recent years, high-throughput robotic setups have been developed to automate the protein crystallization experiments, and imaging techniques are used to monitor the crystallization progress. Images are collected multiple times during the course of an experiment. Huge number of collected images make manual review of images tedious and discouraging. In this chapter, utilizing *trace fluorescent labeling*, we describe an automated system for monitoring the protein crystal growth in crystallization trial images by analyzing time sequence images. Given the sets of image sequences, the objective is to develop an efficient and reliable system to detect crystal growth changes such as new crystal formation and increase of crystal size. This system consists of three major steps—identification of crystallization trials proper for spatiotemporal analysis, spatiotemporal analysis of identified trials, and crystal growth analysis.

6.1 Introduction

The success rate of crystallization can be as low as 2% depending on the protein and how crystallization trials are set up [25]. The crystallization depends on numerous factors such as protein purity, protein concentration, type of precipitant, crystallization techniques, etc. Finding the right combination of these factors that would lead to crystallization is challenging since it requires setting up thousands of crystallization trials with different combination of conditions using techniques such as incomplete factorial design [13]. Moreover, protein crystallization is not an instantaneous process but rather a temporal process. The time required for the growth of crystals can take from several hours to many months. A crystallographer may thus be faced with hundreds of thousands of images of crystallization experiments to be reviewed. Due to the abundance of images collected, the manual review of the images is impractical. To remove the burden of checking every trial image one by one by a group of experts and to automate the crystallization experiments, high-throughput robotic setups coupled with imaging techniques are used. Typically, the crystallization trials are prepared in well-plates with hundreds of wells and multiple droplet locations at each well corresponding to the different crystallization conditions. The robotic arrangement

is made to scan the well-plates capturing images at each droplet location in each well [17]. Most of these automated systems try to classify the crystallization trial images according to the crystallization state of the protein solution. Trials with no positive sign for successful crystallization may be discarded. On the other hand, the likely crystals are further reviewed either to optimize the conditions to prepare better crystals or to determine its suitability for use in X-ray crystallography.

A popular way of helping crystallographers using an automated system is to classify and score images collected by the robotic systems. The number of categories that are identified for crystal trial image classification may give different levels of information about the crystal growth process or optimization of conditions for further trials. Trial images are typically classified into crystal and non-crystal categories to identify experiments leading to successful crystallization (Zuk and Ward (1991) [35], Cumba et al. (2003) [7], Cumba et al. (2005) [5], Zhu et al. (2004) [34], Berry et al. (2006) [3], Pan et al. (2006) [20], Po and Laine (2008) [21]). To provide more information about the crystallization process, the protein crystal growth can be divided into multiple phases. Yang et al. (2006) [33] classified the trials into three categories (clear, precipitate, and crystal). In [26], images were classified into non-crystals, likely leads, and crystal categories. Bern et al. (2004) [2] and Saitoh et al. (2006) [24] classified images into five categories (empty, clear, precipitate, microcrystal hit, and crystal). Likewise, Spraggon et al. (2002) [29] and Cumba et al. (2010) [6] have proposed six categories for the classification. In [28], the focus was on the classification of crystal categories only. In these research studies, each image of a well-plate or droplet location is analyzed individually without considering the previous collected images. For example, it is not possible to determine whether crystals are growing or the number of crystals is increasing without comparing with the history of images collected for that specific well or droplet location.

Besides the accuracy of classifying protein crystallization trial images, another major issue is effective extraction of features for developing real-time stand-alone systems. Systems that utilize high-performance, distributed, or cloud computing extracted as many features as possible hoping that the huge set of features could improve the accuracy of the classification [6]. Some of the most commonly extracted image features are Hough transform features [21], gradient and geometry-related features [21, 28], blob texture features [24], features from differential image [24], Discrete Fourier Transform features [30], features from multiscale Laplacian pyramid filters [32], histogram analysis features [5], Sobel-edge features [31], gray-level co-occurrence matrix features [34], etc. While generating huge number of features may help categorize images at a satisfactory level, it is not suitable for real-time systems where results should be delivered instantly as the plates are scanned. As mentioned in Chap. 5, the image processing and feature extraction have been computationally expensive making it infeasible for real-time processing using stand-alone systems.

This chapter describes the CrystPro system [27] for analysis of trace fluorescently labeled (TFL) protein crystallization results. Prior to setting up the crystallization trials, the proteins are trace fluorescently labeled (TFL) as described elsewhere [9, 18, 22]. The basis of this technique is to covalently label a low proportion, typically 0.2%, of the protein molecules with a fluorescent probe, then removing all probe

molecules that are not attached to the protein. By removing all non-covalently bound probe, all fluorescence originates from a subpopulation of the protein molecules. Thus, the fluorescence is an indicator of the protein's behavior in the crystallization trial. When using the TFL approach, the fluorescence intensity is a function of the packing density of the covalently labeled protein. As the crystalline phase is the most densely packed, this leads to the principle that intensity equals structure. As the fluorescence is in the visible wavelengths, both automated and manual observations are feasible. As image analysis is dependent upon intensity and not wavelength then this approach should be equally useful for images acquired using U.V. fluorescence.

In spatiotemporal analysis of protein crystallization trial images, it is important to note that the growth of regions are not necessarily indicators of crystals. Spatiotemporal analysis may include comparing only sequential images or tracking regions in a series of images and comparing their growth. Not all crystallization trial images are relevant for growth analysis. Hence, in this chapter, we describe identifying trials for spatiotemporal analysis and then show how growth analysis can be performed on those images.

6.2 Is it a Protein—Rule of Thumb

The bane of protein crystallization is the propensity for the nonprotein components of a solution to come out of solution as crystals. Even worse, the crystals are often nicely faceted, satisfying the would-be structural biologists urge to getting highly photogenic crystals onto the beamline for the collection of high resolution diffraction data. Of course, there are several problems with this scenario, the first of which is that good looking does not translate into well diffracting. Nothing ruins a data collection more than putting that well-faceted crystal into the X-ray beam and getting a salt diffraction pattern. To avoid this, crystal growers over the years have developed several rules, most of which are at least 50% true, for avoiding the mounting and diffracting of salt crystals.

The first rule is that if it appears quickly it is probably not a protein crystal. Salt and protein crystal growth kinetics follow the same principles, but their growth kinetics are typically very different. Salt diffusivities will be 1–2 orders of magnitude (or more) greater than protein, so just the transport to the growing surface for proteins will be slower, and thus the growth rate. Salt crystals often appear soon after a plate is set up, and may, likely will, appear within a day. Due to the faster growth kinetics salt crystal growth will stop after a short period of time, so as a result a crystal that appears within the first 24 hours and does not get any larger is most likely salt.

Salt crystals are also mechanically robust. Protein crystals are typically ~ 25 to >65 % solvent channels, and are mechanically very fragile. This difference in properties led to the use of the crush test, pyrrhic determination where one applies mechanical pressure to the crystal. If it maintains its shape, or slips away without being destroyed, then it is likely salt. If, on the other hand, it collapses and turns into a cloud of brown dust, then you had a protein crystal, but do not anymore.

The presence of the solvent channels led to the development of a dye-binding assay. The dye typically employed is methylene blue, sold under the name of Izit by Hampton Research. A small amount of the blue dye solution is added to the crystal containing drop. If it is protein, within a few hours the dye will diffuse into the crystal via the solvent channels and bind to the protein molecules. The higher bound dye concentration in the crystals results in their acquiring a dark blue coloration. As salt crystals have no solvent channels they will not turn blue. A fluorescence variation on this approach has been described [11].

Our approach, as illustrated in this treatise, is to trace fluorescently label the protein prior to setting up the crystallization plate [10, 23]. The fluorescent dye is covalently attached to the protein, with the goal being to keep the labeled population to $\leq 0.5\%$. The key part of the labeling procedure is to then remove all free dyes after the reaction, so that the observed fluorescence is only that of the label attached to the protein. Because of this, one can then track what the protein is doing in response to the added precipitant solutions, and that any increased fluorescence intensity is due to higher (local) protein concentrations. Several other fluorescence-based approaches are extant for identifying the protein, but not the salt, crystals in screening plates [1, 12, 14, 15].

6.2.1 Protein—Get it While it is Fresh

Proteins are derived from living organisms, and as such virtually all have a finite lifetime. Protein stability over the course of a crystallization experiment is a major concern of the crystallographer. While proteins are generally more stable in a closely packed crystal form, they may still degrade over time. Trace amounts of a contaminating protease, insufficient to influence the experimental crystallization outcome, may be present to degrade one's results. The crystallization well may become contaminated with a fungus or other microorganism. For these and other reasons it is best to harvest one's crystals as soon as possible.

Harvesting the crystal at the most opportune time means tracking the crystallization experiment. As mentioned above, salt crystals appear fast and stop growing soon after they appear. Protein crystals, on the other hand, may continue to slowly grow over days or weeks. Thus, while knowing the rate at which a crystal is growing, or not, may be key to determining if it is of salt or protein, this information can also inform when to best harvest the crystal and mount it for diffraction analysis.

6.3 Temporal Analysis of Time Series Images

Analyzing temporal process of protein crystallization may alleviate the burden on the experts. Recently, Mele et al. 2013 [17] described an image analysis program called Diviner that takes a time series of images from a crystallization trial and returns an

estimate of the crystallographically relevant change between the images of the time course. Mele et al. used the intensity change between images in a temporal image sequence as an indicator of crystallization (hopefully not dissolving of crystals). Diviner proposes the difference score for the overall image as a means to determine the importance of a crystallization trial. Their work leads researchers to work on images where a change in the crystallization droplet has been observed. However, the computational cost for the analysis is very high and not feasible for real-time analysis. Moreover, the pre-processing step of droplet (or drop boundary) detection proposed may not be effective for many crystallization trial images. For example, it may not be possible to detect the drop boundary because either the drop boundary is not present or the droplet is not circular.

Despite a few limitations, Mele et al. [17] have shown that time series analysis may hint the experts about images to explore further. By analyzing the temporal images, the changes in images can be visualized and some useful metrics regarding crystal growth can also be derived. Information such as the appearance of new crystals, changes in the size of crystals and change in the orientation can be tracked from the temporal image sequences. In addition, time series analysis can also provide a relatively easier method for determining whether a temporal image sequence leads to crystal formation or not. In this chapter, the goal is to extract further information beyond the detection of change from time series images of protein crystallization trial images. The detection of change between images may not always carry useful image for the experts. Identifying whether (1) crystals are growing and (2) the number of crystals is increasing could be helpful information for crystallographers. Such spatiotemporal analysis of crystallization trials involves identification of regions, their growth, and matching them in the previous images. Since it depends on region segmentation and analysis of regions, there are further complications when time series images are analyzed due to the following reasons.

- (i) Varying illumination, improper focusing, nonuniform shapes and varying orientation of crystals impose complexity in image analysis.
- (ii) Since too many experiments are set up and the images are collected many times during the course of the experiment, image processing can take significant time. Existing automated systems have very high computational cost.

6.3.1 Stages of Temporal Analysis

Time series analysis for checking the number of crystals and change in size of crystals requires comparison of images collected at different time instances for a specific well or a droplet location. Consider a sequence of protein crystallization trial images captured at n different time instances represented as $I = \{I_1, I_2, \dots, I_n\}$ where I_k represents the image collected at the k^{th} time instant and $1 \leq k \leq n$. In Fig. 6.1, crystallization trial images corresponding to 3 protein wells that are captured at 3 time instances are shown. In the first image sequence (Fig. 6.1a), all 3 images consist

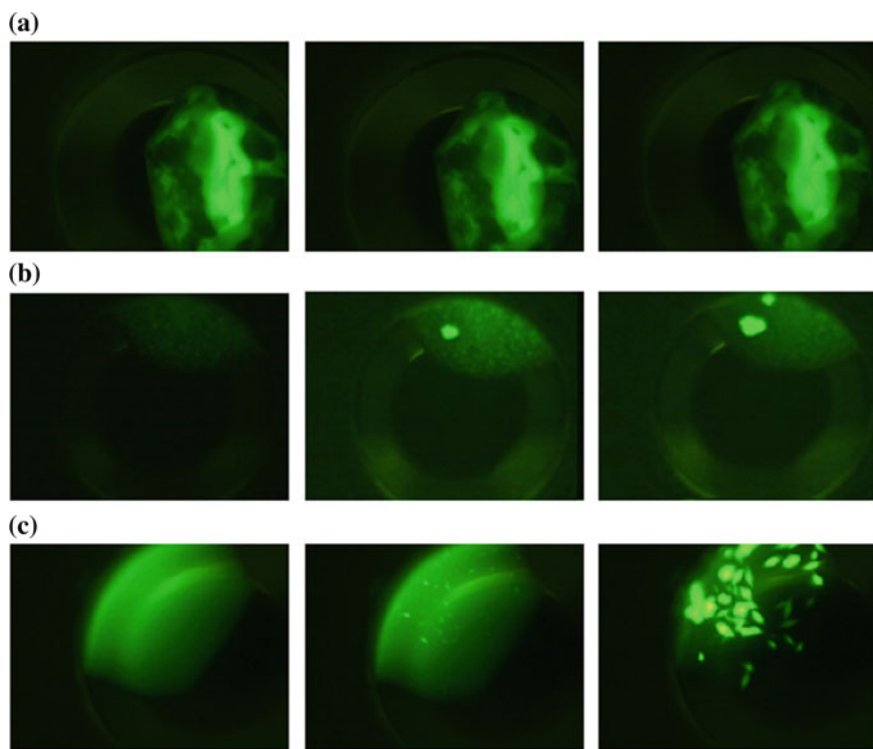


Fig. 6.1 Sample temporal images of crystallization trials. Reprinted (adapted) with permission from *Crystal Growth and Design* 2013 15 (11), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 5254–5262. Copyright (2015) American Chemical Society

of cloudy precipitates without crystals. Hence, this sequence is not interesting for the crystallographers. In the second (Fig. 6.1b) and third sequences (Fig. 6.1c), crystals are formed in the later stages of the experiments. In these images, the increase in the number or growing size of crystals can provide important information for the crystallographers. Since protein crystallization is very rare, only few experiments lead to crystals. Therefore, it is important to first correctly identify such sequences. Once such sequences are identified, an image in the sequence can be compared with any other image in the sequence leading to a combination of $C(n, 2)$ comparison of images.

Figure 6.2 shows the basic components of CrystPro, the crystal growth analysis system. CrystPro consists of three major steps: (1) identification of trials for spatiotemporal analysis, (2) spatiotemporal analysis of identified trials, and (3) crystal growth analysis. Spatiotemporal analysis is especially relevant and useful on crystallization trials having crystals. Therefore, first the trials that have crystals having fair size in any image of a crystallization trial are identified. By doing so, the trials which do not have crystals are eliminated or the crystals are very small where



Fig. 6.2 Overview of CrystPro system. Reprinted (adapted) with permission from *Crystal Growth and Design* 2013 15 (11), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 5254–5262. Copyright (2015) American Chemical Society

spatiotemporal analysis is hard to achieve. Next, spatiotemporal processing is done on the selected trials and spatiotemporal metrics such as change in intensity, foreground area, number of crystals, etc., are obtained. These metrics are used to identify crystal growth such as new crystal formation, crystal size increase, etc., and to perform further analysis.

6.3.2 Sample Dataset and Experimental Setup

The images of crystallization trials are collected using Crystal X2 software from iXpressGenes, Inc. The proteins are trace covalently labeled with a fluorescent probe. As previously shown [9, 18, 22], the use of fluorescence significantly speeds the crystallization plate review process. The primary search criteria is intensity, not straight lines, which simplifies results interpretation by either software or direct visualization. Additionally, crystals that are obscured by other features in the well, such as buried in precipitate, growing in an out of focus location along the edge, etc., are readily apparent simply by the presence of their fluorescence intensity.

In these experiments, green light emitting diodes (LEDs) were used as the excitation source. Carboxyrhodamine is used as the covalent labeling. Spatiotemporal analysis has been performed on three datasets—PCP-ILopt-11, PCP-ILopt-12 and PCP-ILopt-13. Each dataset is captured from a 96-well plate with 3 sub-wells scanned 3 times at different dates leading to 864 images per dataset. In the experiments, the time gap between the first and second captures for PCP-ILopt-11, PCP-ILopt-12, and PCP-ILopt-13 are 2 days. The interval between the second and third captures for PCP-ILopt-11, PCP-ILopt-12, and PCP-ILopt-13 are 3, 3, and 4 days, respectively. The size of the dataset is limited due to the number of crystalline outcomes that can have gradual growth in crystal sizes and new formation of crystals. In these experiments, the hyperthermophile-derived protein pcp (pyroglutamate amino peptidase) was being subjected to optimization trials using a series ionic liquids at 0.1M concentration. The trace fluorescent labeling approach was necessary as the ionic liquids also often crystallized, giving false positives under white light, but not fluorescence, illumination. Expert scores are obtained by a knowledgeable observer (MLP) first closely examining and then scoring each well under white light microscopy. The scores are then reviewed by reference to the latest fluorescent images, to eliminate salt crystals from the high scoring outcomes. Regions of high fluorescence intensity

that are not reflected in the first pass white light score lead to a second scrutiny under white light to resolve the source of the intensity. Thus for manual scoring, fluorescence is used as a feedback mechanism to maximize the correct scoring.

The image processing routines are implemented in MATLAB. The user interface for image visualization is developed using Windows Presentation Framework (WPF). Microsoft Visual Studio 2012 is used as the IDE. On a Windows 7 Intel Core i7 CPU @2.8 GHz system with 12 GB memory, it took around 40min (or 2400 sec) to process (extract features from) 2592 images (3*864 images per dataset) in the first stage. The average processing time per image is about 1s. Hence, the analysis for spatiotemporal analysis between a pair of images takes around 2s.

6.4 Identifying Trials for Spatiotemporal Analysis

The first stage of spatiotemporal analysis is the selection of suitable crystallization trials for the analysis. Image thresholding is often used in image analysis for the separation of foreground regions from the background. Obtaining a good binary image is very critical in image analysis because any error in the binary image will propagate to further processing steps. For example, regions that belong to a crystal should not be cropped in the thresholding stage. If a crystal region is partially detected, then the growth of a crystal may not be analyzed properly. Varying illumination, improper focusing, nonuniform shapes and varying orientation of crystals impose complexity in separating crystals correctly. Moreover, images with skin formation should be identified in this phase and should not be evaluated for crystal growth in size or new crystal formation. Otsu's thresholding [19] is a very popular method for thresholding images. Likewise, Canny edge detection [4] has been widely used in the literature to detect shapes of objects. In this study, since new crystal formation and growth of crystal size for spatiotemporal analysis are considered, the size of crystals should be comparable and not sensitive to poor binarization of images. Therefore, the likely crystals are expected to have closed regions and within a certain size range for the spatiotemporal analysis. A single technique may not work properly for all images. To identify images with crystals, the results from both techniques are used for detecting trials with crystals. Instead of extracting large number of image features and applying a training based system, this method is quick and effective.

6.4.1 Image Thresholding

Segmenting an image into regions and then determining the important regions for further processing are important processes in image analysis. A thresholding algorithm typically classifies pixels into two classes: background pixels (pixels having intensity lower than a certain threshold) and the foreground pixels. Numerous image thresholding techniques have been proposed in the literature. Otsu's method [19]

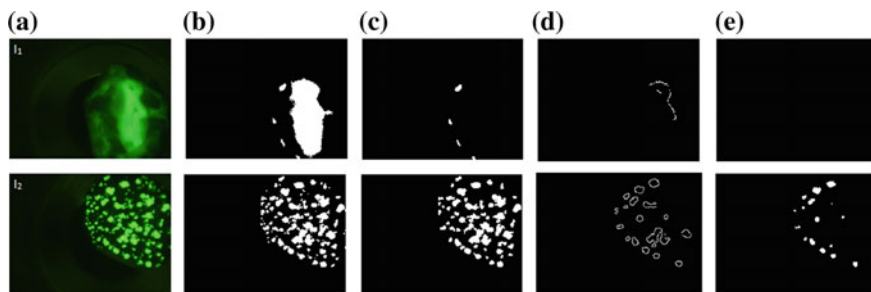


Fig. 6.3 Identifying images with crystals **a** original images, **b** Otsu thresholded images, **c** eliminating very small and very large regions, **d** Canny edge image, and **e** regions with closed components in the edge image. Reprinted (adapted) with permission from *Crystal Growth and Design* 2013 15 (11), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 5254–5262. Copyright (2015) American Chemical Society

iterates through all possible threshold values and calculates a measure of spread of the pixel levels in foreground or background region. Otsu's image segmentation method selects an optimum threshold by maximizing the between-class variance in a grayscale image. Figure 6.3b provides results using Otsu's threshold on 2 images. Here, the first image (I_1) does not have crystals but the second image (I_2) does. By eliminating the very large non-crystal regions (for example, region size above 2.5% of image size), the binary images shown in Fig. 6.3c are obtained. These results show that non-crystal regions might appear as the foreground as an outcome of the thresholding method.

6.4.2 Canny Edge Detection

Canny edge detection algorithm [4] is one of the most reliable algorithms for edge detection. The results show that for most cases, the shapes of crystals are kept intact in the resulting edge image. In this study, crystals regions that have closed components are considered. With this assumption, the unclosed edges are eliminated. Figure 6.3d provides the Canny edge image for the original images in Fig. 6.3a. Figure 6.3e shows the result after eliminating the unclosed regions. The final image (Fig. 6.3e) for the first image is blank, which suggests that there are no crystals. The final image for the second image has likely crystal regions.

6.4.3 Merging Results of Thresholding and Canny Edge Detection

The results from Otsu's thresholding and Canny edge detection are used to identify trials with large crystals. For reliable detection of crystal growth analysis, the region of a crystal should have at least 10 pixels for 320x240 image resolution. The size can be matched to other resolutions proportionally. Figure 6.4 shows the basic flow of the method. For every image, Otsu's thresholding image is applied first and very small and very large regions are eliminated. If the resulting binary image has foreground objects, Canny edge image is applied and likely crystal regions are determined. An image is considered to have a crystal if it has likely crystal regions from the results from Otsu's binary image as well as Canny edge image. A crystallization trial is considered to have crystals if at least one instance of the trial image is found to have crystal. This method provides a quick and accurate approach to identify trials suitable for spatiotemporal analysis.

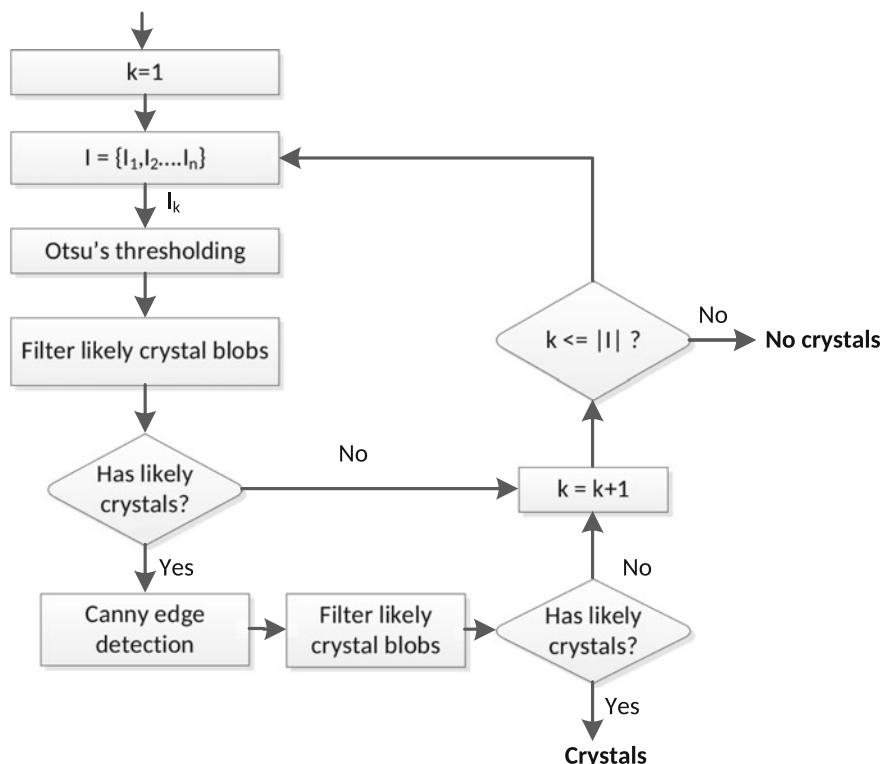


Fig. 6.4 Process flow to identify crystallization images with crystals. Reprinted (adapted) with permission from *Crystal Growth and Design* 2013 15 (11), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 5254–5262. Copyright (2015) American Chemical Society

Table 6.1 Results of crystal detection on the 3 datasets

Dataset	TN	FP	FN	TP	ACC	SENS	PREC
PCP-ILopt-11	261	2	3	22	0.98	0.88	0.92
PCP-ILopt-12	273	6	1	8	0.98	0.89	0.57
PCP-ILopt-13	273	3	1	11	0.99	0.92	0.79

6.4.4 Evaluation

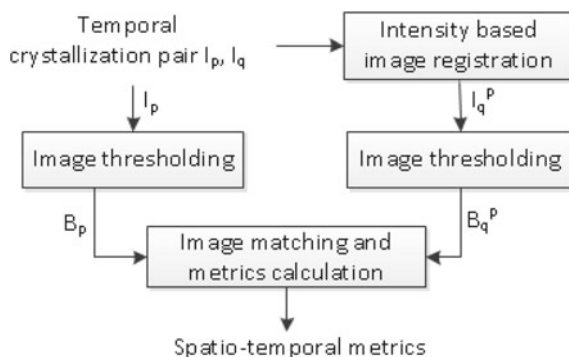
The first stage of new crystal detection or crystal growth in size is the detection of crystal regions in images. The classification results of identifying candidate trial images for 3 datasets are provided in Table 6.1. Each dataset has 864 images collected from 288-well-plate-collected at 3 time instances. The predicted results from this method are compared against expert scores. In Table 6.1, true negative (TN), false positive (FP), false negative (FN), and true positive (TP) refer to the number of trials correctly predicted as non-crystals, the number of non-crystal trials predicted as crystals, the number of trials incorrectly predicted as non-crystals, and the number of trials correctly identified as crystals, respectively.

For PCP-ILopt-11, the system predicts 24 (2 FP + 22 TP) trials with crystals. Similarly, for PCP-ILopt-12, the system predicts 14 trials (6 FP + 8 TP) with crystals. Likewise, PCP-ILopt-13 has 14 (3 FP + 11 TP) crystallization trials predicted to have crystals. The accuracy of the system is above 98% for all 3 datasets. Likewise, the sensitivity for detecting trials with crystals is also very high. There are few false negatives (missing trials with crystals). Those trial images are either blurred or have illumination problem. The system is able to reject non-crystals trials with very high accuracy. As successful trials are rare, this will help eliminate large proportion of unsuccessful trials.

6.5 Spatiotemporal Analysis of Protein Crystal Growth

Given a set of images of a protein well captured at different time instances, the goal is the analysis of the growth of protein crystals. This analysis requires matching crystals in two images and identifying the changes. Information such as appearance of new crystals, dissolving of crystals, changes in the size of crystals, etc. gives useful information about the growth of crystals. Figure 6.5 shows the overview of the spatiotemporal analysis system. The first stage of this analysis is the registration of images that are collected at different time instances. Since the robotic microscope may not have the same exact position each time a well is captured, the images should

Fig. 6.5 Overview of spatiotemporal analysis. Reprinted (adapted) with permission from Crystal Growth and Design 2013 15 (11), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 5254–5262. Copyright (2015) American Chemical Society



be aligned to make proper analysis. For an image pair (I_p, I_q) , image I_q is aligned with respect to I_p and is represented as I_q^p . Next, binary images are obtained for the images I_p and I_q^p represented as B_p and B_q^p respectively. The binary images are then matched and spatiotemporal metrics related to crystal growth changes are extracted. These spatiotemporal metrics are used to predict the crystal growth changes. In this system, in addition to comparing consecutive images, the first image in the sequence is also compared with the last image in the sequence. This helps to detect overall change in the solution. For n images in a sequence, this results in n comparisons of images in the sequence. The steps of spatiotemporal analysis of protein crystal growth are described next.

6.5.1 Identifying Crystallographically Important Regions

In general, the protein crystallization images containing crystals have 4 different intensity regions: background region, droplet (the solution), high-intensity regions around crystals and the crystal regions with the highest intensity. The background is the least illuminated region. The droplet region has higher illumination than the background. If crystals are present in an image, crystal regions will have the highest intensity. Also, the regions around the crystals have high intensity because of the presence of crystals. The spatiotemporal analysis of crystallization images include analyzing the shape, size, and growth of the crystals. The main purpose is to separate crystal regions from the rest of pixels. In the experiments, a single threshold produced poor results for some images.

Otsu's method [19] has been extended to generate multiple thresholds and categorize multiple classes of pixels in an image. For example, given an image, k thresholds can be used or computed to differentiate the image pixels into $k + 1$ classes according to intensity. To separate pixels in an image into four classes, three thresholds need to be calculated. Therefore, multilevel thresholding is applied with three threshold levels to identify four intensity regions. The binary image is obtained by considering

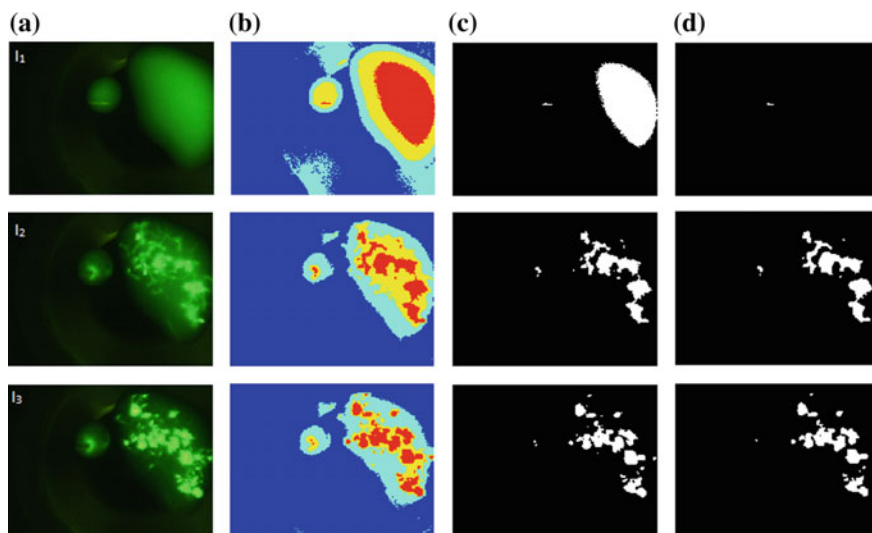


Fig. 6.6 Image binarization using multi-level Otsu's thresholding, **a** original image, **b** image segmentation into 4 classes using Otsu's thresholding, **c** binary image obtained by selecting the last class pixels as foreground and the rest as background, and **d** final binary image after filtering minimum (5 pixels) and maximum (2.5% of image size) sized regions. Reprinted (adapted) with permission from *Crystal Growth and Design* 2013 15 (11), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 5254–5262. Copyright (2015) American Chemical Society

the pixels above the highest threshold as the desired foreground. The rest of the pixels are considered background. Furthermore, region segmentation is applied and very small and very large regions are eliminated.

Figure 6.6 shows the result of applying multilevel Otsu's method with three threshold levels on a sequence of crystallization trial images. In Fig. 6.6b, it is possible to observe how the pixels are separated into 4 intensity regions. The red pixels have the highest intensity and represent the foreground whereas the blue, cyan, and yellow pixels represent the low-intensity regions. By considering the red pixel regions as foreground, the binary images shown in Fig. 6.6c are obtained. It should be noted that the white regions in the binary image (Fig. 6.6c) do not necessarily represent crystals. Generally, if an image does not have crystals, the partial illumination of the protein droplet yields large foreground region in the binary image. Those regions are discarded by applying region segmentation and filtering out the regions that are larger than a certain threshold (e.g., 2.5% of the image area). Figure 6.6d shows the final binary images after filtering the regions based on the region size. These binary images are used for further analysis. For the majority, this method correctly identifies the crystals. However, if in an image there are multiple crystals with different illuminations, it is possible to miss the crystals having less illumination.

6.5.2 Image Registration and Alignment

Since images are collected at multiple time instances, and also because of the inaccuracy in the acquisition system, the images are not exactly aligned. Even though there may not be any change in the image sequence, the pixels may not appear at the same position in the consecutive images. To compare pixels of different images, intensity-based image registration [16] is applied. Very slight rotation is possible if the plate is not located exactly as it was positioned before. Rotation is ignored in the registration a) since neither the plate nor the microscope can rotate and b) rotation by positioning the plate is minor or can be covered by translations. The microscope lens can usually maintain its distance to the plate properly, so scaling parameters are also not considered. Hence, image registration only involves computation of translation parameters (t_x , t_y) for the next image in the sequence. Figure 6.7a,b shows images of a protein well captured at two time instances. Figure 6.7c shows the overlaying of image in Fig. 6.7b on top of image in Fig. 6.7a after alignment. The translation parameters are $t_x = -8$ (8 pixels to the left) and $t_y = 9$ (9 pixels to the bottom) for this example. For proper alignment, the second image is shifted to the left and bottom with respect to the first image.

6.5.3 Spatiotemporal Features

Spatiotemporal processing and feature extraction from image pairs are applied to analyze the crystal growth changes. Spatiotemporal features are compared for a pair of images (I_o and I_n). B_o and B_n represent the corresponding binary images. In this system, the dominant feature for indication of forming crystals is the increase in the intensity. The change of intensity ($\mu_d^{o,n} = \mu_n - \mu_o$) on the overall image can be used to determine whether crystals are forming. Similarly, the percentage change in the size of foreground area is computed as $A_d^{o,n} = 100 * |A_n - A_o| / A_o$. Since,

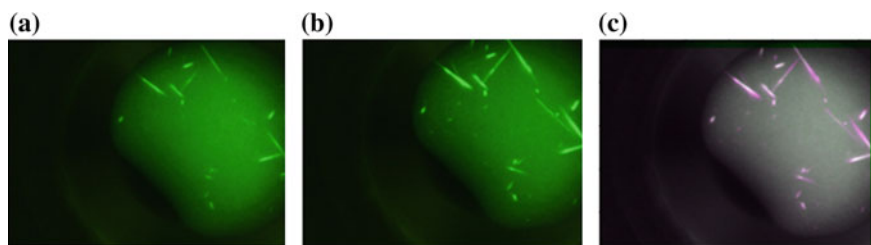


Fig. 6.7 Spatial alignment using intensity- based image registration **a** Image at time instance t_1 , **b** Image at time instance t_2 , and **c** Image 2 mapped on top of Image 1. Reprinted (adapted) with permission from Crystal Growth and Design 2013 15 (11), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 5254–5262. Copyright (2015) American Chemical Society

μ_d and A_d provide rough information about the overall change, the analysis can be improved based on the segmented regions. The change in the number of regions is computed as $\overline{N}^{o,n} = N_n - N_o$. To get further information whether the size of the regions increased or decreased, the number of matched pixels ($N_m^{o,n}$), the number of pixels in additional regions ($N_a^{o,n}$), and disappeared regions ($N_r^{o,n}$) are determined. Table 6.2 provides the list of spatiotemporal features and their descriptions. These features are extracted for every pair of consecutive images and also for the first and last image in the sequence. If there are k images in the sequence, $7 * k$ features are generated. These features are useful in the image review process.

Figure 6.8 provides a sample sequence of images for spatiotemporal analysis. Rather than providing a simple scenario having clear large crystals forming or growing, this example with small crystals and possible errors of segmentation is provided to show that the combination of these spatiotemporal features can provide useful information even for such cases. In Fig. 6.8, new crystals appear in the second image (I_2) and the growth of crystals is observed in the third image (I_3). In I_1 , there is no crystal but thresholding method identifies a foreground region (Fig. 6.8d) which has higher intensity than its surrounding. Images I_2 and I_3 have many small crystals that may or may not match (Fig. 6.8e,f). It is possible to see growth of some crystals in I_3 . Therefore, the pair of (I_1, I_2) can be used to test new crystal formation, and the pair of (I_2, I_3) can be used to test crystal size growth.

Table 6.3 provides the spatiotemporal features for 3 image pairs. While the average intensity difference (μ_d) is not significant, other measures such as the percentage change in foreground area (A_d), difference in the number of regions (\overline{N}), etc., provide important information about changes in the sequence. In I_2 , the number of regions is increased by 15 compared to I_1 . Similarly, there are 3 new regions in I_3 compared to I_2 . The change between I_1 and I_3 is more significant. The number of additional regions in I_3 compared to I_1 is 17. The increase in the foreground area indicates the presence of crystal growth.

Table 6.2 Spatiotemporal analysis metrics

#	Metric	Description
1	$\mu_d^{o,n}$	Average intensity difference
2	$A_d^{o,n}$	Percentage of change in the foreground area
3	$\overline{N}^{o,n}$	Difference in number of regions $ N_n - N_o $
4	$N_m^{o,n}$	No of matched object pixels
5	$N_a^{o,n}$	No of new object pixels
6	$N_r^{o,n}$	No of disappeared object pixels
7	$I_f^{o,n}$	Size increment factor

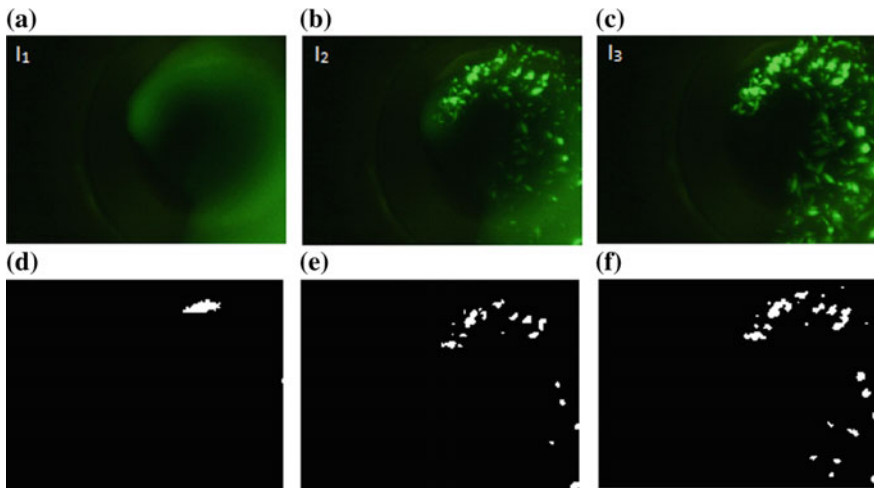


Fig. 6.8 Sample image sequence with new crystals as well as crystal size growth. Reprinted (adapted) with permission from *Crystal Growth and Design* 2013 15 (11), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 5254–5262. Copyright (2015) American Chemical Society

Table 6.3 Spatiotemporal metrics for image sequence in Fig. 6.8

Symbol	(I_1, I_2)	(I_2, I_3)	(I_1, I_3)
(t_x, t_y)	(-8,12)	(-7,11)	(-6,13)
μ_d	1.8	-0.6	1.2
A_d	105	63	240
\bar{N}	15	3	17
N_m	35	824	52
N_a	770	655	1287
N_r	358	82	341
I_f	23	1.8	25.8

From results in Table 6.3, it is possible to infer that new regions are forming in I_2 and I_3 , and crystal growth in size is observed in I_3 . Such information is useful for experts to gain knowledge about the phases of crystallization process after further review of the corresponding images. It is important to extract such information in challenging sequences as shown above where thresholding method may make mistakes. In the next section, these features are used to determine rules for predicting crystal growth.

6.6 Determining Crystal Growth

Consider an image pair (I_o, I_n) from a time series image trial, where I_n is a new image collected later than an old image I_o . To compare the changes in crystal growth, the binary images B_o and B_n are used. B_n^o represents the binary image after B_n is aligned with respect to B_o . The background pixels are represented by 0 and object pixels are represented by 1. The method from [8] is used for comparison of two regions. To find the common object pixels, new object pixels in B_n^o and the disappeared object pixels from B_o , first a sum image $B_{o,n}$ is computed using Eq. 6.1:

$$B_{o,n} = B_o + 2B_n^o \quad (6.1)$$

As in [8], the binary image B_n^o is multiplied by 2 and added with the binary image B_o . The sum image $B_{o,n}$ consists of 4 different values—0, 1, 2, and 3. The object pixels common in both images have value 3. They are referred as matched pixels. The object pixels in B_n but not in B_o have the value 2. Such pixels are added pixels. The object pixels in B_o but not in B_n^o have the value 1. These are considered removed pixels. Likewise, the background pixels in both the images have the value 0. Using this information, the following statistics is computed.

- The number of matched pixels (N_m) = $\sum_{i=1}^W \sum_{j=1}^H (B_{o,n} == 3)$
- The number of added pixels (N_a) = $\sum_{i=1}^W \sum_{j=1}^H (B_{o,n} == 2)$
- The number of removed pixels (N_r) = $\sum_{i=1}^W \sum_{j=1}^H (B_{o,n} == 1)$
- Size increment factor (I_f) to evaluate the growth of crystals:

$$I_f = \begin{cases} 0 & \text{if } N_m < \tau \text{ (}\tau \text{ is the minimum threshold to consider} \\ & \text{a match; default value is 10)} \\ N_a/N_m & \text{otherwise} \end{cases}$$

Figure 6.9a,b provides an image pair I_1 and I_2 collected at two time instances. Image I_2 is aligned with respect to image I_1 . Figure 6.9c,d shows the binary images B_1 and B_2 obtained using multilevel Otsu's thresholding method. Figure 6.9e shows the regions after B_1 and B_2 are matched. Here, the matched object pixels are shown in white, the additional pixels in B_2^1 are shown in green. The object pixels in B_1 missing in B_2^1 are shown in red. Numerically, the count of matched pixels is 2232 and the number of added pixels is 1223. Likewise, the size increment factor is equal to 0.55. This provides an estimation for increase in the size of the crystals in the second image.

Formation of new crystals or growing crystals lead to three types of changes: 1) the overall foreground area, 2) the number of crystals, and 3) the size of a matching region (hopefully a crystal). Therefore, this information is used to guide in identifying the type of protein crystal growth. Table 6.4 provides a list of rules used to predict if there is a crystal growth in a pair of images. The first rule is used to identify new crystal formation. The rule indicates that there must be some change in the number

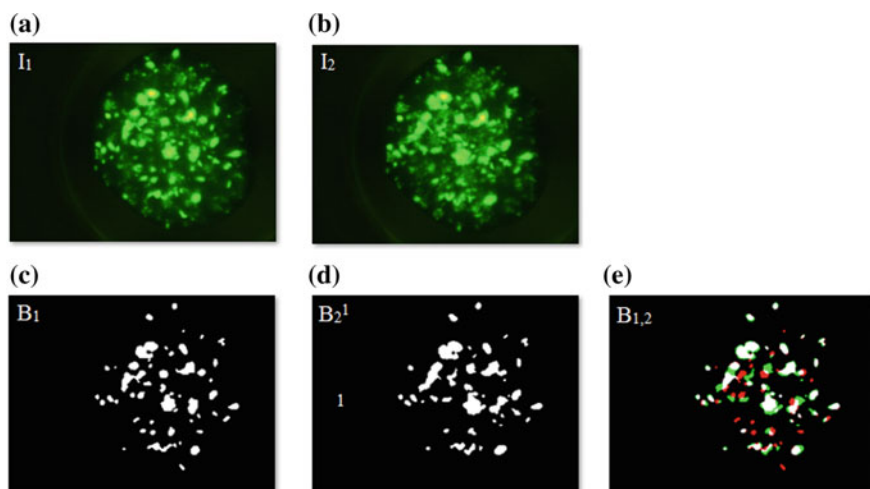


Fig. 6.9 Matching images to determine crystal growth changes, **a** original image I_1 , **b** original image I_2 , **c** binary image B_1 , **d** binary image B_2 (aligned with respect to I_1), and **e** matching B_1 and B_2^1 (matched object pixels in white, new object pixels in B_2 in green and removed object pixels shown in red). Reprinted (adapted) with permission from *Crystal Growth and Design* 2013 15 (11), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 5254–5262. Copyright (2015) American Chemical Society

Table 6.4 Prediction rules for crystal growth

Rule #	Description	Rule signifying growth
1	New crystal formation	$\bar{N} > 1$
2	Increase in crystal size	$A_d > 50\%$ AND $I_f > 50\%$

of crystals forming or disappearing. The second rule is used to identify changes in the size of crystals. This has two conditions that should be both satisfied. The first part states that the foreground area should increase by at least 50% for clear crystal growth. Similarly, the second part states that the area of the matched regions should increase by at least 50%. Any image sequence in which one of the rules given in Table 6.4 is satisfied is considered to have crystal growth and thus important from the crystallographer's point of view.

6.7 Detection of New Crystals

Formation of new crystals is an important outcome in crystallization experiments. Once this system detects formation of crystals and growing crystals, a crystallographer may start from these cases for his or her analysis. In Fig. 6.10, each column

shows sample image pairs of crystallization trial images. In the image pairs shown in Fig. 6.10a,b, no additional crystals are formed in the second image in comparison to the first image. On the other hand, Fig. 6.10c,d shows the image pairs where new crystals formed in the second image. This crystal growth analysis system can distinguish these two scenarios and also provide the count of new crystals appearing in the crystallization trials.

For the crystallization trials identified with crystals from Sect. 6.4.4, temporal analysis for new crystal formation is carried out using *Rule 1* in Table 6.4. Let I_1 , I_2 , and I_3 correspond to images collected for a crystallization experiment at 3 different time instances. For each dataset (PCP-ILOpt-11, PCP-ILOpt-12, and PCP-ILOpt-13), the temporal analysis for 3 image pairs (I_1, I_2), (I_2, I_3), and (I_1, I_3) are provided in Table 6.5. The objective is to determine the performance of the system to correctly identify new crystal formation by comparing the predicted outcome with the actual

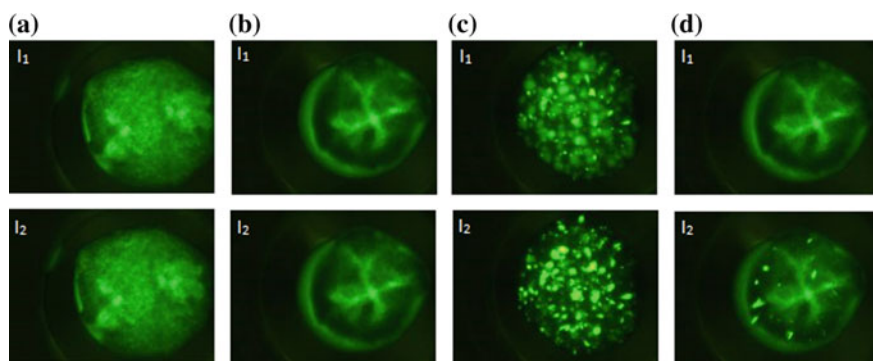


Fig. 6.10 Sample temporal trial images **a–b** Image pairs with no new crystals, and **c–d** Image pairs with new crystals. Reprinted (adapted) with permission from *Crystal Growth and Design* 2013 15 (11), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 5254–5262. Copyright (2015) American Chemical Society

Table 6.5 Experimental results for new crystals detection

Dataset	Image pair	TN	FP	FN	TP	ACC	SENS
PCP-ILOpt-11	I1, I2	7	6	0	11	0.75	1.00
	I2, I3	8	5	0	11	0.79	1.00
	I1, I3	5	7	0	12	0.71	1.00
PCP-ILOpt-12	I1, I2	5	4	0	5	0.71	1.00
	I2, I3	7	3	0	4	0.79	1.00
	I1, I3	0	6	0	8	0.57	1.00
PCP-ILOpt-13	I1, I2	3	2	0	9	0.86	1.00
	I2, I3	5	1	1	7	0.86	0.88
	I1, I3	3	1	0	10	0.93	1.00

expert score. Since this step is based on the candidate trial identification for analysis, the candidate trials include both the true positive and the false positive trials. The PCP-ILOpt-11 and PCP-ILOpt-12 datasets did not yield any false negatives. The results show that the system exhibits very high sensitivity for new crystal formation. Out of 78 total cases where new crystals are formed, 77 are identified correctly leading to satisfactory accuracy (performance) of the system.

6.8 Detection of Crystal Size Increase

Analysis of the changes in the size of crystals is another important factor for crystallographers. After identifying candidate crystallization trials for analysis, a spatiotemporal analysis between image pairs is carried out for detecting increase in size of crystals. Figure 6.11a,b shows sample image pairs with no growth in the size of crystals. On the other hand, Fig. 6.10c,d shows the image pairs where crystals grow in size in the second images. This crystal growth analysis system can detect crystal growth in size and distinguish these two cases using *Rule 2* in Sect. 6.6. The results for experiments for identifying increase in crystal size are provided in Table 6.6. The number of image pairs with observed crystal growth in size is 9, 5, and 4 for PCP-ILOpt-11, PCP-ILOpt-12, and PCP-ILOpt-13, respectively, considering pairs (I_1, I_2) , (I_2, I_3) , and (I_1, I_3) . CrystPro correctly identifies 7 out of 9 for the first dataset, 3 out of 5 for the second dataset, and 2 out of 4 for the third dataset.

Besides growth detection, CrystPro also determines other metrics such as the percentage of area change in the matched crystals. An expert may use these metrics for further analysis of the crystallization conditions.

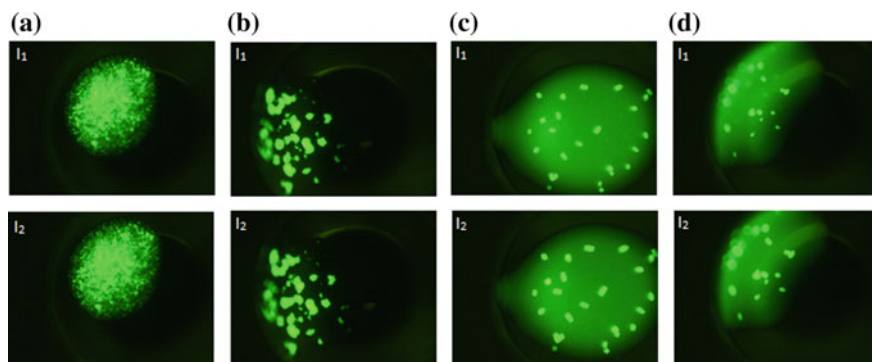


Fig. 6.11 Sample image sequences with increase in size of crystals in successive scan. Reprinted (adapted) with permission from *Crystal Growth and Design* 2013 15 (11), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 5254–5262. Copyright (2015) American Chemical Society

Table 6.6 Experiments with crystal size increase

Dataset	Image pair	TN	FP	FN	TP	ACC	SENS
PCP-ILopt-11	I1, I2	20	2	0	2	0.92	1.00
	I2, I3	18	1	2	3	0.88	0.60
	I1, I3	20	2	0	2	0.92	1.00
PCP-ILopt-12	I1, I2	10	2	0	2	0.86	1.00
	I2, I3	8	5	0	1	0.64	1.00
	I1, I3	11	1	2	0	0.79	0.00
PCP-ILopt-13	I1, I2	12	1	1	0	0.86	0
	I2, I3	12	0	1	1	0.93	0.50
	I1, I3	12	1	0	1	0.93	1.00

6.9 Discussion

6.9.1 Trace Fluorescent Labeling

Influence of Labeling on Crystallization. The trace fluorescently labeled (TFL) approach requires the covalent attachment of a fluorescent probe to a subpopulation of the protein. In the experiments, it is shown that at the trace derivatization level ($< 1\%$) the presence of the probe does not affect the nucleation process, crystal growth, or crystal diffraction resolution. [9, 22] Other advantages that accrue from this approach are that the probe wavelengths can be selected to avoid interfering substances, one can use this method with direct visualization, and one can employ more than one color of fluorescent probe for the crystallization of complexes. Trace fluorescent labeling did not have an adverse effect on spatiotemporal analysis in the experiments.

Possibility of Quenching. Protein crystallization screening involves trials over a wide range of components in chemical space. Some of these are likely to be fluorescence quenchers. While strong quenching has not been observed, it is highly likely that some may be taking place, manifested as reduced fluorescence intensity. Quenching is a collisional process, with the degree of quenching being dependent upon the strength and concentration of the quenching species and the accessibility of the fluorescent probe. Fluorescent probes become shielded from the bulk solution environment upon incorporation into a crystal. Thus, while quenchers may be active on the probe in solution they will be far less effective once the probes are buried in a crystal. Nonetheless, care should be taken when selecting a probe for TFL applications, particularly with respect to its pH sensitivity. In the experiments carried out to date no evidence of quenching has been observed.

Comparison with UV Fluorescence. Plate imaging systems primarily use visible light and UV fluorescence, which is dependent upon the protein having a tryptophan residue, which is not always present. Protein crystallization plate analysis using UV

fluorescence requires UV transmissive optics, a UV light source, a suitably UV sensitive camera, and which for safety reasons cannot be used for direct visual observation. However, the UV fluorescence produces concentration-dependent intensity, the same as for the TFL approach discussed here. Regardless, for both the UV and TFL methods, the signal of interest is the emitted intensity. Thus software methods developed for intensity-based image analysis and interpretation will be equally applicable with both UV and visible fluorescence. For this reason the CrystPro system should be equally suitable for UV fluorescence-based image analysis.

6.9.2 Spatiotemporal Analysis

Increasing accuracy versus overfitting. This system is not a supervised method, and hence is not affected by noise in expert scores. To improve the accuracy of results, one might be tempted to extract many features and train a classifier. However, it is very critical not to overfit the data by using irrelevant features to build the model. Even careful selection of image features may not be sufficient to prevent overfitting for classification models. For example, to detect new crystals, the difference in the number of crystals is the only relevant feature. Likewise, to determine the growth of crystals, the changes in the size of matched crystals is a useful feature while the difference in crystal count is not. Table 6.7 provides the classification results using decision tree classifier with the features in Table 6.3 as the input. Although this decision tree has good accuracy, the classification model is finely tuned by using irrelevant features automatically (observed after analyzing the decision tree). Such a model may work on the training set but not on new crystallization trial images. The performance of this system mentioned in the previous subsections is a better indicator of the accuracy of this system. In addition, the performance can get better as the number of simple image pairs increases or challenging image pairs are excluded.

Size of crystals. CrystPro focuses on first identifying the crystallization trials that have crystals having some minimum size. Extremely small crystals are hard to match and susceptible to matching incorrect regions due to poor thresholding. Figure 6.12a shows sample trial images where small sized crystals are grown but are missed by CrystPro. However, if crystals grow in size, such images will be detected.

New crystal formation. CrystPro detects whether new crystals form or not. Hence, as soon as the plate is set up, the plate should be scanned and images should be captured before crystals grow. If the crystals are already grown in the first image,

Table 6.7 Classification results using decision tree classifier

Dataset	Test	Accuracy	Sensitivity
PCP-ILOpt-11	New crystals	0.94	0.73
	Growth	0.97	0.83
PCP-ILOpt-12	New crystals	0.89	0.90
	Growth	0.99	0.91

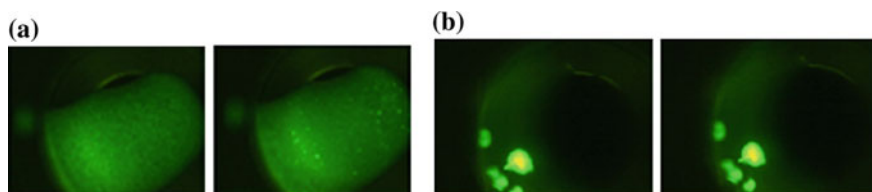


Fig. 6.12 Sample crystal images with no distinct crystal growth. Reprinted (adapted) with permission from *Crystal Growth and Design* 2013 15 (11), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 5254–5262. Copyright (2015) American Chemical Society

CrystPro does not detect new crystal formation. Figure 6.12b shows sample trial image pair where crystals are fully grown in the first image and there is no apparent growth in crystals in the second image. To use CrystPro effectively, the images should be collected as soon as the plate is set up.

Proper thresholding. Having a proper binary image is very critical in automated image analysis systems. If the images in a temporal sequence are not thresholded properly, crystal regions may not be segmented properly. Hence, it cannot be determined whether those regions are growing or not. For some images, the proposed thresholding technique does not provide the correct binary image. Varying illumination and improper focus are some of the factors for incorrect thresholding. Normalizing the illumination is not always a good idea since crystal formation is indicated by the intensity increase in the experiments. If two images are normalized to have the same illumination, crystal formation or growth may not be detected. Therefore, effective thresholding is essential for the best results.

6.10 Summary

This chapter described the CrystPro system for automated analysis of protein crystallization trial images using time sequence images. This approach involves generating proper binary images, applying image segmentation, matching the regions and then comparing the matched areas to determine the changes. Given the images of crystallization trials collected at different time instances as the input, first the candidate trials for spatiotemporal analysis are identified. Secondly, spatiotemporal analysis is done on the selected trials and changes such as new crystal formation and crystal size increase are identified. This information is helpful for crystallographers to find out important conditions for crystal growth. The system has a very high accuracy and sensitivity for detecting trials with new crystal formation. Likewise, the system exhibits reasonable accuracy and high sensitivity for crystal growth detection. By analyzing the temporal images, useful metrics are derived to help the crystallographers review the images.

The image thresholding plays a critical role for spatiotemporal analysis described in this chapter. Proper image thresholding will improve spatiotemporal analysis.

Normalization of images based on intensity causes problems in identifying growing crystals. If the images are normalized, the growing crystals may look like they shrink in time series analysis. The background area can be used for normalization rather than the complete image area. Such normalization may help generate proper thresholded images.

Acknowledgements The majority of this chapter is Reprinted (adapted) with permission from *Crystal Growth and Design* 2013 15 (11), Madhav Sigdel, Marc L. Pusey, and Ramazan S. Aygun, 5254–5262. Copyright (2015) American Chemical Society. Some modifications have been made to fit into this book.

References

1. Asanov, A. N., McDonald, H. M., Oldham, P. B., Jedrzejak, M. J., & Wilson, W. W. (2001). Intrinsic fluorescence as a potential rapid scoring tool for protein crystals. *Journal of Crystal Growth*, 232(1), 603–609.
2. Bern, M., Goldberg, D., Stevens, R. C., & Kuhn, P. (2004). Automatic classification of protein crystallization images using a curve-tracking algorithm. *Journal of applied crystallography*, 37(2), 279–287.
3. Berry, I. M., Dym, O., Esnouf, R., Harlos, K., Meged, R., Perrakis, A., et al. (2006). Spine high-throughput crystallization, crystal imaging and recognition techniques: current state, performance analysis, new technologies and future aspects. *Acta Crystallographica Section D: Biological Crystallography*, 62(10), 1137–1149.
4. Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 679–698.
5. Cumbaa, C., & Jurisica, I. (2005). Automatic classification and pattern discovery in high-throughput protein crystallization trials. *Journal of structural and functional genomics*, 6(2–3), 195–202.
6. Cumbaa, C. A., & Jurisica, I. (2010). Protein crystallization analysis on the world community grid. *Journal of structural and functional genomics*, 11(1), 61–69.
7. Cumbaa, C. A., Lauricella, A., Fehrman, N., Veatch, C., Collins, R., Luft, J., et al. (2003). Automatic classification of sub-microlitre protein-crystallization trials in 1536-well plates. *Acta Crystallographica Section D: Biological Crystallography*, 59(9), 1619–1627.
8. Dinç, I., Dinç, S., Sigdel, M., Sigdel, M. S., Pusey, M. L., & Aygün, R. S. (20014). Dt-binarize: A hybrid binarization method using decision tree for protein crystallization images. In *WORLDCOMP'14*
9. Forsythe, E., Achari, A., & Pusey, M. L. (2006). Trace fluorescent labeling for high-throughput crystallography. *Acta Crystallographica Section D*, 62(3), 339–346.
10. Forsythe, E., Achari, A., & Pusey, M. L. (2006). Trace fluorescent labeling for high-throughput crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 62(3), 339–346.
11. Groves, M. R., Mller, I. B., Kreplin, X., & Mller-Dieckmann, J. (2007). A method for the general identification of protein crystals in crystallization experiments using a noncovalent fluorescent dye. *Acta Crystallographica Section D: Biological Crystallography*, 63(4), 526–535.
12. Judge, R. A., Swift, K., & Gonzalez, C. (2005). An ultraviolet fluorescence-based method for identifying and distinguishing protein crystals. *Acta Crystallographica Section D: Biological Crystallography*, 61(1), 60–66.
13. Luft, J. R., Newman, J., & Snell, E. H. (2014). Crystallization screening: the influence of history on current practice. *Structural Biology and Crystallization Communications*, 70(7), 835–853.

14. Lukk, T., Gillilan, R. E., Szebenyi, D. M. E., & Zipfel, W. R. (2016). A visible-light-excited fluorescence method for imaging protein crystals without added dyes. *Journal of Applied Crystallography*, 49(1), 234–240.
15. Madden, J. T., DeWalt, E. L., & Simpson, G. J. (2011). Two-photon excited UV fluorescence for protein crystal detection. *Acta Crystallographica Section D: Biological Crystallography*, 67(10), 839–846.
16. MATLAB. (2013). *version 7.10.0 (R2013a)*. The MathWorks Inc., Natick, Massachusetts
17. Mele, K., Lekame, B. T., Fazio, V. J., & Newman, J. (2013). Using time-courses to enrich the information obtained from images of crystallization trials. *Crystal Growth & Design*
18. Meyer, A., Betzel, C., & Pusey, M. (2015). Latest methods of fluorescence-based protein crystal identification. *Acta Crystallographica Section F: Structural Biology Communications*, 71(2), 121–131.
19. Otsu, N. (1975). A threshold selection method from gray-level histograms. *Automatica*, 11(285–296), 23–27.
20. Pan, S., Shavit, G., Penas-Centeno, M., Xu, D.-H., Shapiro, L., Ladner, R., et al. (2006). Automated classification of protein crystallization images using support vector machines with scale-invariant texture and gabor features. *Acta Crystallographica Section D: Biological Crystallography*, 62(3), 271–279.
21. Po, M. J., & Laine, A. F. (2008). Leveraging genetic algorithm and neural network in automated protein crystal recognition. In *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, (pp. 1926–1929). IEEE
22. Pusey, M., Barcena, J., Morris, M., Singhal, A., Yuan, Q., & Ng, J. (2015). Trace fluorescent labeling for protein crystallization. *Acta Crystallographica Section F*, 71(7), 806–814.
23. Pusey, M., Barcena, J., Morris, M., Singhal, A., Yuan, Q., & Ng, J. (2015). Trace fluorescent labeling for protein crystallization. *Acta Crystallographica Section F: Structural Biology Communications*, 71(7), 806–814.
24. Saitoh, K., Kawabata, K., & Asama, H. (2006). Design of classifier to automate the evaluation of protein crystallization states. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, (pp. 1800–1805). IEEE
25. Segelke, B. W. (2001). Efficiency analysis of sampling protocols used in protein crystallization screening. *Journal of Crystal Growth*, 232(1–4), 553–562.
26. Sigdel, M., Pusey, M. L., & Aygun, R. S. (2013). Real-time protein crystallization image acquisition and classification system. *Crystal Growth and Design*, 13(7), 2728–2736.
27. Sigdel, M., Pusey, M. L., & Aygun, R. S. (2015). CrystPro: Spatiotemporal Analysis of Protein Crystallization Images. *Crystal Growth and Design*, 15(11), 5254–5262.
28. SIGDEL, M., SIGDEL, M. S., DINÇ, I., DINÇ, S., AYĞÜN, R. S., AND PUSEY, M. L. Chapter 27 - automatic classification of protein crystal images. In *In Emerging Trends in Image Processing, Computer Vision and Pattern Recognition*. Morgan Kaufmann, 2015, pp. 421–432.
29. Spraggon, G., Lesley, S. A., Kreusch, A., & Priestle, J. P. (2002). Computational analysis of crystallization trials. *Acta Crystallographica Section D: Biological Crystallography*, 58(11), 1915–1923.
30. Walker, C. G., Foadi, J., & Wilson, J. (2007). Classification of protein crystallization images using fourier descriptors. *Journal of Applied Crystallography*, 40(3), 418–426.
31. Wilson, J. (2002). Towards the automated evaluation of crystallization trials. *Acta Crystallographica Section D: Biological Crystallography*, 58(11), 1907–1914.
32. Xu, G., Chiu, C., Angelini, E.D., & Laine, A.F. An incremental and optimized learning method for the automatic classification of protein crystal images. (pp. 6526–6529)
33. Yang, X., Chen, W., Zheng, Y. F., & Jiang, T. (2006). Image-based classification for automating protein crystal identification. *Intelligent Computing in Signal Processing and Pattern Recognition* (pp. 932–937). Berlin: Springer.
34. Zhu, X., Sun, S., & Bern, M. (2004) Classification of protein crystallization imagery. In *Engineering in Medicine and Biology Society, 2004. IEMBS'04*, vol. 1, (pp. 1628–1631)
35. Zuk, W. M., & Ward, K. B. (1991). Methods of analysis of protein crystal images. *Journal of crystal growth*, 110(1), 148–155.

Chapter 7

Focal Stacking for Crystallization Microscopy

Abstract Automated image analysis of protein crystallization images is one of the important research areas. For proper analysis of the microscopic images, it is necessary to have all objects in good focus. If objects in a scene (or specimen) appear at different depths with respect to the camera's focal point, objects outside the depth of field usually appear blurred. Therefore, scientists capture a collection of images with different depths of field. Each of these images can have different objects in focus. Focal stacking is a technique of creating a single focused image from a stack of images collected with different depths of field. In this chapter, we analyze focal stacking techniques suitable for trace fluorescently labeled protein crystallization images but also applicable images captured under white light.

7.1 Introduction

Imaging technology has become a critical module of scientific analysis systems in biochemistry, physics, and space sciences. Microscopy imaging enables researchers and experts to visualize and analyze microscopic world. Although there have been significant improvements in many aspects of imaging technology, focusing on objects is still a problem for many applications. Image acquisition systems are usually equipped with a camera that can only capture objects in focus if they lay in the depth of field of a camera. To capture other objects in focus, the microscope lens can be moved up or down to update the depth of field accordingly. Changing the depth of field does not solve the problem since there is no single in-focus image that covers all objects. As such, scientists are required to analyze a series of images since each image has only a section or region in focus.

Depending on the problem domain, focusing problems are dealt with (1) by adjusting the level or focal point of the camera to generate the best in-focus image using a single depth of field, or (2) by fusing in-focus regions from multiple images that are captured with different depths of field. The first method is usually named as “auto-focusing”, while the second one is usually termed as “focal stacking” in the literature.

Auto-focusing and focal stacking methods have limitations on protein crystallization trial images. The microscopic images such as protein images may have 3D objects that can appear at different levels of a solution. If objects appear at different depths, passive auto-focusing methods that select the best image usually fail. Focal stacking algorithms may also fail due to several assumptions made while fusing images:

- (a) The contrast of a region will be higher when it is in focus with respect to when it is out of focus.
- (b) The brightness of a region is higher when it is in focus compared to when it is out of focus.

There are also a few challenges of focal stacking:

- (a) There may be discontinuities in the final image, since pixel values are obtained from a set of images.
- (b) Since images are captured at different times, the lighting conditions may change.
- (c) The size of an object when it is in focus and out of focus might be different. Typically, perspective model as in pinhole camera model is observed when regular cameras capture images. However, the fused image follows orthographic projection model.

Obtaining clear regions in images is important and necessary for image processing needed for feature extraction, classification of crystallization phases, and crystal growth analysis. This chapter provides an overview and evaluation of techniques that could be used for protein crystallization microscopy. Especially, these methods should yield in-focus images for trace fluorescently labeled images as well as images captured under traditional light sources such as white light.

7.2 Typical Viewing Area ~ 2 mm in Diameter

The crystal images shown in Chap. 2 are cropped $\sim 50\%$ from the actually recorded images. Droplet positions for 96-well SBS format plates typically hold a volume of $\sim 2\mu\text{L}$, and are on the order of 2 mm in diameter. Some plates may have larger and/or non-circular well positions. The region of interest is the drop position; for sitting drop plates, this is the well(s) adjacent to the crystallization solution reservoir. The wells may be circular or rectangular in shape, depending on the plate manufacturers design. We favor a circular well design where there are no sharp angles or corners. Inevitably, due to surface tension effects, sharp corners provide a place for liquids to pool, such that the crystallization droplets are not centered in the well.

Several characteristics are highly desirable for a microscope to be used for manually imaging crystallization plates. Dissecting-type microscopes are best for this purpose as they often have most of the desired characteristics and are often relatively low cost. Ideally, the microscope should have a zoom function, where one can go

from viewing an entire set of drops associated with a crystallization condition to a single crystallization drop. Still higher magnifications are also very useful for close examination of features of interest that cannot be resolved at lower magnification levels. An ocular with a built-in scale is useful for estimating crystal sizes. A long working distance, the greater the better, is a necessity when mounting crystals from the well for X-ray diffraction. It is also useful to have a glass platform above the microscope base to provide an air gap for cooling purposes, as the transmission lighting for many systems can cause the base to warm up, potentially leading to crystal dissolution. Having polarization capability is also very helpful, with one polarizer below and one above the sample, the upper polarizer usually being attached to the objective. Protein crystals are birefringent, and rotation of the upper polarizer will result in crystals that become brightly colored. This does not hold true for crystals in the cubic space group. Finally, it is very desirable to have a port for attaching a camera, to record images of one's crystals.

Fewer ancillary capabilities are needed for a microscope to be used for automated drop imaging. Most important of course is that the associated system be able to transport the plate to position each drop position within the viewing area. The automation of the image collection process makes it difficult for collection of "custom" views for each drop.

7.2.1 Objective Characteristics

For our microscopy system, we routinely use either a 5X ultra-long working distance (ULWD) microscope objective or a 35 mm camera lens. The ULWD objective is useful for crystallization plates or systems having an unusual geometry or depth, while the camera lens enables focusing at the lens as well as a variable aperture for controlling intensity and depth of field. All images for this volume were acquired using the camera lens. A zoom function is only useful for collection of single image, particularly since collecting well images at higher magnification means that a lot of the drop volume, and possible features of interest, may be excluded. As a result, we find that when higher magnification is needed, it is easier to switch out the objective for those few images.

7.2.2 Depth of Field

The depth of field refers to that area on the sample side of the lens that is within a nominally acceptable focus, while the term depth of focus refers to the corresponding zone on the imaging side of the lens where the image is in focus. Depth of field is the distance between the nearest and furthest objects that are in acceptably sharp focus. Several factors, some of which are (hopefully) irrelevant such as subject movement,

influence the depth of field for an image. The most important variables for our considerations are the lens focal length and the aperture or lens f-number.

7.2.3 Drop Depth and Your Crystal Probably Isn't Where You Are Looking

Increasing the magnification decreases the depth of field. This informs that routine serial image acquisition should be carried out at a magnification no greater than that which just encompasses the subject region of interest. Increasing the aperture size also decreases the depth of field, and correspondingly decreasing the aperture increases it; think pinhole camera's here. For a digital camera having a 1/2" sensor, using a 5X objective lens and a f/4 aperture, the approximate depth of field is between ~23 and 34 μm , depending upon how one estimates the circle of confusion. A reasonable estimate (obtained by inserting a mounting loop of known diameter) for a crystallization drop depth is ~150 μm . From this, we see that ~1/5 of the drop depth will be in focus.

It can reasonably be assumed that wherever one sets the focus of an automatic imaging system, the crystals will not grow within that associated depth of field. This is not a problem with larger crystals, but many times smaller crystals will grow on the surface of sitting drops, or fall to the bottom of the well, usually under a layer of precipitate. An advantage of fluorescence-based imaging methods is that when out of the immediate depth of field zone the brighter emission light from a crystal will be observed, even though the straight lines that would be used to identify it using white light may not be.

7.3 Take Multiple Images to See Through the Drop

There are several approaches to getting around the single fixed focal point problem for automated imaging. Most directly, one takes a series of images while focusing through the drop. The downside of this is that reviewing the images then takes that much more effort. In passive auto-focusing, once the series of images has been obtained, they can then be processed to select the best for the series, assuming there is a crystal present, and the rest discarded. Alternatively, in focal stacking, they can be combined by one of several methods to produce an image having an increased depth of field.

There are trade-offs with either of the above approaches. At the outset, taking multiple images requires movement along the focusing axis for the imaging system. The images must then be transferred to the controlling system for any subsequent processing. Even before the processing additional time is needed for imaging

each crystallization drop position, along with the concomitant increased wear on the focusing axis movement which must be translated up and down for each crystallization drop.

Once multiple images are acquired, they can be processed to determine which is the best, has features in the best focus, or to produce a composite image having an extended depth of field. Again, both of these post-acquisition processing steps require time. It is not likely that they can be accomplished “on the fly”, between acquisition of those for the next drop position. As a result, they will either temporarily consume memory if the processing is carried out after all images are acquired or they will slow down the overall rate of image acquisition if processing is carried out immediately after acquisition.

7.4 Auto-Focusing

Auto-focusing is a method of capturing an object of interest in focus by determining the depth of the object or by selecting the best image from a series of images with different depths of field. Auto-focusing can be categorized as active or passive depending on whether the camera position is determined ahead of time with respect to the object distance or the selection of the best image from a series of images captured at different depths of field.

7.4.1 Active Auto-Focusing

In an image acquisition system, if the system allows selection of the object of interest and determines where the camera should be positioned with respect to its distance, it is called *active auto-focusing*. An active auto-focusing system is equipped with a special hardware that helps determine the correct position of the camera lens. Stauffer [20] describes an active auto-focusing system in which a beam of modulated energy is projected toward a subject. The system captures the image using a single depth of field that is considered as the best depth of field. Bezzubik et al. [1] show how image contrast varies depending on the position of the stage relative to a microscope objective. Active auto-focusing is generally expensive as it requires expensive hardware modification. This system usually works well if there is a single object of interest.

7.4.2 Passive Auto-Focusing

An alternate to active auto-focusing is passive auto-focusing where the best-focused image is selected from a series of images captured at different depths of field. Let I represent an image set $\{I_1, I_2, I_3, I_4, \dots, I_k\}$ and $|I|$ represent the number of images

Table 7.1 Objective functions ©2016 IEEE

Name	Objective Function ($F_m(I)$)
Vollath-F4	$F_{vol4}(I) = \sum_{x=1}^{W-1} \sum_{y=1}^H I(x, y) \cdot I(x+1, y) - \sum_{x=1}^{W-2} \sum_{y=1}^H I(x, y) \cdot I(x+2, y)$
Vollath F5	$F_{vol5}(I) = \sum_{x=1}^{W-1} \sum_{y=1}^H I(x, y) \cdot I(x+1, y) - W \cdot H \cdot (\bar{I})^2$
Norm Variance	$F_{normvar}(I) = \frac{1}{WH(\bar{I})} \sum_{x=1}^W \sum_{y=1}^H [I(x, y) - \bar{I}]^2$
Laplacian	$F_{lap}(I) = \sum_{x=1}^W \sum_{y=1}^H [I(x-1, y) + I(x, y-1) + I(x+1, y) + I(x, y+1) - 4 \cdot I(x, y)]^2$

in the set I . These images are captured with varying depths of field. All images in I have size $W \times H$. The pixel at (x, y) in i th image I_i is represented as $I_i(x, y)$. In passive auto-focusing, an image is selected as the best-focused image from the image set, I . To define the best-focused image, an objective function is used to provide a value for an image according to its clarity and details. Let $F_m(I)$ be the function that measures the quality of image I using objective function m . Let I_f represent the best-focused image in I and $BF(I, F_m)$ represent the function for finding the best-focused image in I using objective measure $F_m(I)$. Then, $BF(I, F_m) = I_f$ where $F_m(I_f) = \max_{1 \leq i \leq |I|} F_m(I_i)$, $I_f \in I$, and $1 \leq f \leq |I|$.

In the literature, various quality measures have been proposed to evaluate image focus. Objective functions such as Laplacian, variance, Vollath-F4 [24], Vollath F5 [24], entropy, etc. are some basic examples of quality measures. Table 7.1 provides a list of some objective functions with their mathematical expression. Forero et al. [2] stated that objective functions like Laplacian and variance do not benefit from clear and sharp parts that appear in images. A quality measure may not help find the best-focused image for all domains. Among these quality measures, Vollath-F4 has been shown to provide satisfactory results for images for medical and biological images [8, 12, 15, 16, 22]. Mateos-perez et al. [13] evaluated autofocus algorithms and point out that Vollath-F4 and mid-frequency discrete cosine transform measures are suitable for real-time auto-focusing.

7.5 Focal Stacking

Focal stacking is a method of generating a focused image from images captured at varying depths of field by fusing in-focus areas. The objective is to generate a composite image with all regions in focus by selecting the in-focus pixels from the different image slices. Six images captured at varying depths of field using the microscopy system described in [18] are provided in Fig. 7.1. It may be observed that regions R_1 and R_2 are best focused in images I_2 and I_6 , respectively. The goal of focal stacking methods is to fuse these in-focus images from multiple images

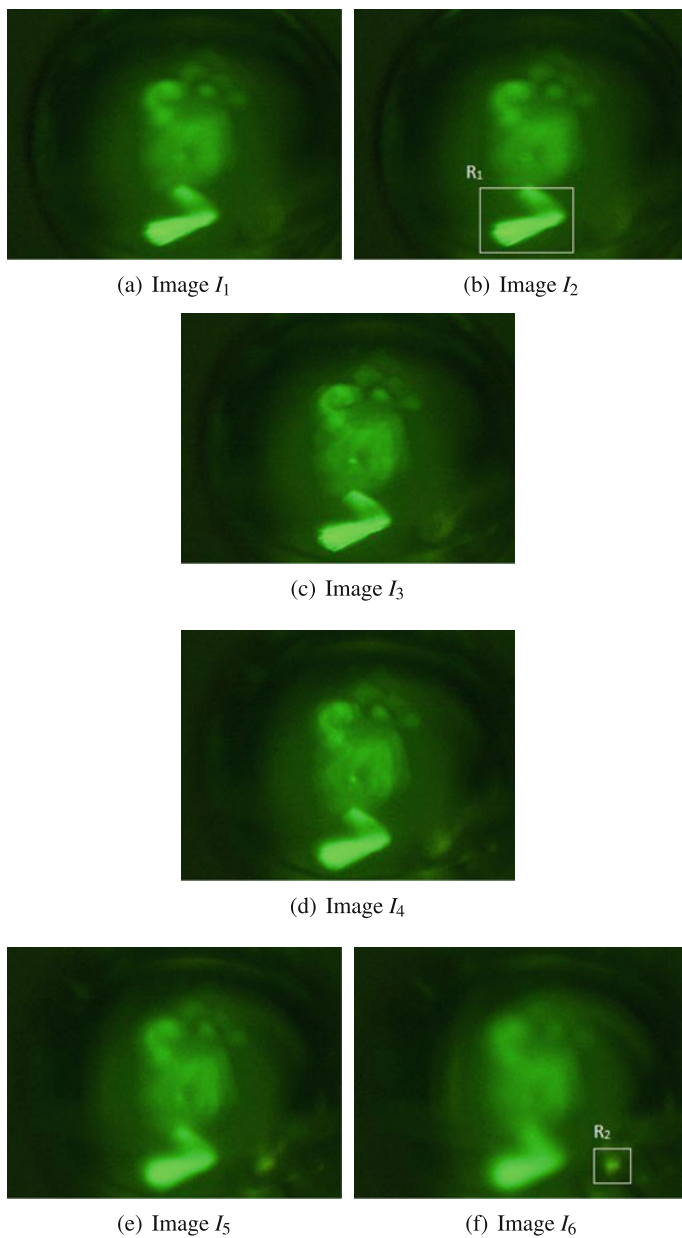


Fig. 7.1 Images of a protein crystallization sample captured with different depths of focus. Image resolution is 320x240 ©2016 IEEE

to generate a single in-focus image. Focal stacking methods can be categorized as pixel-based, neighborhood-based, and transform-based focal stacking methods.

7.5.1 Pixel-Based Focal Stacking (PBFS)

The most basic focal stacking method is the pixel-based focal stacking (PBFS) where each pixel value at the corresponding position in all images is compared to determine the best in-focus pixel value. For an input stack image set I and pixel position (x,y) , the best representative pixel value is determined using an objective function and selection criteria. Laplacian is one of the commonly used objective functions. Using a certain kernel function, Laplacian (L) value for every pixel position (x,y) is calculated. For each image $I_i \in I$, a Laplacian image L_i is created. The maximum selection criteria are then used to determine the best representative pixel for every position. At any position (x,y) , $I_f(x, y) = I_k(x, y)$ where $L_k(x, y) = \max_{1 \leq i \leq |I|} L_i(x, y)$ and $1 \leq k \leq |I|$. This method can be used with different objective functions.

7.5.2 Neighborhood-Based Focal Stacking (NBFS)

The major limitation of PBFS is the discontinuity between neighboring pixels caused by selection of pixels from different images. Neighborhood-based focal stacking (NBFS) algorithms use neighborhood information to get appropriate value of a pixel to minimize the inconsistency [4, 21]. NBFS benefits from surrounding pixels rather than solely relying on pixels on the same projection. As in PBFS, an objective function is necessary to choose the best pixel value.

7.5.3 Transformation-Based Focal Stacking

In this method, each input image in spatial domain is first transformed into another domain. The image quality and details are then compared in that domain using some objective functions and comparison methods. After determining appropriate output results, the image is re-transformed to the spatial domain by applying inverse transform. In the literature, image fusion using various transformation methods such as discrete wavelet transform, complex wavelet transform, and curvelet transform has been proposed [7, 9, 17, 23]. Forster et al. [3] proposed complex-valued wavelet transform-based image fusion algorithm. This method utilizes real and complex wavelet transforms to identify in-focus regions. The complex wavelet-based method is shown to outperform focal stacking using real-valued wavelet. One important thing to note is that there is a trade-off between capability of obtaining spatial details and

the sensitivity to noise in wavelet transform technique [11]. Image fusion algorithm by combining curvelet and wavelet transform is described in [10]. A comparative analysis of different multi-resolution transforms for image fusion has been presented in [11].

7.6 Focal Stacking for Trace Fluorescently Labeling Microscopy

Two assumptions on high intensity and high contrast for identifying in-focus regions are not applicable to trace fluorescently labeled protein crystallization trial images. Since the background area is dark, when the high-intensity regions are out of focus, they form a blurry enlarged intensity around the perimeter of the region. These artificial high-intensity regions are not part of the high-intensity region. These are rather artifacts when objects are not captured properly. As the depth of field changes, the size of the region changes as well. This section introduces an important focal stacking algorithm, FocusALL [19], which has been shown to work well for trace fluorescently labeled images. FocusALL is a neighborhood-based method and selects pixels based on the neighborhood information. Its pixel selection criteria are based on modified Harris corner response measure as described as follows.

7.6.1 Modification of Harris Corner Response Measure (HCRM)

Harris et al. [5] introduced a measure for detecting corners in an image. Harris corner method uses the principal curvatures of a two-dimensional local autocorrelation matrix based on the first derivatives of an image. Let this matrix A be represented as in Eq. 7.1:

$$A = \begin{bmatrix} S_x S_x & S_x S_y \\ S_x S_y & S_y S_y \end{bmatrix} \quad (7.1)$$

where $S_x S_x$, $S_y S_y$, and $S_x S_y$ are obtained using product of first derivatives (S_x , S_y) using a smooth circular window w such as Gaussian as follows:

$$\begin{aligned} S_x &= \left(\frac{\partial I}{\partial x} \right) \otimes w & S_y &= \left(\frac{\partial I}{\partial y} \right) \otimes w \\ S_x S_x &= \left(\frac{\partial I}{\partial x} \right)^2 \otimes w & S_y S_y &= \left(\frac{\partial I}{\partial y} \right)^2 \otimes w \\ S_x S_y &= \left(\frac{\partial I}{\partial x} \frac{\partial I}{\partial y} \right) \otimes w \end{aligned}$$

Then, Harris corner response measure at a specific pixel (x,y) is computed as in Eq. 7.2 where k is a constant:

$$M(x, y) = \text{Det}(A(x, y)) - k(\text{Trace}(A(x, y)))^2 \quad (7.2)$$

The value of $M(x,y)$ is high for corner pixels. In an out-of-focus image, pixels are smoothed by neighboring pixels. In a focused image, the variation from a pixel to its neighbor is expected to be higher than variation in defocused image. Therefore, it is reasonable to use this value as an objective function in focal stacking.

The $M(x, y)$ is actually a function of eigenvalues (α and β) of the matrix A in Eq. 7.1. These eigenvalues are correlated with the principal curvatures of the local autocorrelation function [5]. The determinant in $M(x, y)$ can be computed as $(\alpha * \beta)$, whereas the trace is equal to $(\alpha + \beta)$. The contours of $M(x, y)$ with respect to α and β are shown in Fig. 7.2a. While HCRM can differentiate corners from edges, it gives little weight to edge pixels that has strong gradient in one direction. However, in focal stacking algorithm, both corners and edges are important. If corner and edge pixels

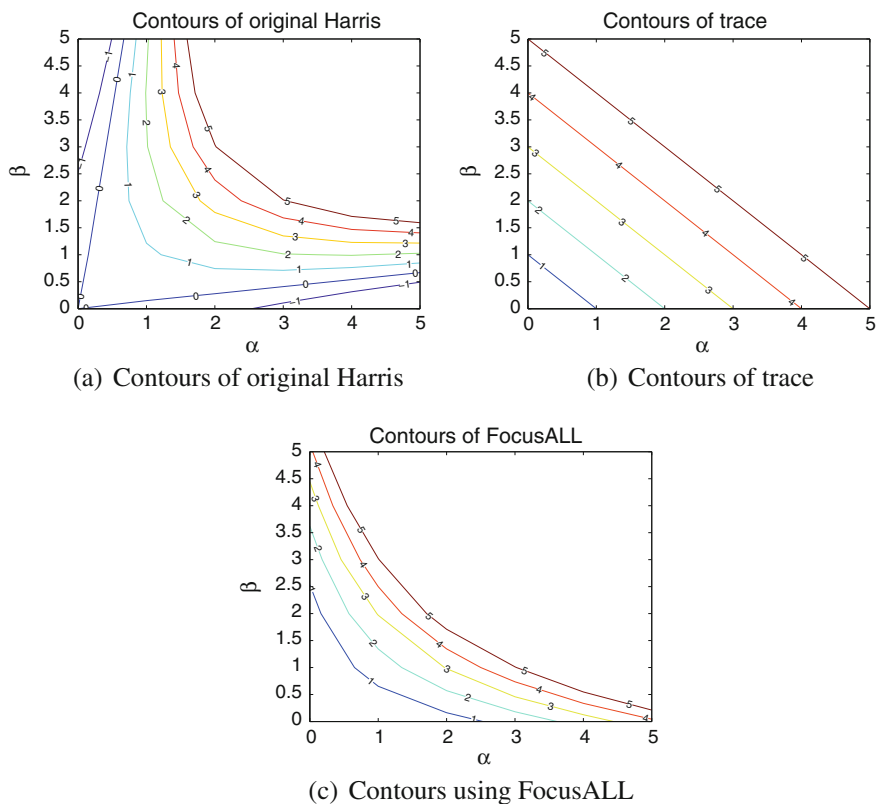


Fig. 7.2 Variation of contours with eigenvalues ©2016 IEEE

are given equal importance, $M(x, y)$ can be represented with the trace of matrix A or the summation of eigenvalues. In such a case, the contours of $M(x, y)$ would be as shown in Fig. 7.2b. However, using the trace only may give more weight to edges. To give more weight to corner pixels than edge pixels, the modified HCRM value that is given in Eq. 7.3 is used as the objective function:

$$M(x, y) = \text{Det}(A(x, y)) + k(\text{Trace}(A(x, y)))^2 \quad (7.3)$$

The contours of this proposed measure are provided in Fig. 7.2c. The curve of the contours is an indication of the emphasis on the corners pixels. A corner pixel with two low eigenvalues may be preferred to an edge with (one) high eigenvalue. FocusALL uses the modified HCRM in Eq. 7.3 as the objective function in this technique. The two major steps in FocusALL are described next.

7.6.2 Calculating Representative HCRM Value

In this step, for all images in the input stack I , HCRM value for every pixel is calculated. Then, the best representative HCRM value is determined for every pixel position. Let $M_i(x, y)$ be the HCRM value for the pixel position (x, y) of an image I_i calculated as in Eq. 7.3. Once all $M_i(x, y)$ values are calculated, maximum selection criteria are applied to determine the best representative M for every position (x, y) : $M(x, y) = \max_{1 < i \leq |I|} M_i(x, y)$. The pseudocode for this algorithm is provided in Algorithm 1. The algorithm takes image stack I as the input and returns a list with the attributes: HCRM value, image index i , and pixel position (x, y) for the best representative HCRM values for all pixel positions.

Algorithm 1 Find representative HCRM value for every position (x, y)

```

1: Input:  $I$  (Image stack)
2: Output:  $ObjList$  (Object array with attributes HCRM, imgIndx, x and y)
3:
4: procedure  $ObjList = \text{REPHCRM}(I)$ 
5:   //  $M_i(x, y)$  is HCRM at pixel  $(x, y)$  for image  $I_i$ 
6:    $i = 0$ 
7:   for  $x = 1$ ;  $x \leq I.Width$ ;  $x++$  do
8:     for  $y = 1$ ;  $y \leq I.Height$ ;  $y++$  do
9:        $Mmax = 0$ 
10:       $maxIndx = 0$ 
11:      for  $k = 1$  to  $|I|$  do
12:        if  $M_k(x, y) > Mmax$  then
13:           $Mmax = M_k(x, y)$ 
14:           $maxIndx = k$ 
15:       $ObjList[i].HCRM = Mmax$ 
16:       $ObjList[i].imgIndx = maxIndx$ 
17:       $ObjList[i].x = x$ 
18:       $ObjList[i].y = y$ 
19:       $i++$ 

```

7.6.3 Generating Focused Image

An image is generated by selecting best pixels from the images in input image stack I . First, the best representative $M(x, y)$ values obtained from the previous step are sorted in descending order based on HCRM values. To obtain the final focused image, the pixels having highest HCRM values are filled by in descending order. In addition to assigning pixel values to the corresponding pixel with high HCRM value, neighborhood pixels are also filled by analyzing the frequency of images used for that neighborhood. Let us assume that the system processes the i th highest HCRM value for the position (x_m, y_n) and its value is obtained from image I_k . Also, consider that the neighborhood window size is $dx \times dy$. To find the best pixels around (x_m, y_n) , the most frequently used image in the region $(x_m-dx/2, y_n-dy/2)$ to $(x_m+dx/2, y_n+dy/2)$ of the final focused image is first determined. In other words, the most repeatedly used image slice is found to fill the pixels around the neighborhood of (x_m, y_n) . If none of the pixels in the region is filled already, the pixel values for this region are obtained from the image slice I_k . Otherwise, the pixels values for all non-filled position in the region are filled with the pixels from mostly used image. Suppose image I_f is the most frequently selected image in this region. Then, the non-filled pixels in the region $(x_m-dx/2, y_n-dy/2)$ to $(x_m+dx/2, y_n+dy/2)$ are filled with the pixel values from I_f . This process is repeated with the next highest HCRM value until all the pixel positions are processed. At the end of the procedure, a focused image which is referred as Full Harris Image (FHI) is generated. Using the neighborhood information helps to maintain the spatial consistency.

The pseudocode for this algorithm is provided in Algorithm 2. The algorithm takes image stack (I), neighborhood size (dx, dy), and the HCRM threshold, and returns the final focused image. The HCRM threshold is used to determine the pixels to be filled in the focused image. Only the pixels having HCRM values higher than the HCRM threshold are filled on the focused image. The focused image obtained using HCRM threshold 0 is called the Full Harris Image (FHI). Using 0 as the HCRM threshold ensures that representative pixels are determined for every pixel in the focused image. Algorithm 3 provides the FindMode function to find the most repeated image within the neighborhood of a pixel coordinate and is called in Line 18 of Algorithm 2. Note that Algorithm 3 may return full Harris image (FHI) for filling regions or output partial Harris image (PHI) for varying illumination images discussed in Sect. 7.8. Figure 7.3a shows the focused image for the protein crystallization trial image set shown in Fig. 7.1. The focused image has very few discontinuities, and all the objects are in focus.

For every pixel position (x, y) , the final focused image I_f contains the pixel from an image I_i in the input stack I . Let C_i represent the color for image I_i . Depth color image can be represented as $C_I(x, y) = C_i$, if pixel (x, y) is chosen as $I_i(x, y)$ where $1 \leq i \leq |I|$. The depth color image gives an insight of the depth view of the objects.

Blue, green, red, cyan, yellow, and pink colors represent pixels selected from images $I_1, I_2, I_3, I_4, I_5,$ and $I_6,$ respectively. Fig. 7.3b shows the corresponding depth color image.

Algorithm 2 Generate final focused image

```

1: Input:  $\mathbf{I}$  (Image stack),  $(dx, dy)$  (Neighborhood size) and  $thres_{HCRM}$  (HCRM
   threshold)
2: Output:  $I_{harris}$  (FHI or PHI)
3: Note: If  $thres_{HCRM} = 0$ ,  $I_{harris}$  is FHI, else  $I_{harris}$  is PHI
4:
5: procedure  $I_{harris} = GENIMGHARRIS(\mathbf{I}, dx, dy, thres_{HCRM})$ 
6:   //  $M_i(x, y)$  is HCRM at pixel  $(x, y)$  for image  $I_i$ 
7:    $ObjList = repHCRM(\mathbf{I})$  //See Algorithm 1
8:   // Sort ObjList in descending order using HCRM value
9:    $Sort(ObjList, 'HCRM', 'Descending')$ 
10:  //Create a 2D array to keep track of selected image indices for each
   coordinate
11:   $track[ ][ ] = NULL$ 
12:  //Generate Harris image
13:  for  $i = 1$  to  $ObjList.size()$  do
14:     $x = ObjList[i].x$ 
15:     $y = ObjList[i].y$ 
16:    if (  $ObjList[i].HCRM \geq thres_{HCRM}$  ) then
17:      //Find index of most repeated image (Algorithm 3)
18:       $modeIdx = findMode(x, y, dx, dy, track, |\mathbf{I}|)$ 
19:      if (  $modeIdx$  is NULL ) then
20:         $modeIdx = ObjList[i].imgIndex$ 
21:        for  $p = -dx/2$  to  $+dx/2$  do
22:          for  $q = -dy/2$  to  $+dy/2$  do
23:            if (  $track(x + p, y + q)$  is NULL ) then
24:               $track(x + p, y + q) = modeIdx$ 
25:               $I_{harris}(x + p, y + q) = (M_{modeIdx}(x + p, y + q) > thres_{HCRM}) ? I_{modeIdx}(x + p, y +$ 
    $q):NULL$ 

```

Algorithm 3 Find mode of representative images in the neighborhood

```

1: Input:  $(x, y)$  (Pixel Position),  $(dx, dy)$  (Neighborhood size),  $track$  (2D array of
   image size that keeps track of selected image index for each coordinate,  $|\mathbf{I}|$ )
   (Number of images in image stack)
2: Output:  $modeIdx$  (Best representative image index at  $(x, y)$ )
3:
4: procedure  $modeIdx = FINDMODE(x, y, dx, dy, track, |\mathbf{I}|)$ 
5:   //Find frequency of image indices in the neighborhood
6:    $countList[ ] = NULL$ 
7:   for  $p = -dx/2$  to  $+dx/2$  do
8:     for  $q = -dy/2$  to  $+dy/2$  do
9:       (  $countList[track(x + p, y + q)]++$  )
10:  //Find the mode image indexes in the neighborhood, return NULL if all
   count is 0
11:   $maxCnt = 0$ 
12:   $modeIdx = NULL$ 
13:  for  $k = 1$  to  $|\mathbf{I}|$  do
14:    if (  $countList[k] > maxCnt$  ) then
15:       $maxCnt = count[k]$ 
16:       $modeIdx = k$ 

```

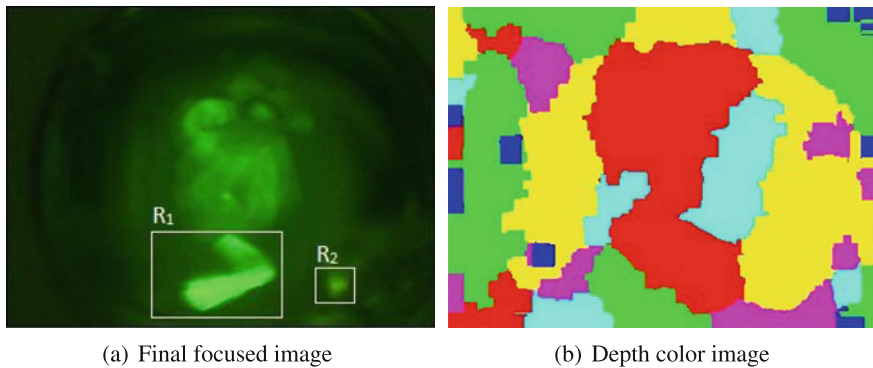


Fig. 7.3 Applying basic FocusALL ©2016 IEEE

7.7 Handling High-Resolution Images

Automatic microscopic systems generally capture images in high resolution. The experts prefer to analyze the images in their original resolution, since some information or details may be lost after resizing or processing the images. As mentioned in Sect. 7.5, focal stacking algorithms require processing every pixel in the image. Hence, applying focal stacking algorithms on high-resolution images is time consuming. In addition, since the intensity difference between neighboring pixels is low in high-resolution images, the objective function used for determining the clarity of the pixels may fail for these images.

As the resolution of an image increases, the intensity difference between two neighboring pixels decreases. Since HCRM measures the change in intensity of neighbor pixels, edges and corner pixels may not be properly detected in high-resolution images. Hence, the basic FocusALL algorithm may not generate desired focused images for high-resolution images. Figure 7.4a shows the final focused image created by using basic FocusALL with 1280x960 resolution. Two regions are highlighted and the zoomed in versions are provided in Fig. 7.4b, c, which shows discontinuities in the final focused image.

The FocusALL for high-resolution images (FocusALL-HR) is an enhanced version of the basic FocusALL technique to solve this problem. FocusALL is applied on a base low-resolution image as an initial step to obtain focused image in high resolution. The base resolution that FocusALL works properly with is determined empirically. First, the depth color map of the base resolution image is generated. Next, the depth color image is resized from base resolution to high resolution using interpolation. This step helps to generate appropriate depth color map for high-resolution image. Then, using the enlarged depth color map and image slices in high resolution, final focused image is generated. Figure 7.5a shows the focused image of the base resolution. Depth color map of the base resolution image is shown in Figs. 7.5b, c, which shows the depth color image of high-resolution image. Using the enlarged depth color map, the focused image is generated (Fig. 7.5d).

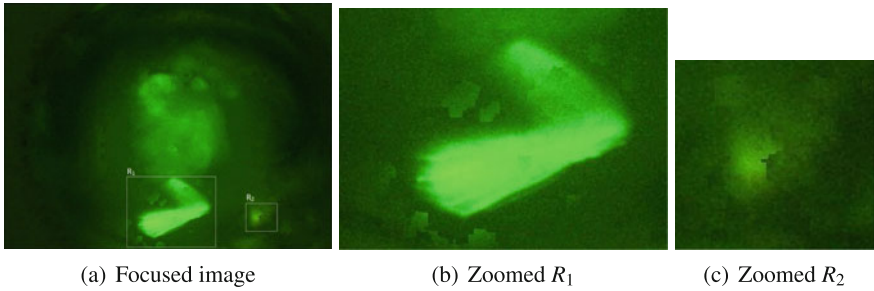


Fig. 7.4 Applying FocusALL to high-resolution image ©2016 IEEE

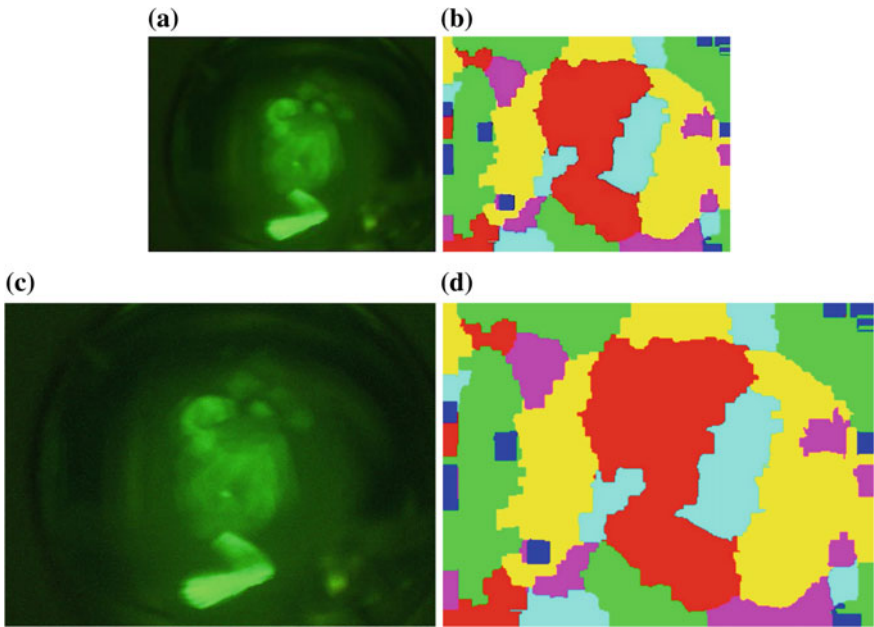


Fig. 7.5 Applying FocusALL-HR on a high-resolution image. **a** Focused image at base resolution, **b** Depth color image at base resolution, **c** Enlarged depth color image, and **d** Focused image at high resolution ©2016 IEEE

7.8 Handling Varying Illumination

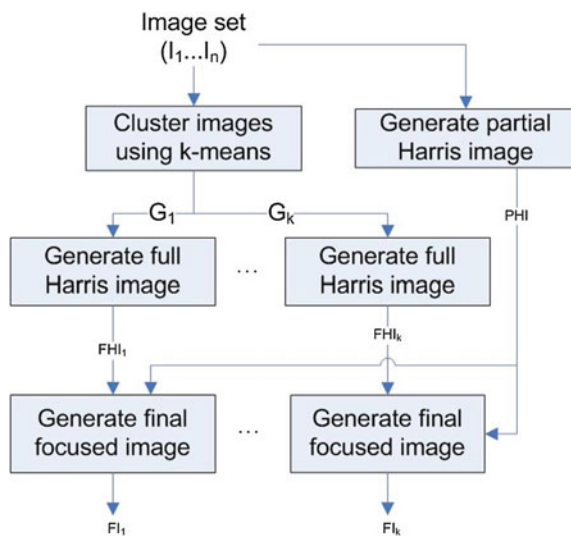
Another challenge in focal stacking is that the lighting conditions may change while capturing the images. If focal stacking is applied on such image set, high discontinuity may be observed in the focused image due to pixels picked from images with different illuminations. For such cases, basic FocusALL generates a focused image with discontinuities and artifacts.

Combining pixels from images with varying illumination to generate smooth focused image is quite challenging. Like any focal stacking algorithm, FocusALL also may not produce proper focus for these types of images. Because of illumination changes, the resulting focused image may consist of artifacts and discontinuities. Figure 7.8c shows a set of six images collected under different illuminations. Here, the top two images in Fig. 7.8c have high illumination, while the bottom four images in Fig. 7.8c have comparatively lower illumination. This may result in several discontinuities and artifacts in the background. Discontinuities are critical if they are observed inside object. FocusALL for varying illumination images (FocusALL-VI) is an enhanced version of the basic FocusALL technique to deal with such cases.

There are two aspects of varying illumination handling. First, for each cluster of images, the original FocusALL algorithm is applied and a fused image per cluster is generated. Second, a template image that separates final focused image pixels into three groups as object, background, and holes is obtained. This template image is named as partial Harris image (PHI). The holes are inside an object and filled based on the image of the closest pixel. The background pixels are filled with the full Harris images generated for each cluster. The PHI and FHI generation is very similar to each other. The main difference between them is that the PHI requires a threshold higher than 0 and generates holes. Algorithm 2 is used for PHI generation but with a threshold higher than 0.

Fig. 7.6 provides the basic flow of handling images for varying illumination. First, partial Harris image is obtained which separates the image pixels as object, background, and holes. Next, images with similar illumination are grouped under each cluster and full Harris image (FHI) is obtained from each cluster. To obtain the complete focused image, object pixels are obtained from the partial Harris image (PHI), holes are obtained using pixels from neighboring object pixel image and background

Fig. 7.6 Generating focused image for varying illumination ©2016 IEEE



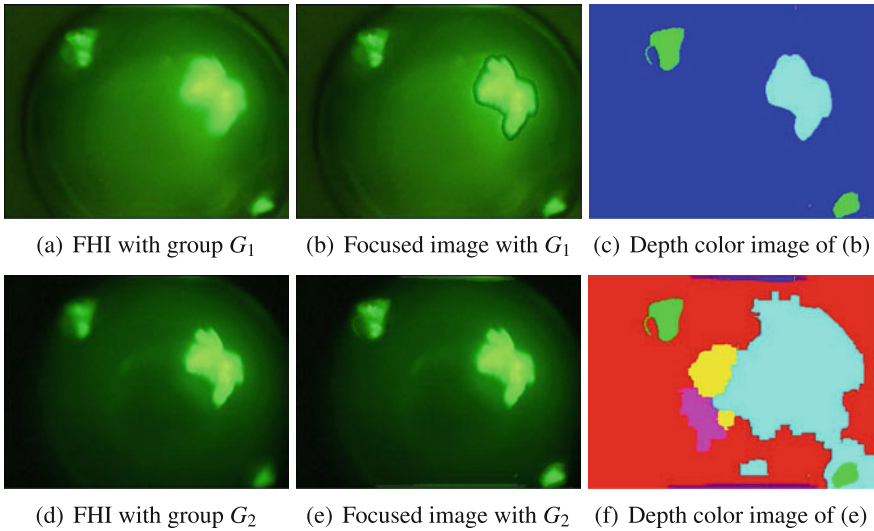


Fig. 7.7 Results of FocusALL-VI ©2016 IEEE

is filled using the full Harris image. k image clusters will yield k focused images. The expert can select one of these images as the best-focused image.

The images in image stack I are grouped using k-means clustering [6]. First, intensity histogram is obtained for each image. The intensity histogram is input to the k-means algorithm as a set of features. After providing the desired number of clusters as input to k-means clustering algorithm, the images are grouped into each cluster. For the image set shown in Fig. 7.8c, k-means clustering with intensity histogram with 25 bins and 2 clusters is applied. Using this procedure, the first two images in Fig. 7.8c fall under group G_1 , and the rest of the images in Fig. 7.8c fall under group G_2 . For each cluster of images, full Harris image is generated.

To generate the final focused image, the steps of basic FocusALL are followed with some modifications. First, the representative HCRM values are calculated. Then, the Full Harris image (focused image) is generated for each group using Algorithm 2. Here, the image stack G_1 or G_2 is the input to the FHI generating algorithm. Figures 7.7a, d show the FHI generated from group G_1 and group G_2 separately. Let the FHI generated using group G_1 be FHI_{G_1} and the FHI generated using group G_2 be FHI_{G_2} . The background regions in PHI are obtained from FHI_{G_1} or FHI_{G_2} . The holes inside objects are filled using the pixels of images of object pixels closest to the hole pixel. The detailed explanation of this method is provided in [19].

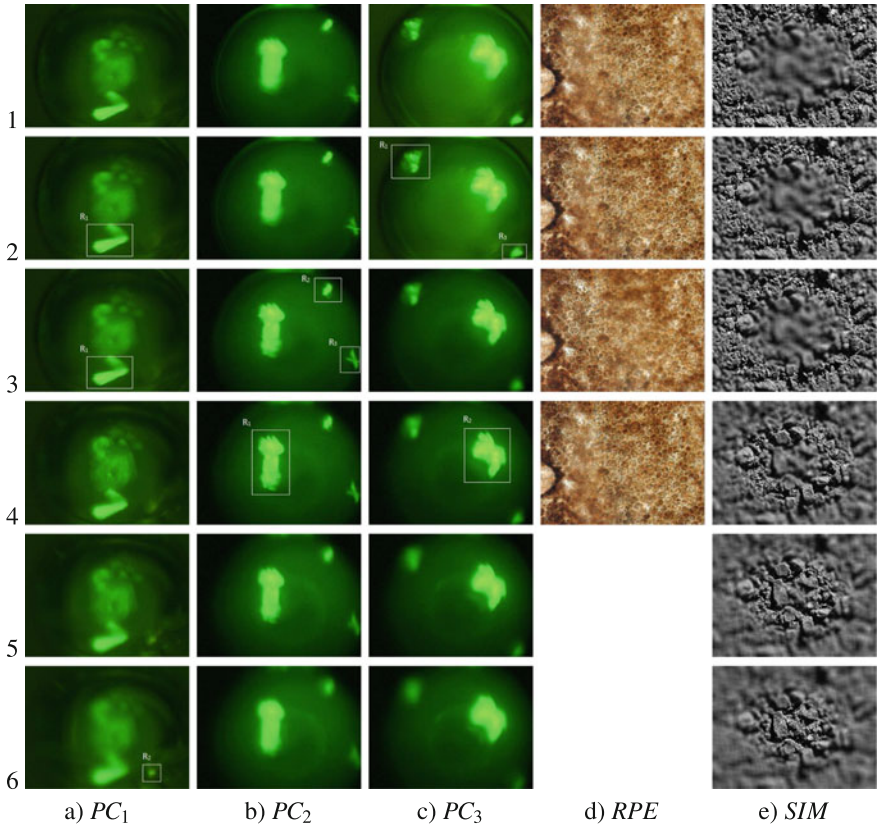


Fig. 7.8 Experimental dataset (images captured with different depths of field **a** Protein images 1 (PC_1), **b** Protein images 2 (PC_2), **c** Protein images 3 (PC_3), **d** Retinal pigment epithelial (RPE) images, and **e** Simulated texture images ©2016 IEEE

7.9 Evaluation of Focal Stacking Methods

The FocusALL algorithm has been evaluated on three protein crystallization image test cases: PC_1 , PC_2 , and PC_3 shown in Fig. 7.8. The images for protein crystallization trial sets were captured using the acquisition system described in [18]. The images are collected at a resolution 2560x1920. Each dataset consists of six images collected with different depths of field. The protein crystallization datasets used in evaluation contain random scattered noise pixels. Thus, median filtering with window size 3x3 is applied prior to using focusing algorithms. Figures 7.8a, 7.8b, and 7.8c provide the images after median filter for the test cases PC_1 , PC_2 , and PC_3 ,

respectively. FocusALL method has also been evaluated on retinal pigment epithelial (RPE) images.¹ There are four images in the RPE image set provided in Fig. 7.8d.

Simulated data with different focal depths of a microscope from a single a texture image² was created for objective evaluation of methods. Gaussian smoothing is applied for varying depth of field. The image is first mapped to 3D normal distribution model to create different focus levels for a 2D texture image. Then, using the height of each pixel as a smoothing parameter, smoothing is applied partially to different parts of the image. Figure 7.8e shows the set of six images with simulated different focal depths. The resolution of the images is 320x240.

The performance of the FocusALL technique is compared with other focusing algorithms. Since Vollath-F4 [24] has usually performed well in diverse domains, it is chosen as the objective function for the best-focused image selection method. As a transformation-based method, the complex wavelet transform (EDF-CWT) method is selected since it provided good results in fluorescence microscopy [3]. To evaluate this, the extended depth of field (EDF) plugin for ImageJ application [3] is used. In addition to the EDF-CWT method in the EDF program, the results using Sobel-based method (EDF-Sobel), variance-based method with window size 5 (EDF-Var5) and real-valued wavelet transform (EDF-RW) are evaluated. For the real wavelet method, the medium quality option is selected since it provided better result compared to the real wavelet medium high-quality option. For the FocusALL algorithm, the default neighborhood size is 15x15 pixels. HCRM threshold value is determined empirically and chosen as 20.

7.9.1 Low-Resolution Image

For low resolution, the images in Fig. 7.8 are downsampled to 320x240 and then the focusing algorithms are applied. The RPE images are of size 321x256. Figure 7.9 provides the focusing results using different techniques on four image sets (PC_1 , PC_2 , RPE, and SIM). The PC_1 image set (Fig. 7.8a) has mainly two regions of interest highlighted as region R_1 in the second and third images, and region R_2 in the sixth image. The results of focusing results for this dataset is provided in Fig. 7.9a. In other words, R_1 is best focused in the second or third image, and R_2 is the best focused in the sixth image of the set. The Vollath-F4 method selects the third image in the input set as the best-focused image. The selected image has only one region in focus and the other region is barely noticeable. The focused images using EDF-Sobel and EDF-Var5 methods introduce significant noise in the final images. Moreover, the region R_2 is not clear. The focused images using EDF-RW and EDF-CWT have both the regions in focus. However, around the borders of region R_1 , there are noise pixels and artifacts. The focused image using FocusALL has the regions of interest

¹Images obtained from http://bigwww.epfl.ch/demo/edf/demo_5.html (Courtesy of Peter Lundh von Leithner and Heba Ahmad, Institute of Ophthalmology, London).

²http://www.textureking.com/content/img/stock/big/DSC_3518.JPG.

in good focus and has a good contrast with the background. Figure 7.10 provides a zoomed in view of region R_1 from the focused images using EDF-RW, EDF-CWT, and FocusALL methods. The result from EDF-RW method shows artifacts around the region. The EDF-CWT method performed comparatively better than the EDF-RW method. However, there is random noise around the object. The R_1 region using FocusALL has smooth boundary of the object and the discontinuity is minimized.

Figure 7.9b provides the focusing results using different techniques for PC_2 (Fig. 7.8b). This image set has mainly three regions of interest represented as R_1 , R_2 , and R_3 . The region R_1 is best focused in the fourth image of the set. Similarly, regions R_2 and R_3 are best focused in the third image of the set. The Vollath-F4 method selects the third image from the set as the best-focused image. This image looks satisfactory although the edges in region R_1 are not very sharp. The focused images from EDF-Sobel and EDF-Var5 have additional layers in R_1 region. There are lots of noise pixels around the regions of interest and the objects are distorted. The EDF-RW and EDF-CWT methods perform reasonably well on this image set. However, if R_1 region is looked into closer, it is possible to see additional layers around the borders of the object. In the focused image from FocusALL method, all the regions of interest are clear. The edges of the objects are more noticeable compared to other results.

On the retinal epithelial images (Fig. 7.8d), it is difficult to select the regions of interest. Figure 7.9c provides the focusing results using different techniques. Here, the major problematic regions in the result images are highlighted. The focusing result using Vollath-F4 has the most blurred regions. The EDF-Var5 method has the best result. Other methods, EDF-Sobel, EDF-RW, EDF-CWT, and FocusALL have relatively small blurred regions. All these methods result in a good focused image compared to any single image in the input set.

On the simulated dataset (Fig. 7.8e), each image has different regions blurred. It is difficult to show the regions of interest in this set. The focusing methods can be evaluated by comparing the resulting focused images with the original texture image. Similarly, the clarity of details and overall image sharpness can be analyzed. Figure 7.9d provides the focusing results with different techniques. For each image, the problematic regions are shown in rectangular box. The outcome using best image selection method with Vollath-F4 is the most problematic. Similarly, the results with EDF-RW and EDF-CWT methods have large regions that are out of focus. The focused images with EDF-Sobel and FocusALL (neighborhood size 3x3) have small blurred regions in different parts of the images. Nevertheless, these are satisfactory results and do not affect the details in the images very much. The focusing outcome with EDF-Var5 has the least image portion that is out of focus. Therefore, variance method provides the best outcome, and the results from EDF-Sobel and FocusALL methods are of acceptable quality.

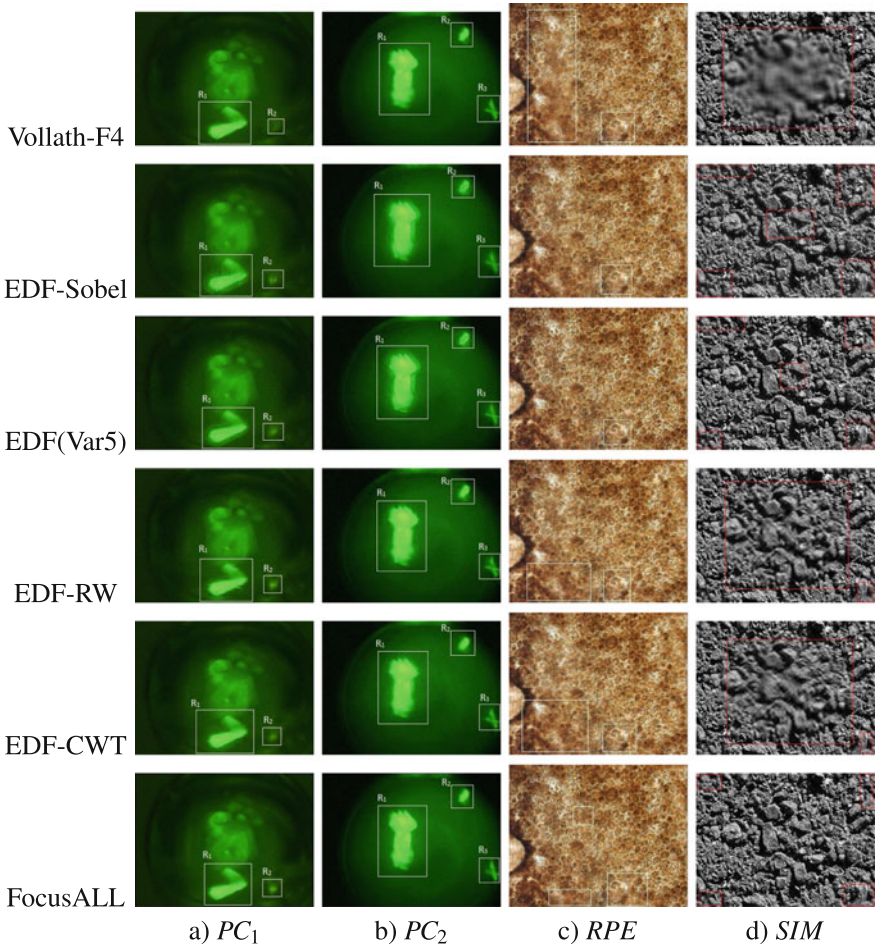


Fig. 7.9 Focusing results using different techniques **a** Protein crystallization images 1 (PC_1), **b** Protein crystallization images 2 (PC_2), **c** Retinal pigment epithelial (RPE) images, and **d** Simulated texture images ©2016 IEEE

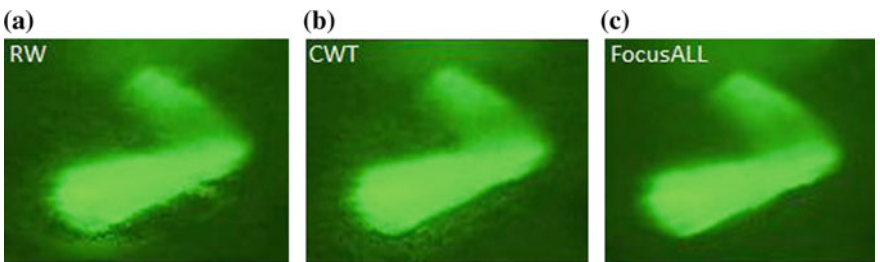


Fig. 7.10 Comparison of region R_1 in focused images on PC_1 **a** EDF-RW, **b** EDF-CWT, and **c** FocusALL ©2016 IEEE

7.9.2 High-Resolution Image

To evaluate the performance on high-resolution images, FocusALL-HR is tested on PC_1 and PC_2 image sets at 1280x960 resolution. Figure 7.11 provides the focusing results on PC_1 and PC_2 for different techniques. To highlight the problems, only the region R_1 is provided for both the image sets. Since the best image selection method does not benefit from focused regions in different image slices, the result from best image selection method is not provided. Likewise, the EDF-CWT method performed better compared to the EDF-RW method. Therefore, the result from EDF-RW is not shown. The EDF-Sobel and EDF-Var5 methods introduce significant noise around the objects. This can be observed in Fig. 7.11a, b and Fig. 7.11e, f. It is difficult to distinguish the object boundary because of several artifacts around the object. This is true for both the image sets. The results from EDF-CWT method and FocusALL-HR provide good contrast between the foreground and background. For PC_1 , the results from EDF-CWT and FocusALL are similar. On PC_2 , the EDF-CWT has some noise on the border of the object (Fig. 7.11g). FocusALL performed better on this data as the edges are clear, and the noise around the object is less. The outputs of EDF-RW, EDF-CWT, and FocusALL on low-resolution images look to be like the lower resolution of outputs generated from high-resolution images. When EDF-Var5 and EDF-Sobel are applied on a high-resolution image, it was observed that the outputs had more noise than the low-resolution outputs.

In terms of the computation time, the Vollath-F4 best image selection (Vollath-F4), Sobel-based (EDF-Sobel), variance-based (EDF-Var5), and FocusALL methods complete in similar times. On a Windows 7 Intel Core i7 CPU @2.8 GHz system with 4 GB memory, the processing time for all these methods for 1280x960 image resolution was less than 10 sec. The EDF-RW method took around 20 sec to process

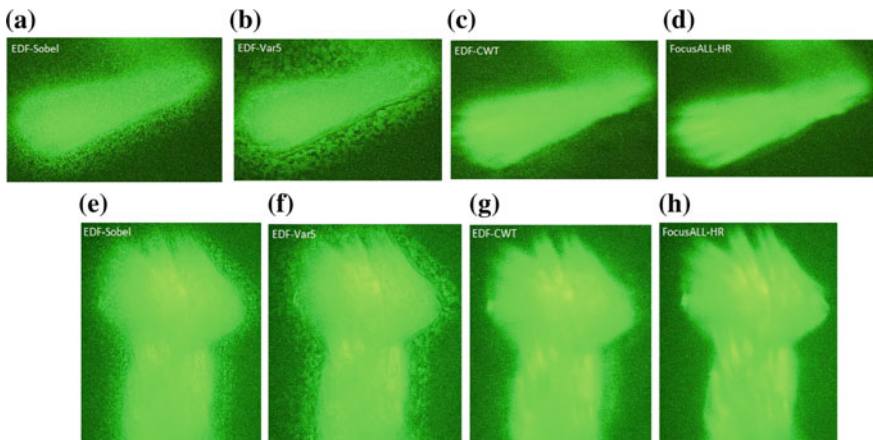


Fig. 7.11 Comparison of focusing results on high resolution **a-d** Results on region R_1 of PC_1 dataset, and **e, f** Results on region R_1 of PC_2 dataset ©2016 IEEE

the same resolution, while the EDF-CWT method took around 40 sec. As the image resolution goes higher, the computation time for the RW and CWT methods increases significantly. For image resolution 2560x1920, the CWT technique takes at least 10 mins to generate the focused image. The complexity of the FocusALL algorithm does not increase with the increase in image resolution. This is because the main processing is done in base resolution. The depth color image obtained for base resolution is enlarged to determine the pixel selection on the desired high resolution.

7.9.3 Varying Illumination Images

For varying illumination analysis, the protein crystallization image set PC_3 shown in Fig. 7.8c for evaluating algorithms is considered. This test case has three regions of interest. The image resolution is 320x240. Using the best-focused image selection method using Vollath-F4, the second image in the set (Fig. 7.12a) is selected as the best-focused image. Here, the regions R_1 and R_3 are in good focus but region R_2 could be improved if it were picked from the fourth image in the set. Using the Sobel technique, the resulting image shown in Fig. 7.12b introduces significant noise throughout the image. The focused images using the variance method (EDF-Var5) (Fig. 7.12c), real wavelet (EDF-RW) (Fig. 7.12d), and complex wavelet method (EDF-CWT) (Fig. 7.12e) all have dark regions around regions R_1 and R_3 . The problematic regions are marked by red rectangle. Using two clusters, the FocusALL-VI generated two focused images as shown in Figs. 7.12f, g. Using this method, all three regions are in good focus. The image in Fig. 7.12g looks better than the image in Fig. 7.12f since it does not have an artificial boundary around the large object region R_2 . The expert can make selection among the two images for further analysis. Experiments were also conducted on varying illumination for high-resolution images and get results similar to Fig. 7.12a–g. The region R_2 for varying illumination on high-resolution images is shown for EDF-Var5, EDF-RW, EDF-CWT, and FocusALL (from G_2 cluster) techniques in Fig. 7.12h–k. FocusALL generates sharper object regions than EDF-RW and EDF-CWT, and it does not have the noisy regions in the background as in EDF-Var5. However, FocusALL may generate artificial boundaries in the final focused image. Therefore, if the accuracy of the complete image is more critical than individual in-focus regions, EDF-RW may be preferred to FocusALL.

7.9.4 Comparison of Different Methods

In the experiments provided earlier, Vollath-F4 method picks up the overall best image from a given image set. The main problem for other methods is to pick up the best pixel for each pixel position. While CWT and RWT use wavelet coefficients, Sobel and variance use intensity change within neighborhood. The FocusALL method utilizes corner information to select the best pixel. For the discontinuity

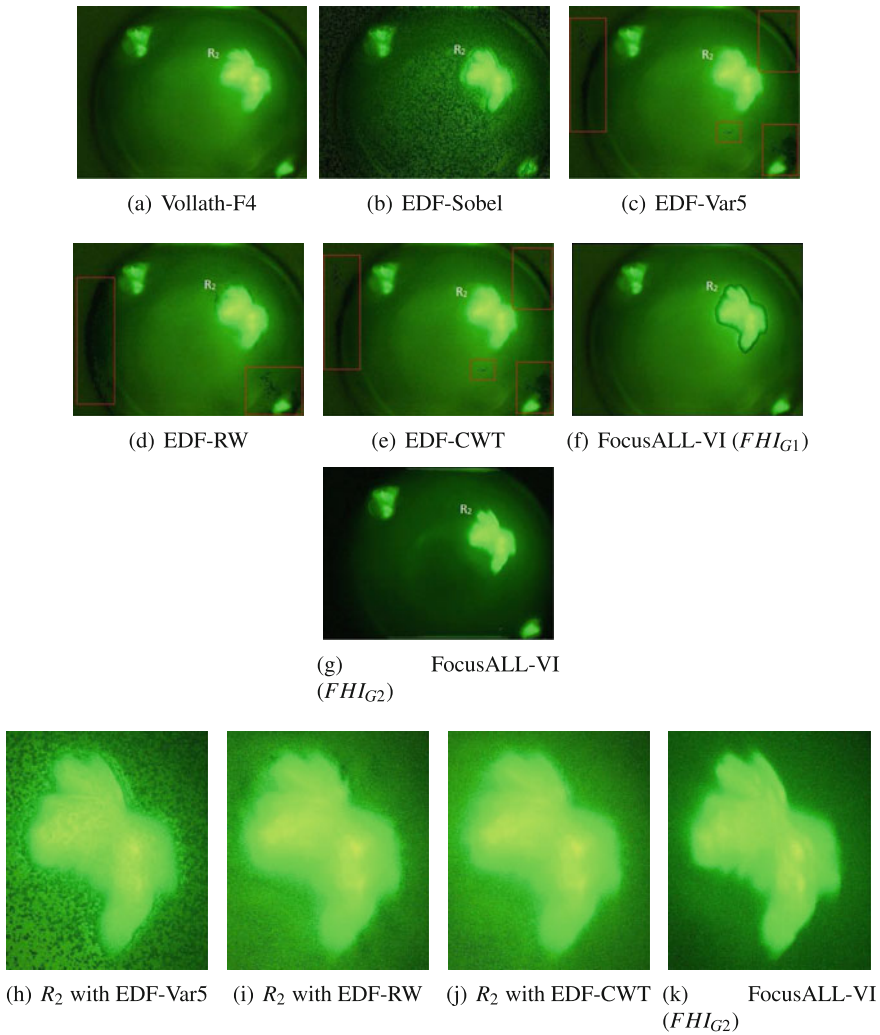


Fig. 7.12 Varying illumination results on PC_3 (Fig. 7.8c), **a–g** Results on low resolution (320x240), and **h–k** Region R_2 in high resolution (1280x960) ©2016 IEEE

problem, CWT method checks consistency in sub-bands and spatial context (3x3 neighborhood). FocusALL method uses a window to fill the regions around a corner. In addition, the window size in FocusALL is used to deal with blurriness caused by high-intensity regions. These choices are the major differences between the techniques. If a method does not perform well for a specific dataset, the pixel selection strategy and/or dealing with the discontinuity problems by that method does not work well for that dataset.

Focal stacking algorithms benefit by combining the in-focus pixels in different images to get a clear composite image. However, the focal stacking algorithms have added complexity and chances for discontinuities in the final focused image compared to the best image selection method. The Sobel, variance, real wavelet transform, and complex wavelet transform-based focal stacking available in Extended Depth of Field (EDF) [3] are evaluated in the experiments. Likewise, on the protein crystallization images, the complex wavelet transform method performed good for some images, while several discontinuities and artifacts were produced in other images. On simple images all methods perform well. However, if images have artifacts that affect the neighboring pixel values in an image, the basic methods such as Sobel and variance start to perform poorly. EDF-RWT, EDF-CWT, and FocusALL can handle image datasets with complexities due to blurring of pixels better than Sobel and variance methods. However, EDF-RWT and EDF-CWT cause an additional layer or border around the high-intensity regions.

7.10 Summary

Focusing is an important problem for protein crystallization analysis as crystals may float at different depths in a liquid solution. Due to possible presence of multiple crystals, in-focus images need to be generated for proper image analysis, feature extraction, and classification methods.

In this chapter, focusing techniques for protein crystallization microscopy are analyzed. Focal stacking techniques may yield discontinuities in the final image. Minimization of discontinuities using the neighborhood information is explained. Especially, it has been noted that two assumptions for finding in-focus regions may not be true always: a) high-contrast regions belong to in-focus regions and b) high-intensity regions belong to in-focus regions. FocusALL method could generate good in-focus images in a reasonable time (< 10 sec for high-resolution images), while some methods generate results in minutes. For varying illumination images, transform-based methods generated good results. FocusALL method yielded proper results for trace fluorescently labeled images as well as other biological and synthetic images.

Acknowledgements ©2016 IEEE. Reprinted, with permission, from M. S. Sigdel, M. Sigdel, S. Dinç, I. Dinc, M. L. Pusey and R. S. Aygün, “FocusALL: Focal Stacking of Microscopic Images Using Modified Harris Corner Response Measure,” in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 2, pp. 326–340, March-April 1 2016. doi: <https://doi.org/10.1109/TCBB.2015.2459685>

References

1. Bezzubik, V., Ustinov, S., & Belashenkov, N. (2009). Optimization of algorithms for autofocusing a digital microscope. *Journal of Optical Technology*, 76(10), 603–608.
2. Forero, M., Sroubek, F., & Cristóbal, G. (2004). Identification of tuberculosis bacteria based on shape and color. *Real-time imaging*, 10(4), 251–262.
3. Forster, B., Van De Ville, D., Berent, J., Sage, D., & Unser, M. (2004). Complex wavelets for extended depth-of-field: A new method for the fusion of multichannel microscopy images. *Microscopy Research and technique*, 65(1–2), 33–42.
4. Goldsmith, N. (2000). Deep focus; a digital image processing technique to produce improved focal depth in light microscopy. *Image Anal Stereol*, 19, 163–167.
5. Harris, C., & Stephens, M. A. (1988). combined corner and edge detector. In *Alvey vision conference* (vol. 15, p. 50). Manchester: UK,
6. Hartigan, J., & Wong, M. (1979). Algorithm as 136: A k-means clustering algorithm. *Applied statistics*, 100–108.
7. Hill, P., Canagarajah, C., & Bull, D. (2002). Image fusion using complex wavelets. In *BMVC* (pp. 1–10) Citeseer.
8. Junior, A., Costa, M., Costa F., C. F., Fujimoto, L., & Salem, J. (2010). Evaluation of autofocus functions of conventional sputum smear microscopy for tuberculosis [c]. In *IEEE International Conference on Engineering in Medicine and Biology Society (EMBS)* (pp. 3041–3044).
9. Lewis, J. (2007). O Callaghan, R., Nikolov, S., Bull, D., and Canagarajah, N. Pixel-and region-based image fusion with complex wavelets. *Information fusion*, 8(2), 119–130.
10. Li, S., & Yang, B. (2008). Multifocus image fusion by combining curvelet and wavelet transform. *Pattern Recognition Letters*, 29(9), 1295–1301.
11. Li, S., Yang, B., & Hu, J. (2011). Performance comparison of different multi-resolution transforms for image fusion. *Information Fusion*, 12(2), 74–84.
12. Liu, X., Wang, W., & Sun, Y. (2007). Dynamic evaluation of autofocusing for automated microscopic analysis of blood smear and pap smear. *Journal of microscopy*, 227(1), 15–23.
13. Mateos-Pérez, J., et al. (2012). Comparative evaluation of autofocus algorithms for a real-time system for automatic detection of mycobacterium tuberculosis. *Cytometry Part A*, 81(3), 213–221.
14. Moravec, H. (1980). *Obstacle avoidance and navigation in the real world by a seeing robot rover*. DTIC Document: Tech. rep.
15. Osibote, O., Dendere, R., Krishnan, S., & Douglas, T. (2010). Automated focusing in bright-field microscopy for tuberculosis detection. *Journal of microscopy*, 240(2), 155–163.
16. Redondo, R., et al. (2012). Autofocus evaluation for brightfield microscopy pathology. *Journal of biomedical optics*, 17(3), 0360081–0360088.
17. Shi, W., Zhu, C., Tian, Y., & Nichol, J. (2005). Wavelet-based image fusion and quality assessment. *International Journal of Applied Earth Observation and Geoinformation*, 6(3), 241–251.
18. Sigdel, M., Pusey, M., & Aygün, R. (2013). Real-time protein crystallization image acquisition and classification system. *Crystal growth & design*, 13(7), 2728–2736.
19. Sigdel, M. S., Sigdel, M., Din, S., Dinc, I., Pusey, M. L., & Aygn, R. S. (2016). Focusall: Focal stacking of microscopic images using modified harris corner response measure. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(2), 326–340.
20. Stauffer, N. (1983). Active auto focus system improvement, US Patent 4,367,027.
21. Sugimoto, A., & Ichioka, Y. (1985). Digital composition of images with increased depth of focus considering depth information. *Applied optics*, 24(14), 2076–2080.
22. Sun, Y., Duthaler, S., & Nelson, B. (2004). Autofocusing in computer microscopy: selecting the optimal focus algorithm. *Microscopy research and technique*, 65(3), 139–149.
23. Valdecasas, A., Marshall, D., Becerra, J., & Terrero, J. (2001). On the extended depth of focus algorithms for bright field microscopy. *Micron*, 32(6), 559–569.
24. Vollath, D. (1988). The influence of the scene parameters and of noise on the behaviour of automatic focusing algorithms. *Journal of microscopy*, 151(2), 133–146.

Chapter 8

Crystal Image Region Segmentation

Abstract In general, a single thresholding technique is developed or enhanced to separate foreground objects from the background for a domain of images. This idea may not generate satisfactory results for all images in a dataset, since different images may require different types of thresholding methods for proper binarization or segmentation. To overcome this problem, this chapter explains “super-thresholding” method that utilizes a supervised classifier to decide an appropriate thresholding method for a specific image. This method provides a generic framework that allows selection of the best thresholding method among different thresholding techniques that are beneficial for the problem domain. A classifier model is built using features extracted priori from the original image only or posteriori by analyzing the outputs of thresholding methods and the original image. This model is applied to identify the thresholding method for new images of the domain.

8.1 Introduction

Protein crystallization is a critical approach to understand the functionality and the structure of a particular protein [16]. The images of protein solutions are acquired and it is very important to detect well-shaped crystals since they provide important information about the structure. Since the shapes of crystals are important for determining the usability of crystals for further analysis, proper segmentation is critical. Moreover, image segmentation and thresholding may help determine the phase of a protein image in automated systems. Usually, crystal images are expected to have distinguishable features such as high intensity, sharp clear edges, and proper geometric shapes. However, in some cases, these features may not be dominant due to focusing or reflection problems even if there is a protein crystal in the image [28]. Therefore, a single type of thresholding technique may not provide an informative binary image for classifying images. Moreover, binary images may lose some important information or it may keep some unnecessary information leading to incorrect classification. For example, incorrect thresholding method may not detect a blurred crystal in an image. In [27], three thresholding techniques (Otsu’s threshold, 90th percentile green intensity threshold, and max green intensity threshold) were used

together to classify protein crystallization images not to lose any informative feature. All these binary images were used regardless whether they were proper or not. However, when features of these three binary images are included, this may also involve unnecessary features that may yield incorrect classification for some of the samples.

When each of these thresholding techniques is tried one at a time, it was noticed that there is at least one thresholding technique that works for a sample image in general. However, there is no single consistent technique that works for all images. This leads to the idea to construct a system that selects proper thresholding method for a specific sample. In this way, a protein crystallization analysis system may not be bound to limitations of a single thresholding technique.

Protein crystallization images is a challenging problem domain for thresholding due to following reasons:

1. No single thresholding technique works for all images in the protein crystallization image dataset that is evaluated,
2. Since images are collected from different phases of protein crystal growth, crystals may have varying sizes, shapes, and intensities,
3. The sizes and the number of crystals may vary,
4. Images may be captured under different illuminations, and
5. Since crystals may have 3D shapes or they may appear at different depths from the camera, some crystals may be blurred or out of focus.

This chapter explains a supervised thresholding methodology that selects the best thresholding technique for a particular image using a classifier. Super-thresholding has two different feature extraction approaches to select the thresholding method: priori and posteriori. In priori feature extraction approach, features are extracted from original images only. In posteriori feature extraction approach, firstly different thresholding methods are applied to original images. Then, the thresholded image is mapped to the original image to extract some features from foreground, background, and borders of the regions. Once the features are ready, the classifier is trained by these features to select the best thresholding method. Super-thresholding technique tries to select the most informative and reliable thresholding method for each protein crystal image. This approach provides a generic framework for a set of thresholding techniques that are suitable for the domain.

8.2 Image Binarization Methods and Limitations

There has been significant research on image thresholding (binarization) and segmentation techniques. The thresholding techniques can be roughly categorized as local thresholding and global thresholding. In global thresholding, a single threshold is used for all pixels in the image. In local thresholding, the threshold value may change based on the local spatial properties around a pixel.

Global thresholding generally depends on maximizing variances [19, 33] or entropy [5, 12, 13] between the classes and minimizing the error within the classes. However, it does not use spatial information in an image [18]. Generally, the global thresholding techniques benefit from the histogram peaks of the intensities of the image. If there are two distinct peaks in the histogram of the intensities, finding the optimal threshold value turns out to be straightforward. However, there are some cases where it is not possible to obtain two separate peaks in the histogram. In such cases, thresholding by iterative partitioning might be a good solution [25, 26].

Unlike global thresholding, local thresholding uses spatial features of a neighborhood in an image [1, 3, 17, 21, 23]. Although local thresholding techniques look more generic and superior to global thresholding, tuning parameters, partitioning the image, and the time complexity are some issues to be considered [21]. First, the parameters of non-automated local thresholding techniques are required to be set by the user for images taken under different conditions. Second issue about local thresholding is that it may classify background pixel as object pixel for poorly illuminated images, even though there is no object in the sub-image.

This chapter focuses on binarization of the crystal images, which contains 3 types of the crystal objects (as described in Chap. 2): 2D plates, small 3D crystals, and large 3D crystals. We have evaluated a number of thresholding methods. Thresholding methods such as thresholding using component tree (Silva, 2011) [30], image segmentation using double local thresholding (Chuang, 2011) [3], edge sensitive thresholding [21], thresholding based on iterative partitioning [26], Otsu's thresholding [19], and Pylon [14] neither generated proper binary images for protein images nor improved super-thresholding accuracy. Therefore, these methods were not included in the experiments for super-thresholding method. Nevertheless, individual performances of these methods are still provided in the experiments. For super-thresholding, three thresholding techniques are used: green percentile image thresholding with $p = 97$ and $p = 99$ ($g97$ and $g99$) as explained in Sect. 4.4.2 and modified Howe's method [10], which is explained next. Note that we have used $g99$ and $g100$ interchangeably in the past.

Howe et al. proposed an automated document binarization method using Laplacian energy [10, 11]. This technique tries to minimize the global energy function which depends on the Laplacian of the image as well as edge discontinuities information using Canny edge operator. Since this technique was proposed for document binarization, it is hard to get proper results without any pre- or post-processing on the image. Before this method is applied to the protein crystallization image the dataset, the samples are negated, since the images have black background. When a negative image is binarized, a frame effect is observed at the border of the image. Those artifacts are removed from binary images. Interestingly, this method produced proper binary images for 56% of the images. Figure 8.1 (j–l) shows some of the resulting binary images for this method. Since the image is reversed (or negated) and pre-processed, this adapted method is referred as ($R - Howe$) in the rest of the chapter.

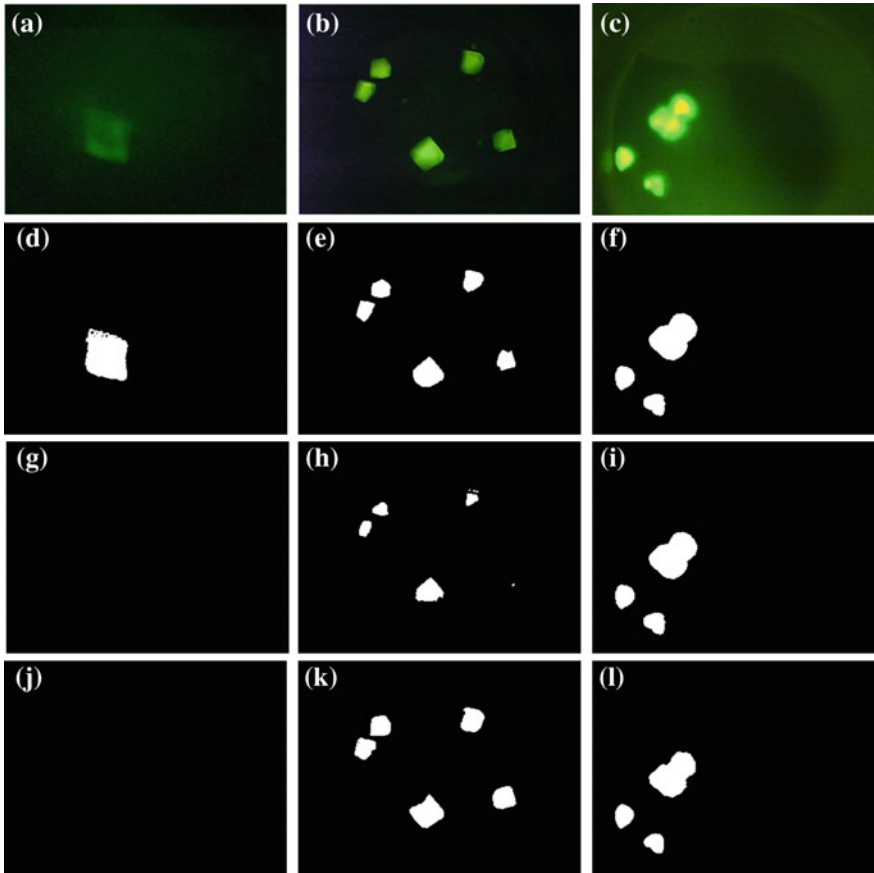


Fig. 8.1 Binarization results of different techniques: **a–c** original images, **d–f** *g97*, **g–i** *g99*, and **j–l** *R – Howe* ©2017 IEEE

8.3 Supervised Thresholding

Binarization techniques are usually constructed based on some assumptions which may or may not be suitable for every image on a dataset. Almost every thresholding technique fails under some specific circumstances, and usually there is a better alternative to that technique in the literature [24]. It is observed that some techniques may generate better results for some images while others do a better job for other images. The main goal was to exploit the powerful features of different binarization methods and use them whenever they perform well.

8.3.1 Building the Training Set

Since super-thresholding uses supervised classifiers before image binarization, a training set is needed for building a model. After running available thresholding techniques, the labeling can be done manually with the assistance of domain experts for all images in the dataset. For instance, if one protein image is binarized more accurately with the *R – Howe’s* method, that image is labeled as “1;” if the best method is *g97*, the image as is labeled as “2.” If the best method is *g99*, the image is labeled as “3.” Such a training set is satisfactory to build the model. In addition, the correct regions of the foreground are manually identified to generate the actual ground-truth binarized images. These ground-truth images are used to quantify how effective the thresholding algorithms are. Since the ground-truth images are available, the labels of images are generated automatically using the correctness measurement provided next. Note that ground-truth images are not needed to build the classifier.

8.3.2 Correctness Measurement

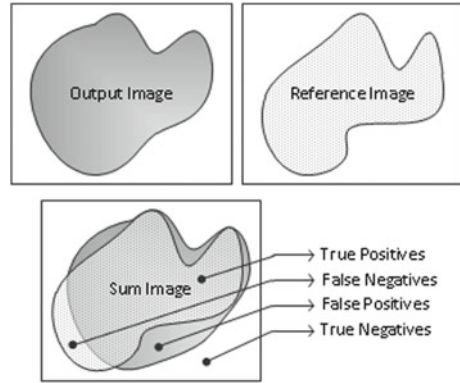
It is usually a subjective task to evaluate the results of a binarization process. Since a simple visual comparison of each binary image would not provide objective and dependable results, in this study, the reference (ground-truth) binary images are generated for all protein images in the dataset. The protein instances have been manually identified using an image editing software [31] that has the capability of auto selection of objects on the image. Once the rough object region is selected by the software, domain experts manually edit the borders for fine level corrections.

Once the reference images are ready, it is possible to calculate the correctness of any binary image by comparing with the reference image. The similarity between an output binary image (generated by a binarization method) and the corresponding reference binary image is measured using “weighted sum” of the images. Suppose the pixels of protein instances (foreground) are represented by “1,” and the background area is represented by “0” in a binary image. When the reference binary image is multiplied by 2 and added to the output binary image, the sum image that can represent all the pixels on the image as correctly classified or misclassified is obtained. Following equation shows this idea:

$$I_S = 2 \times I_R + I_O \quad (8.1)$$

where I_S , I_R , and I_O are the sum image, reference binary image, and the output binary image, respectively. The sum image includes 4 regions. These regions can easily be referred as True Positive (*TP*), False Positive (*FP*), True Negative (*TN*), and False Negative (*FN*). If the value of pixel p_{ij} on the sum image is “3,” it is a *TP* where both output image and reference image have foreground pixel. If the pixel value is “2,” it

Fig. 8.2 Sum image ©2017
IEEE



is a *FN*. Similarly, if the pixel value is “1,” it is a *FP*. Finally, if the pixel value is “0,” it is a *TN*. Figure 8.2 presents a sample sum image and its 4 regions.

TP, *TN*, *FN*, and *FP* are used to measure the correctness of an output binary image. In the literature, there are several measures that offer correctness measures from different perspectives. It is often a significant factor to select a proper measure that is more relevant to the characteristics of the problem. For example, the classical accuracy measure may not be a proper measure in this type of evaluation. Because in a typical protein binary image, there are usually very few number of foreground pixels compared to the background pixels. In other words, *TN* pixels can easily suppress the accuracy even if there are no *TP* pixels. In order to avoid bias toward a specific measurement method, 4 well-known measures are used: Accuracy (*ACC*), F-Score (*F1*) [22], Matthews’s correlation coefficient (*MCC*) [15], and Jaccard similarity (*JACC*) [2].

8.3.3 Feature Extraction

Dinc et al. [7] previously used only three thresholding techniques and one classifier (Decision Tree) using only one statistical feature. In the work presented here, more number of thresholding techniques and features are analyzed to see whether new methods and features can improve the accuracy of the results. 4 different feature sets (*FS*) are generated to test the performance of the priori approach and 1 feature set for the posteriori approach.

Table 8.1 shows brief descriptions and formulas of the features where I_{Gray} , I_{Green} , F , B , F_{in} , and F_{out} represent gray level image, green channel of original image, foreground image, background image, inner boundary image, and outer boundary image, respectively. i , j , and k represent indices of the corresponding set or image. In addition, G represents the set of connected graphs of the canny edge image, and l_i represents the length of the i th line in the set of lines, L , extracted from

Table 8.1 Definitions of features for priori and posteriori approach ©2017 IEEE

	Feature name	Description	Formulas
FS ₁	$H(X)[2]$	Measures of vertical (given formula) and horizontal autocorrelation of gray level co-occurrence matrix	$H(X) = - \sum p(x_i) \log p(x_i)$
	$\sigma(I_{Gray})$	Standard deviation of the gray level image	$\sigma(I_{Gray}) = \sqrt{\frac{\sum_{(i,j) \in I_{Gray}} (I_{Gray}(i,j) - \mu(I_{Gray}))^2}{ I_{Gray} - 1}}$
	$r_k[2]$	Measure of horizontal and vertical autocorrelation of gray level co-occurrence matrix	$r_k = \frac{\sum_{(i,j) \in I_{Gray}} (I_{Gray}(i,j) - \mu(I_{Gray}))(I_{Gray}(i-k,j) - \mu(I_{Gray}))}{\sum_{(i,j) \in I_{Gray}} (I_{Gray}(i,j) - \mu(I_{Gray}))^2}$
	\hat{L}	Sum of all edge lengths in the canny edge image	$\hat{L} = \sum_{i \in L} l_i$
FS ₂	r_k	Measure of horizontal autocorrelation of gray level co-occurrence matrix	$r_k = \frac{\sum_{(i,j) \in I_{Gray}} (I_{Gray}(i,j) - \mu(I_{Gray}))(I_{Gray}(i-k,j) - \mu(I_{Gray}))}{\sum_{(i,j) \in I_{Gray}} (I_{Gray}(i,j) - \mu(I_{Gray}))^2}$
	$\mu(I_{Gray})$	Average intensity level of the grayscale image	$\mu(I_{Gray}) = \frac{\sum_{(i,j) \in I_{Gray}} I_{Gray}(i,j)}{ I_{Gray} }$
	$\sigma(I_{Gray})$	Standard deviation of the gray level image	$\sigma(I_{Gray}) = \sqrt{\frac{\sum_{(i,j) \in I_{Gray}} (I_{Gray}(i,j) - \mu(I_{Gray}))^2}{ I_{Gray} - 1}}$
	k	Measure of peakedness of the histogram of the gray level intensity of the image	$k = \frac{\sum_{(i,j) \in I_{Gray}} (I_{Gray}(i,j) - \mu(I_{Gray}))^4}{(I_{Gray} - 1) (\sigma(I_{Gray}))^4}$
	$H(X)$	Measure of horizontal spatial disorder or spatial randomness of gray level co-occurrence matrix	$H(X) = - \sum p(x_i) \log p(x_i)$
FS ₃	$ G $	Number of connected edges (lines) in the edge image	$ G $
	\tilde{G}	Number of graphs with perpendicular edges in the canny edge image	$\tilde{G} = \sum \perp(G_k) \text{ where } \perp(G_k) = \begin{cases} 1 & \exists l_i \in L_k \text{ and } \exists l_j \in L_k \text{ and } 70 \leq \alpha(l_i, l_j) \leq 90 \\ 0 & \text{otherwise} \end{cases}$
	$\mu(L)$	Average length of all edges in the canny edge image	$\mu(L) = \frac{\sum_{i \in L} l_i}{ L }$

(continued)

Table 8.1 (continued)

	Feature name	Description	Formulas
	\hat{L}	Sum of all edge lengths in the canny edge image	$\hat{L} = \sum_{i \in L} l_i$
	\bar{G}	Sum of all edge lengths in the graphs with no perpendicular edges	$\bar{G} = \sum_{i \in L_k} l_i$ where $\perp G_k = 0$
	$\max(L)$	Length of the longest edge in the canny edge image	$\max_{1 \leq i \leq L } (l_i)$
FS_4	$H(X)[2]$	Measures of vertical (given formula) and horizontal autocorrelation of gray level co-occurrence matrix	$H(X) = - \sum p(x_i) \log p(x_i)$
	k	Measure of peakedness of the histogram of the gray level intensity of the image	$k = \frac{\sum_{(i,j) \in I_{Gray}} (I_{Gray}(i,j) - \mu(I_{Gray}))^4}{(I_{Gray} - 1) (\sigma(I_{Gray}))^4}$
	l_o	1 if $\eta_p > 0$, 0 otherwise	$l_o = \exists l_i \in L_k$ and $\exists l_j \in L_k$ and $70 \leq \alpha(l_i, l_j) \leq 90$
	η_c	Number of graphs whose edges form a cycle	$\eta_c = G_i $, where G_i is cyclic graph
	η_{hc}	Number of Harris corners	[9]
FS_5		For each binary image:	
	$\mu(F)$	Mean intensity of foreground region	$\mu(F) = \frac{\sum_{(i,j) \in F} I_{Green}(i,j)}{ F }$
	$\sigma(F)$	Standard deviation of foreground region	$\sigma(F) = \sqrt{\frac{\sum_{(i,j) \in F} (I_{Green}(i,j) - \mu(F))^2}{ F - 1}}$
	$\mu(B)$	Mean intensity of background region	$\mu(B) = \frac{\sum_{(i,j) \in B} I_{Green}(i,j)}{ B }$
	$\sigma(B)$	Standard deviation of background region	$\sigma(B) = \sqrt{\frac{\sum_{(i,j) \in B} (I_{Green}(i,j) - \mu(B))^2}{ B - 1}}$
	$\mu(F_{in})$	Mean intensity of inner pixels of the foreground region	$\mu(F_{in}) = \frac{\sum_{(i,j) \in F_{in}} I_{Green}(i,j)}{ F_{in} }$
	$\mu(F_{out})$	Mean intensity of inner pixels of the foreground region	$\mu(F_{out}) = \frac{\sum_{(i,j) \in F_{out}} I_{Green}(i,j)}{ F_{out} }$

the edge image. In the beginning, 17 histogram features [8] and 12 edge features [29] were extracted from the dataset. These features were tested and they generated satisfactory results in earlier studies [27, 29]. However, to reduce the number of features 2 feature selection methods were applied in priori approach experiments. Random Forest feature selection was used in the first 3 feature sets. The first feature set (FS_1) contains a subset of histogram and edge features. It has 1 edge feature, 4 texture features and 1 histogram feature. For FS_2 and FS_3 , 5 of the histogram features and 6 of the edge features were selected, respectively. In FS_4 , 6 of 29 combined features were selected using minimal-redundancy-maximal-relevance criterion (mRMR) feature selection method [20]. Finally, 6 statistical features were extracted for each binary image in FS_5 using posteriori approach.

8.4 Framework of Super-Thresholding

Once images are labeled, the features of the images were analyzed if there is a relationship between some features of the image and the thresholding techniques. After trying some basic features such as mean, standard deviation of intensity, autocorrelation of the images, it was noticed that some of these features could be informative to establish the relations between protein images and thresholding techniques. For instance, Dinc et al. [7] concluded that if the standard deviation of the image is less than 12.86, g_{90} thresholding method usually generates the best results. Similarly, if the standard deviation is more than 40.22, Otsu's method produces the most promising results [7].

Presence of a relation between image features and thresholding methods lead to an analysis of this relation. Thus, supervised classifiers (Bayesian classifier (*BYS*), Decision Tree (*ID3*), Random Forest (*RF*), and Artificial Neural Network classifiers (*ANN*)) were employed in order to construct a training model [32]. Since the classification process is sensitive to the factors such as data type or distribution, four classifiers having different characteristics were examined. These methods can be categorized as follows: Bayesian is a probability-based classifier, Random Forest is an ensemble classifier, Decision Tree is a rule-based classifier, and finally, Neural Networks is a powerful classifier particularly for non-linearly distributed data. The goal is to determine the one that offers the best classification results for the dataset.

Super-thresholding can binarize fast compared to complex segmentation methods. Figure 8.3 provides a general overview of super-thresholding. As shown in the figure, super-thresholding consists of four main stages: preprocessing stage, training stage, testing stage, and binarization stage. In the preprocessing stage, the dataset is labeled by an expert. Later, the dataset is divided into training and test sets. In the training stage, a classifier model is built using the features extracted from images. Feature extraction is done by two approaches called "priori" and "posteriori." Either of these approaches can be used in the feature extraction stage based on the preference.

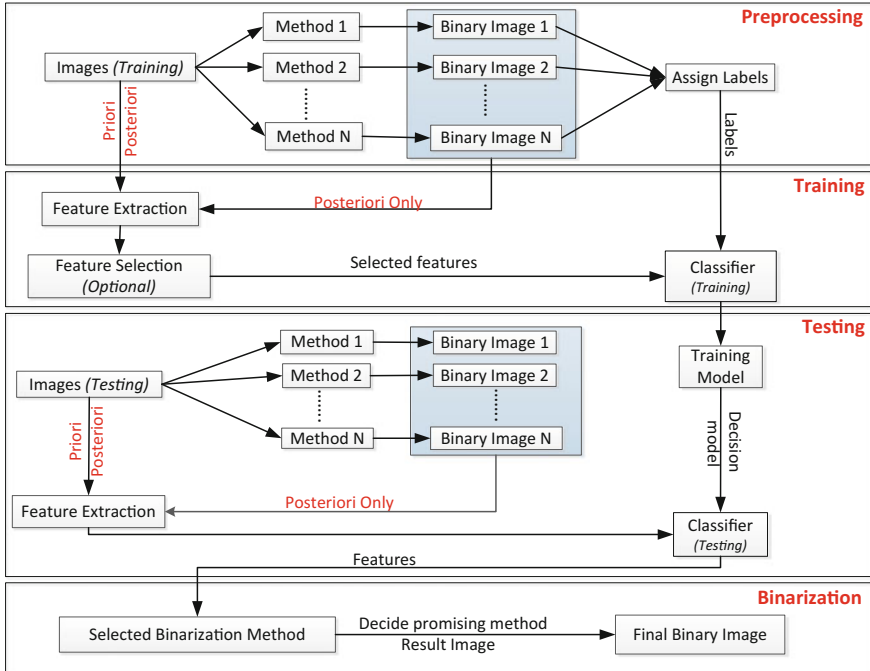


Fig. 8.3 The framework of Super-thresholding ©2017 IEEE

The classification model is trained based on the features coming from the preferred approach. Table 8.1 presents the features used in this chapter for both approaches.

8.5 Priori Approach

In the priori approach, the features are extracted from original images only. Any type of feature extracted such as the mean intensity, standard deviation, etc. from the original image can be included in this approach. This approach is relatively fast for feature extraction, since no information is extracted from the output binary images.

The priori approach just analyzes the original image without applying any thresholding method. Since no thresholding method is used in this approach, it is relatively fast.

8.6 Posteriori Approach

The posteriori approach requires running all thresholding methods to extract features. When all thresholded images are generated, they are mapped to the original images. Then foreground, background, inner and outer pixels of the object regions are detected (see Fig. 8.4). Later, a set of statistical features are extracted from these regions to feed classifiers (see FS_5 in Table 8.1). This approach is less efficient than the priori approach due to the necessity of all binary images for feature extraction, however, it can easily be parallelized, since each thresholding method can be run independently.

The main idea behind the posteriori approach is that inner and outer boundary regions can be used as an indicator whether a thresholded image is an accurate binary image or not. Normally, a significant intensity change is expected between inside and outside of the objects. Therefore, each image is both dilated and eroded using 5×5 structuring element to obtain information around the boundary pixels of the foreground as in Eqs. 8.2 and 8.3:

$$F_{out} = I_Bin \oplus S = \bigcup_{s \in S} I_Bin_s \quad (8.2)$$

$$F_{in} = I_Bin \ominus S = \bigcap_{s \in S} I_Bin_{-s} \quad (8.3)$$

where I_Bin is the input binary image and S is the structuring element. Figure 8.4f shows the total region that is of interest around the boundary.

Once features are extracted from the dataset using either the priori or posteriori approach, a classifier model is generated in training stage. The same features are used for classifying test images to determine the best thresholding method. In order to evaluate the correctness of binary images, the results are compared with ground-truth binary images generated by the research group at the University of Alabama in Huntsville. This evaluation with respect to the ground-truth binary images is explained in Sect. 8.7.

Alternatively, the binary image results could be combined or fused using a weighted sum for the final decision. However, in the experiments, it was noticed that this idea did not yield satisfactory binary images since in many cases, only one method provides the correct result while all other methods fail (see Fig. 8.5). Moreover, the way of assigning weights to each method is not obvious and it may cause biased decision toward higher weighted method even though it may fail.

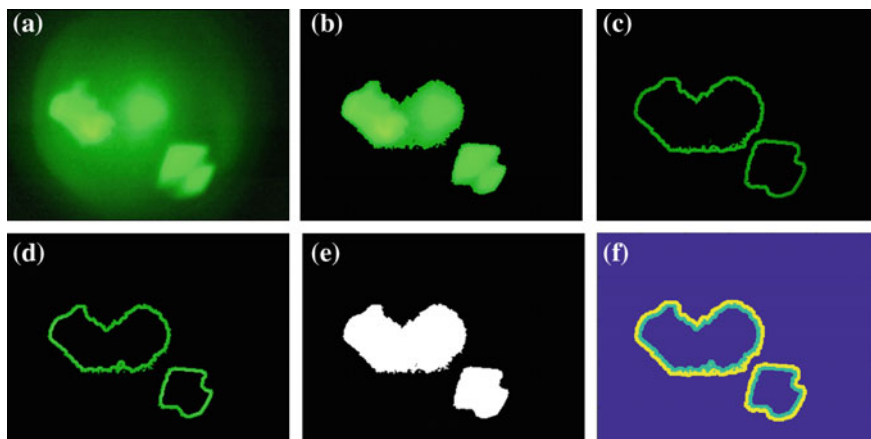


Fig. 8.4 Posteriori feature extraction: **a** original image, **b** foreground image, **c** outer pixels, **d** inner pixels, **e** thresholded image, and **f** inner and outer boundaries of foreground (**b**) ©2017 IEEE

8.7 Evaluation of Super-Thresholding

In the experiments, three different thresholding methods (g_{97} ; g_{99} ; $R - Howe$) and four classifiers (Bayesian classifier (BYS), decision tree ($ID3$), random forest (RF), and artificial neural networks (ANN)) are evaluated to binarize protein crystal images. The experiments were run using MATLAB 2014b on a 16 GB 3.4 GHz Quad-Core CPU. For random forest classifier, the source code¹ that is published by Jaiantilal et al. was used. The number of trees for random forest classifier was set as 500, and the square root of the total number of features is selected as the number of candidate features at one node of a decision tree [4]. In addition, MATLAB built-in neural network toolbox was used with two layers. The hidden layer has $n - 1$ nodes where n is number of features in the dataset. Super-thresholding technique is compared with some other thresholding methods (g_{97} ; g_{99} ; $R - Howe$; Chuang, 2011; Silva, 2011; and Otsu's method [19]).

The dataset consists of 170 protein crystal images of size 320×240 , and all images have been captured by using Crystal X2 of iXpressGenes, Inc. The dataset was labeled with 3 different thresholding techniques such that 29% of them were labeled as g_{99} , 15% of them were labeled as g_{97} , and 56% of them were labeled as $R - Howe$. In order to evaluate the size of the training set, the model is trained with 25, 50, and 75% of the data, respectively. The remaining are reserved for testing.

¹<https://code.google.com/p/randomforest-matlab/>.

8.7.1 Results

Dinc et al. [7] had a relatively small dataset and used only 3 thresholding methods (g_{90} , g_{100} , and $Otsu$). When the dataset was extended and more thresholding methods were provided to the system, the best results were obtained using 3 methods (g_{97} ; g_{99} ; $R - Howe$), and the other methods that do not contribute to the overall performance were eliminated. 5 different feature sets are generated to evaluate the performances of priori and posteriori approaches on super-thresholding. The first four feature sets (i.e., FS_1 , FS_2 , FS_3 , and FS_4) in Table 8.1 were used to test the priori approach. FS_5 was used to evaluate the posteriori approach. Visual results for 3 sample images are given in Fig. 8.5, which clearly shows the superiority of super-thresholding over other methods.

In order to evaluate the performance of the methods, a comprehensive experimental setup has been performed. The super-thresholding was tested for three different training set sizes, four correctness measures, and five feature sets. For each case, the experiments were repeated five times to avoid biased results. Table 8.3 shows the *mean* values of different correctness measures. According to the table, super-thresholding gives the best results using Bayesian classifier on feature set FS_5 (posteriori approach) regardless of the training set size. Super-thresholding achieved $ACC = 0.99$, $F1 = 0.86$, $MCC = 0.87$, and $JACC = 0.77$ on the average (highlighted bold in the table). These results are also the best results in overall experiments. With respect to the best single thresholding method ($R - Howe$), the improvements are $86.2 - 81.0\% = 5.2\%$, $86.2 - 78.6\% = 7.6\%$, and $85.5 - 75.1\% = 10.4\%$ using the $F1$ measure for training sizes of 25, 50 and 75%, respectively.

According to the results, the posteriori approach gives higher accuracy than the priori approach. The priori approach yields best results using FS_1 set. The $F1$ measures using Bayesian and random forest classifiers for FS_1 are calculated as 0.811 and 0.805, respectively. Considering the feature extraction efficiency of the priori approach, these results are also significant for real-time systems. Employing only histogram (FS_2) or edge (FS_3) features do not improve the performance significantly. Similarly, FS_4 , which is generated from both histogram and edge features using mRMR, did not improve performance as well. However, FS_4 provides very close to or slightly higher than $R - Howe$ method. To compare super-thresholding with previous study DT-Binarize [7], the experiments were repeated for 3 different training sizes. The results also show that super-thresholding following the posteriori approach outperforms DT-Binarize around 5–6% in terms of $F1$ measure. These results show that including new features, thresholding methods, and classifiers improves the binarization accuracy.

Classification Accuracy. Considering only the classification accuracy might be misleading in this problem domain. The classification accuracy is not a major indicator in this problem, since the actual labels of images are considered based on only the highest $F1$ measure. For example, for an image I , assume that $F1$ measures are $F1_{g97} = 0.865$, $F1_{g99} = 0.678$, and $F1_{R-Howe} = 0.854$. Based on this information, the actual class label of the image I will be $g97$. However, if the system selects $R - Howe$

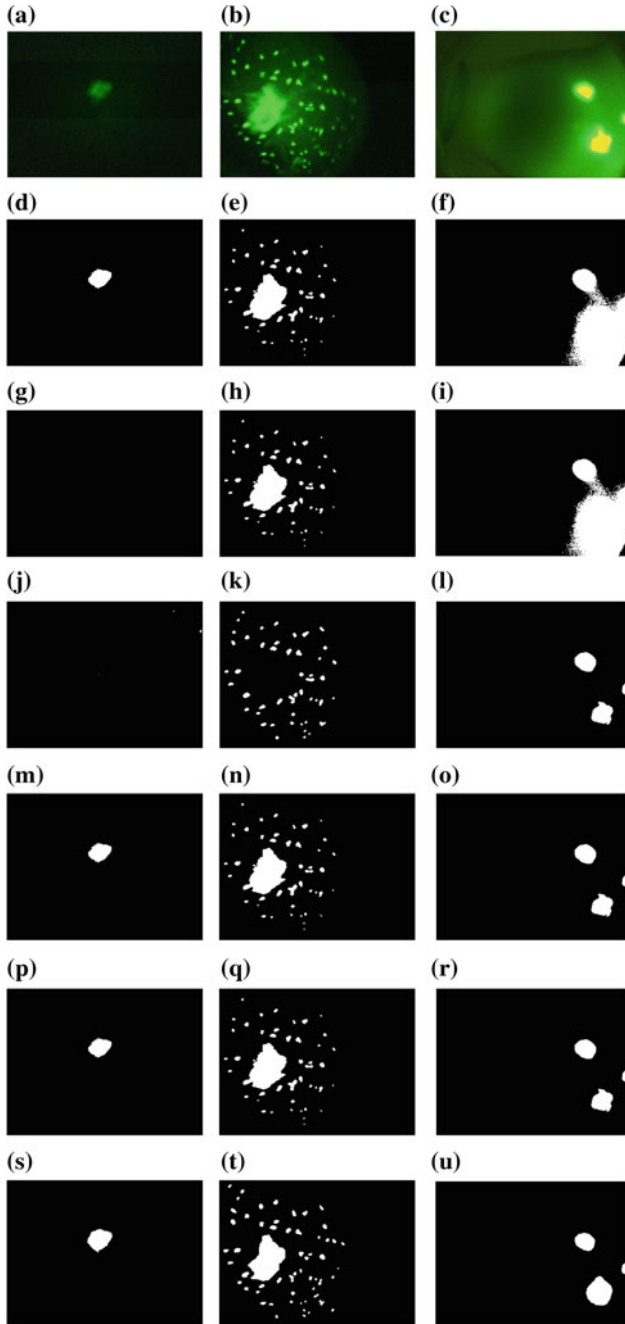


Fig. 8.5 Results of super-thresholding: **a–c** original images; **d–f** g_{97} , **g–i** g_{99} , **j–l** $R-Howe$, **m–o** super-thresholding priori, **p–r** super-thresholding posteriori, and **s–u** ground truth images ©2017 IEEE

Table 8.2 Sample confusion matrix of the experiment using FS_5 and Bayesian classifier ©2017 IEEE

		Actual		
		$G99$	$G97$	$R - Howe$
Predicted	$G99$	9	1	2
	$G97$	1	6	0
	$R - Howe$	2	1	20

method for that image, it is also acceptable in terms of thresholding. Giving higher weight to a thresholding method may not improve the accuracy as well since there are cases where one method is the only one that generates the correct binarized image. Nevertheless, we provide the confusion matrix for the Bayesian classifier on FS_5 feature set which resulted in 83.3% accuracy in Table 8.2. However, this table may not be a proper performance indicator since it does not check the completeness and not measure how good thresholding is for each image.

Soundness and Completeness. Another important issue about the binarization of protein crystal images is the soundness and completeness. It is very likely to generate improper binary images due to illumination or reflection problems. For some cases, binary images may have minor problems, which are acceptable for this problem domain unless it affects the performance of the system that will use these results. However, it is possible to have complete black or white images for some of the binarization methods if the image has a blurred or a very bright large sized object. This causes the system to miss those crystals in the analysis, which cannot be acceptable. In this context, *soundness* is related whether the output of thresholding is acceptable or not for an image. Here, soundness does not imply 100% correctness. On the other hand, *completeness* is related whether a specified method is able to generate *sound* (or acceptable) results for all images in the dataset. Here, completeness is measured as the ratio of the number of sound outputs to the number of all images. For our dataset, none of the thresholding methods have a completeness ratio of 100%. The calculations show that the completeness ratio of $R - Howe$'s method, $g97$, and $g99$ are calculated as 83, 40 and 70%, respectively. According to the results, super-thresholding gave the best accuracy, and it also did not generate any unacceptable results for the dataset with Bayesian classifier on feature set FS_5 and Bayesian classifier on feature set FS_1 as long as the problematic images (mentioned in the beginning of Sect. 8.7), which all thresholding methods failed were not in the test set. Using Bayesian classifier, super-thresholding generally generated the best results in the experiments. Moreover, super-thresholding for these sets has generated unacceptable binary images for only 4% of the dataset (when problematic images are included in the test set), while $R - Howe$'s method generated improper binary images for 21% of the dataset. As stated before, generating proper binary images is as important as the overall accuracy.

Upper Bound Performance Analysis. The performance of classification to select the best technique depends on the success of the binarization methods that are selected

Table 8.3 Correctness measure results of the experiments for each feature set and classifier ©2017 IEEE

Training size	25%				50%				75%			
	ACC	F1	MCC	JACC	ACC	F1	MCC	JACC	ACC	F1	MCC	JACC
<i>g99</i>	0.980	0.718	0.741	0.615	0.977	0.700	0.726	0.599	0.971	0.663	0.691	0.563
<i>g97</i>	0.981	0.771	0.789	0.661	0.979	0.761	0.781	0.652	0.972	0.727	0.752	0.614
<i>Otsu</i>	0.899	0.634	0.663	0.550	0.900	0.634	0.663	0.551	0.880	0.589	0.623	0.508
<i>R – Howe</i>	0.985	0.810	0.815	0.725	0.984	0.786	0.792	0.701	0.985	0.751	0.756	0.671
Silva, 2011	0.973	0.630	0.660	0.495	0.973	0.620	0.652	0.486	0.971	0.596	0.629	0.465
Chuang, 2011	0.968	0.697	0.717	0.564	0.968	0.690	0.710	0.559	0.969	0.669	0.691	0.534
Dinc et al. [7]	0.975	0.818	0.828	0.725	0.980	0.811	0.821	0.720	0.973	0.790	0.801	0.701
<i>BYS, FS₁</i>	0.985	0.825	0.832	0.735	0.985	0.832	0.838	0.740	0.986	0.811	0.817	0.720
<i>ID3, FS₁</i>	0.984	0.824	0.833	0.732	0.984	0.815	0.823	0.725	0.985	0.798	0.806	0.709
<i>RF, FS₁</i>	0.985	0.829	0.836	0.739	0.984	0.823	0.829	0.733	0.986	0.805	0.811	0.718
<i>ANN, FS₁</i>	0.981	0.766	0.786	0.662	0.978	0.726	0.749	0.625	0.972	0.706	0.730	0.606
<i>BYS, FS₂</i>	0.985	0.824	0.830	0.736	0.985	0.833	0.838	0.743	0.986	0.812	0.817	0.723
<i>ID3, FS₂</i>	0.985	0.820	0.827	0.727	0.982	0.793	0.801	0.704	0.983	0.762	0.769	0.677
<i>RF, FS₂</i>	0.985	0.833	0.839	0.744	0.984	0.823	0.829	0.736	0.985	0.774	0.778	0.691
<i>ANN, FS₂</i>	0.981	0.781	0.797	0.679	0.979	0.757	0.774	0.659	0.972	0.709	0.732	0.609
<i>BYS, FS₃</i>	0.985	0.802	0.812	0.712	0.983	0.796	0.805	0.708	0.984	0.768	0.778	0.677
<i>ID3, FS₃</i>	0.982	0.786	0.797	0.695	0.981	0.766	0.778	0.674	0.983	0.737	0.747	0.650
<i>RF, FS₃</i>	0.984	0.799	0.808	0.711	0.982	0.770	0.780	0.682	0.984	0.736	0.746	0.652
<i>ANN, FS₃</i>	0.982	0.741	0.761	0.638	0.978	0.714	0.737	0.612	0.972	0.688	0.713	0.588
<i>BYS, FS₄</i>	0.985	0.811	0.818	0.723	0.984	0.803	0.808	0.718	0.985	0.765	0.770	0.684
<i>ID3, FS₄</i>	0.984	0.808	0.815	0.717	0.987	0.800	0.808	0.711	0.984	0.767	0.775	0.678
<i>RF, FS₄</i>	0.985	0.815	0.821	0.729	0.984	0.800	0.806	0.714	0.985	0.750	0.757	0.669
<i>ANN, FS₄</i>	0.981	0.750	0.768	0.650	0.978	0.729	0.748	0.630	0.975	0.688	0.708	0.595
<i>BYS, FS₅</i>	0.992	0.862	0.867	0.774	0.992	0.862	0.866	0.774	0.991	0.855	0.859	0.765
<i>ID3, FS₅</i>	0.987	0.833	0.841	0.744	0.989	0.840	0.845	0.751	0.985	0.807	0.812	0.721
<i>RF, FS₄</i>	0.988	0.845	0.850	0.758	0.985	0.842	0.847	0.756	0.986	0.824	0.828	0.740
<i>ANN, FS₅</i>	0.981	0.768	0.786	0.664	0.977	0.772	0.790	0.671	0.973	0.743	0.765	0.639
Max-Limit	0.993	0.888	0.890	0.809	0.993	0.884	0.886	0.804	0.992	0.870	0.873	0.786

for the problem domain. This means that there is a practical limit of the performance of super-thresholding. In other words, if none of the selected methods are able to generate a proper binary image for a specific image, super-thresholding does not produce an accurate binary image, as well. Figure 8.5 shows sample cases where each method fails. The upper bound was computed by selecting the best three thresholding methods for each image and compared with the results. In Table 8.3, the last row shows the upper bound for each correctness measure. Correctness measures of the upper bound are calculated using the best binarization method for all images. Results of super-thresholding are within 97.3% ($0.765 \div 0.786$) of the upper bound for Bayesian classifier using 75% of training data with respect to the Jaccard coefficient.

Table 8.4 Timings of feature extraction, classification, and binarization methods^a ©2017 IEEE

Category	Method	Time per image (ms)
Binarization	<i>g99</i>	110.500
	<i>g97</i>	108.900
	Otsu	12.400
	<i>R – Howe</i>	130.000
	Silva, 2011	25.000
	Chuang, 2011	83.000
Testing	<i>BYS</i>	0.097
	<i>ID3</i>	0.006
	<i>RF</i>	0.051
	<i>ANN</i>	0.005
Feature Extraction	<i>FS</i> ₁	48.800
	<i>FS</i> ₂	3.190
	<i>FS</i> ₃	399.900
	<i>FS</i> ₄	443.800
	<i>FS</i> ₅	35.700

^aThe total running time of an experiment is calculated by adding the times of feature extraction, testing, and binarization stages. For example, in priori approach, if the selected method is *R – Howe* using *BYS* on *FS*₂, the total time of binarization for an image will be $130 + 0.097 + 3.190 = 133.287$ milliseconds. However, in posteriori approach, the total time of the binarization will be $110.5 + 108.9 + 130 + 0.097 + 35.7 = 385.197$ milliseconds using *BYS* on *FS*₅. Please note that in posteriori approach features are extracted using the output of all thresholding methods

Time Analysis. The run-time performance of super-thresholding has also been evaluated on a 3.40GHz Intel i7 Quad Core 16GB RAM system using 320×240 images. Table 8.4 provides the timings of feature extraction, classification, and binarization for an image in milliseconds. According to the table, the feature sets having more edge features take more time than the others (i.e., *FS*₃ consists of only edge features). Once the classifier model is built, an image can be binarized in 133 milliseconds using *BYS* on *FS*₂ (the priori approach), and in 385 ms using *BYS* on *FS*₅ (the posteriori approach), and these timings are feasible for Crystal X2 system developed at iXpressGenes, Inc.

8.7.2 Discussion

Performance of Thresholding Methods. A proper thresholding method is needed for each image. If *g97* and *g99* methods are compared, *g99* works better when the foreground is separated better than the background. In protein crystallization images, protein crystal regions are expected to have the highest intensity. Whenever the protein crystal regions have higher intensity than other regions, *g99* works fine. Large 3D crystals are usually distinguishable in terms of intensity and have higher

intensity than other regions. $g99$ works best for images containing large 3D crystals. Since crystals float in a solution, the depth of crystals from the microscope may differ. Only crystals at the depth-of-field appear in focus. Other crystals may be blurred and may have lesser intensity than crystals in focus. In those cases, $g99$ may not provide good binarization. Whenever the foreground intensity is not high, the sizes of crystals are smaller, and crystals appear at different depths, the $g97$ method is likely to perform better than the $g99$ method. R-Howe's method has three components: minimizing global energy for labeling pixels, use of Laplacian to distinguish ink from the background, and use of edge detection to handle discontinuities. The edges are critical factors on separation of crystals. The straight boundaries of crystal regions are one of the important indicators for a crystal. For regions with clear boundaries, *R – Howe* generally provides better results. If the intensity is lower or image is blurred, $g97$ may be preferred. The advantage of $g99$ is that it can easily remove the background since any pixel with low intensity is considered as the background.

Performance of Feature Sets. The feature sets for FS_1 , FS_2 , FS_3 , and FS_4 are used for the priori approach. FS_3 contains mostly edge related features and performed worst among these feature sets. Relying only edge related features is not satisfactory for this domain. FS_1 and FS_2 containing texture-related features perform similarly due to the similarity between feature sets. FS_2 slightly outperforms FS_1 . Note that FS_2 has histogram related feature and does not have edge related features. This difference between FS_1 and FS_2 has a positive impact on the accuracy for FS_2 . FS_4 was generated using mRMR feature selection method. Although FS_4 performs better than FS_3 , it does not perform as well as FS_1 and FS_2 . It looks like features based on intensity statistics is important for the accuracy. The feature set for the posteriori approach performs best among all feature sets. Although FS_5 relies on intensity features, it performed better than any other feature set. Comparison of pixels in the foreground and background as well as comparing pixels at the boundaries of regions are better features for analyzing the performance of thresholding methods.

8.8 Summary

This chapter presented a generic framework to combine different kinds of thresholding techniques using a supervised classifier. A classifier model is constructed using some image features such as autocorrelation, standard deviation, etc. of the protein crystallization images that are labeled by the experts. The labels (or classes) of images correspond to a binarization method which is proper for the image. A binarization method is selected for a given test image using the same classifier, and the selected method is applied to the protein crystallization image to generate a binary image.

Several results are concluded at the end of this study:

1. Single thresholding techniques may not be enough for some of the datasets that have poorly illuminated, noisy and unfocused images,

2. Super-thresholding can be considered the best in terms of soundness and completeness since it generated more proper binary images for protein crystal images than any other method,
3. The success of super-thresholding depends on the success of the thresholding techniques which are selected for the problem domain, and its success also depends on performance of the classifier,
4. Since super-thresholding produces single binary image using only a few simple features of the images for the priori approach, it is feasible for most of real-time classification systems, and
5. The posteriori feature extraction approach of super-thresholding can be easily parallelized since each thresholded image can be generated independently.

It is difficult to generalize or verify the soundness and completeness based on the algorithmic approaches involved in developing the thresholding methods. Expert opinion is usually needed to determine the correctness (or soundness) of a thresholding method. When thresholding techniques are used in automatic analysis systems, incorrect thresholding may lead improper decision making. Therefore, completeness is a critical factor in this domain. Another issue is regarding the choice of the best thresholding method. When building the training set, a number of methods generated good results for a specific image. In those cases, the best one is again selected using the ground-truth images although the second-best is as good as the first one. This significant similarity between methods for some images makes the training difficult. This is the reason why we believe the optimal model is not reached.

Additional features can be extracted by comparing the output binary images and the original images. These features can be used to build a more advanced model to build a classifier and checked how much they improve the accuracy.

Acknowledgements ©2017 IEEE. Reprinted with Permission, from I. Dinç, S. Dinç, M. Sigdel, M. S. Sigdel, M. L. Pusey; R. S. Aygün, “Super-thresholding: Supervised Thresholding of Protein Crystal Images,” in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 4, pp. 986–998, July–Aug. 1 2017. doi: <https://doi.org/10.1109/TCBB.2016.2542811>.

References

1. Chang, S.G., Yu, B., & Vetterli, M. (2000). Spatially adaptive wavelet thresholding with context modeling for image denoising. *IEEE Transactions on Image Processing*, 9(9), 1522–1531.
2. Chowdhury, G. (2010). *Introduction to modern information retrieval*. Facet publishing.
3. Chuang, M.-C., Hwang, J.-N., Williams, K., and Towler, R. (2011). Automatic fish segmentation via double local thresholding for trawl-based underwater camera systems. In *Image Processing (ICIP), 18th IEEE International Conference on 2011* (pp. 3145–3148).
4. Cumbaa, C.A., & Jurisica, I. (2010). Protein crystallization analysis on the world community grid. *Journal of Structural Function Genomics*, 11(1), 61–9.
5. de Albuquerque, M. P., Esquef, I., Mello, A. G., & de Albuquerque, M. P. (2004). Image thresholding using tsallis entropy. *Pattern Recognition Letters*, 25(9), 1059–1065.

6. Dinc, I., Dinc, S., Sigdel, M., Sigdel, M., Pusey, M.L., & Aygun, R.S.(2017). Super-thresholding: Supervised thresholding of protein crystal images. *IEEE/ACM Transactions on Computational Biology and Bioinformatics PP*, 99, 1–1.
7. Dinç, I., Dinç, S., Sigdel, M., Sigdel, M.S., Pusey, M.L., & Aygün, R. S.(2014). Dt-binarize: A hybrid binarization method using decision tree for protein crystallization images. In *Proceedings of The 2014 International Conference on Image Processing, Computer Vision & Pattern Recognition, IPCV'14*, pp. 304–311.
8. Haralick, R. M., Shanmugam, K., & Dinstein, I. H. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 6, 610–621.
9. Harris, C., & Stephens, M.(1988). A combined corner and edge detector. In *Alvey vision conference*, Vol. 15, Citeseer, p. 50.
10. Howe, N.(2011) A laplacian energy for document binarization. In *Document Analysis and Recognition (ICDAR), International Conference on 2011* pp. 6–10.
11. Howe, N. R. (2013). Document binarization with automatic parameter tuning. *International Journal on Document Analysis and Recognition (IJ DAR)*, 16(3), 247–258.
12. Johannsen, G., & Bille, J. (1982). A threshold selection method using information measures. *In ICPR*, 82, 140–143.
13. Kapur, J., Sahoo, P., & Wong, A. (1985). A new method for gray-level picture thresholding using the entropy of the histogram. *Computer Vision, Graphics, and Image Processing*, 29(3), 273–285.
14. Lempitsky, V., Vedaldi, A., and Zisserman, A.(2011). Pylon model for semantic segmentation. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems 24*, Curran Associates, Inc., pp. 1485–1493.
15. Matthews, B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2), 442–451.
16. McPherson, A., & Gavira, J. A. (2014). Introduction to protein crystallization. *Acta Crystallographica Section F: Structural Biology Communications*, 70(1), 2–20.
17. Niblack, W. (1985). *An introduction to digital image processing*. Strandberg Publishing Company.
18. Oh, W., & Lindquist, W. (1999). Image thresholding by indicator kriging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(7), 590–602.
19. Otsu, N. (1975). A threshold selection method from gray-level histograms. *Automatica*, 11(285–296), 23–27.
20. Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), 1226–1238.
21. Ray, N., & Saha, B.(2007). Edge sensitive variational image thresholding. In *IEEE International Conference on Image Processing, ICIP 2007 (Sept 2007)*, Vol. 6, pp. VI – 37–VI – 40.
22. Sasaki, Y.(2007). The truth of the f-measure. *Teach Tutor mater*, pp. 1–5.
23. Sauvola, J., & Pietikäinen, M. (2000). Adaptive document image binarization. *Pattern recognition*, 33(2), 225–236.
24. Sezgin, M., & Sankur, B. (2004). Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, 13(1), 146–168.
25. Shaikh, S., Maiti, A., & Chaki, N.(2011). Image binarization using iterative partitioning: A global thresholding approach. In *Recent Trends in Information Systems (ReTIS), International Conference on*, pp. 281–286.
26. Shaikh, S. H., Maiti, A. K., & Chaki, N. (2013). A new image binarization method using iterative partitioning. *Machine vision and applications*, 24(2), 337–350.
27. Sigdel, M., Pusey, M. L., & Aygun, R. S. (2013). Real-time protein crystallization image acquisition and classification system. *Crystal Growth and Design*, 13(7), 2728–2736.
28. Sigdel, M., Sigdel, M., Dinc, S., Dinc, I., Pusey, M., & Aygun, R. (2015). Focusall: Focal stacking of microscopic images using modified harris corner response measure. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 99, 1–1.

29. Sigdel, M., Sigdel, M.S., Dinç, İ., Dinç, S., Aygün, R.S., & Pusey, M.L.(2015) Chapter 27 - automatic classification of protein crystal images. In *In Emerging Trends in Image Processing, Computer Vision and Pattern Recognition*. Morgan Kaufmann, pp. 421–432.
30. Silva, A.(2011). Region-based thresholding using component tree. In *18th IEEE International Conference on Image Processing (ICIP)*, pp. 1445–1448.
31. Starks, J. L., & Fehl, A. (2012). *Adobe Photoshop CS6: Comprehensive* (1st ed.). Boston, MA, United States: Course Technology Press.
32. Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining* (1st ed.). Boston, MA, USA: Addison-Wesley Longman Publishing Co.
33. Zhang, J., & Hu, J. (2008). Image segmentation based on 2d otsu method with histogram analysis. In *2008 International Conference on Computer Science and Software Engineering*, Vol. 6, pp. 105–108.

Chapter 9

Visualization

Abstract As high throughput, crystallization screening and analysis systems automate the processes starting from setting up plates to scoring, this enables conducting thousands of experiments in a short time. Analysis of crystallization trial experiments in the past has been cumbersome due to the physical environment where an expert needs to look crystallization trial images one by one using a microscope with the likelihood of the majority of experiments yielding unsuccessful outcomes. The visualization of crystallization experiments on a display with some highlighted information along with annotation capability can provide experts a user-friendly and shared environment of collaborative analysis. In this chapter, we summarize the methods and information displayed on various visualization software for protein crystallization analysis.

9.1 Introduction

High throughput systems are capable of setting up many plates and analyzing the solutions in those plates over a time course. With automation of setting up experiments for crystallization trials, crystallographers are exposed to thousands of protein crystallization images. Looking each plate well one by one through a microscope to detect crystals has been a tedious task for crystallographers especially considering very low success rate for some difficult proteins. Expert analysis is still at the core of protein crystallization process despite automation of many states of crystallization experiments. Expert analysis is especially required to validate scorings done by the system, distinguish false classifications, eliminate false hits, and most importantly detect crystals that could be missed by the system. Moreover, based on the results of experiments, experts may design new screens to yield crystalline conditions. Experiment visualization interfaces may help experts during this process.

The visualization software for protein crystallization experiments at least enables experts to look at crystallization trial images on a browser or application interface and annotate them in a relaxed environment than through a microscope in a somewhat cold and fatiguing lab environment [4]. Besides providing a cozy environment, the visualization system should assist experts in:

1. browsing thousands of crystallization trial images in a single session [4],
2. identifying plate wells-containing crystals,
3. displaying images at high resolutions with zooming and repositioning of images [4],
4. identifying cocktails that lead to crystalline conditions,
5. retrieving related or similar images from plates of the same protein [4],
6. manually scoring or updating automated scoring,
7. visualizing crystals under different lighting sources,
8. analyzing growth of crystals, and
9. developing new successful screens.

We have analyzed features of visualization software for crystallization trial images. We briefly look into plate view, single well (or drop) view, scoring support and visualization, time course view for growth of crystals, sequential view of a plate, multiple light source support, and chemical space mapping. In the following examples, we briefly explain how various visualization software presents experiment results and provide sample interfaces of some systems such as xtalPIMS [4], Visual-X2 [6], and RoCKS [1]. We finally provide some future directions for visualization of experiment results.

9.2 Plate Visualization

The main idea of plate visualization is to provide an overall view of an experiment result typically on a grid layout similar to the experiment plate. Plate visualization may quickly help crystallographers identify the wells containing crystals. Plate visualization interface acts as an index. Once interesting wells are identified on plate view, the expert may select a specific well and then get more detailed information about the contents of that well.

In the literature, the most common of type of plate used for visualization displays is a 96-well plate composed of 8 rows and 12 columns. When 96 images are resized to fit on a screen, lots of details in those images are not visible whatsoever. Displaying supportive information for each well such as the score of a well is a necessity. Cocktail information does not fit into the screen. However, Visual-X2 [6] developed at the University of Alabama in Huntsville for iXpressGenes, Inc. uses hovering option to display the cocktail when the mouse hovers over a well on plate view. We briefly describe several strategies of plate visualization below.

1. *Thumbnails*. Original images can be displayed as thumbnails on the plate. The thumbnail images are bounded by a colored boundary to indicate the category or score of an image [14] (Fig. 9.1). The viewing software in Hiraki et al. allows users to select sizes of images to be displayed. The score of specific well is displayed as a colored well-id on the top of each scored well (Fig. 9.2). Since images can be selected at different sizes, the users have to scroll the window if large sizes are selected. These methods work at an acceptable level if 96-well

plate or smaller sizes are used. However, if 1536-well plate is used, displaying all images on screen at the same time has little value to experts since images would be too small to glean useful information. To solve this problem, MacroScope [9] displays 1536-cocktail screen as 16 arrays of 96 thumbnail images.

- 2. *Color coded grid.* Each well on a plate could be represented with a color corresponding to the score of that well (Fig. 9.3). The expert may choose a specific colored well to see details of that well. This works satisfactorily if the number of scores or classes is low. As the number of scores increases, the user may be challenged to remember what each color represents.

Fig. 9.1 Scores as colored boundaries for thumbnail images [14]

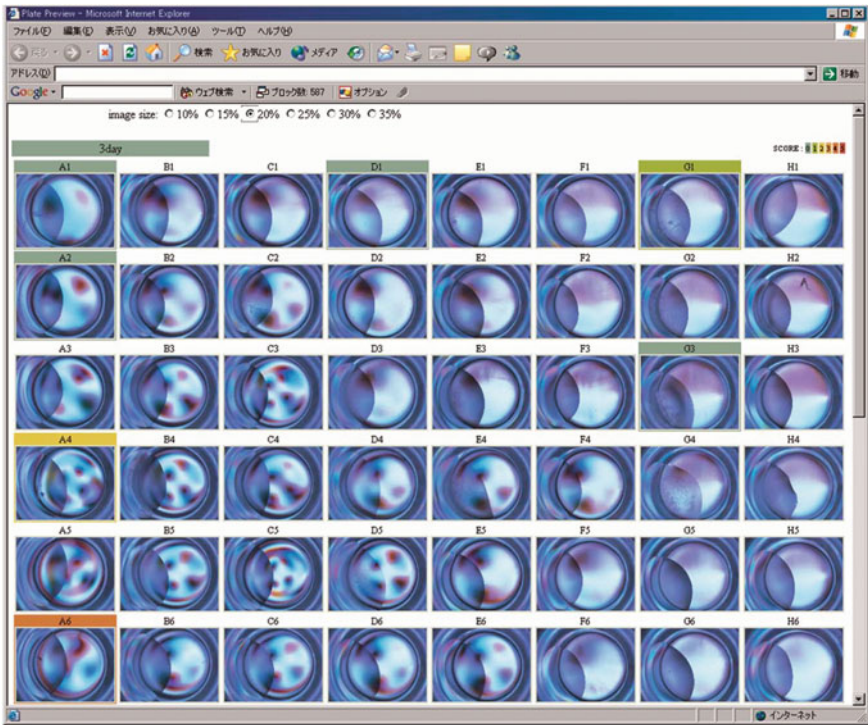
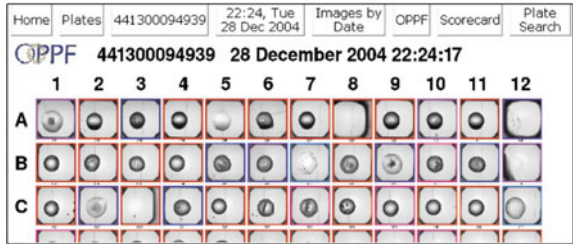


Fig. 9.2 The plate view interface in [7]

3. *Glyphs*. Alternatively, glyphs that represent different types of scores could be used (Fig. 9.4). The thumbnails may be too small to visualize any well. In that case, glyphs could be used to determine wells containing crystals (Fig. 9.5). The glyph representation should be informative about the scores or phases of crystallization.
4. *Region-of-Interest*. Thumbnails might be too small to see anything in them and glyphs are just symbolic representations that do not show the actual content.

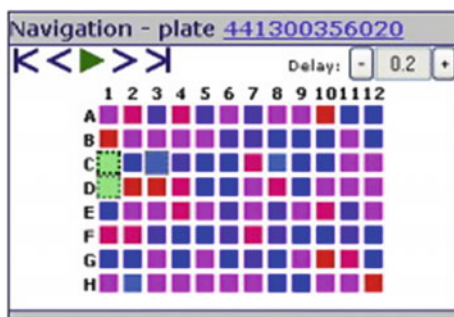


Fig. 9.3 Color coded plate view in xtalPIMS



Fig. 9.4 The glyphs used in Visual-X2



Fig. 9.5 Plate view in Visual-X2

Alternatively, image processing tools can be used for detecting regions that are likely to contain the crystal in the image. Those regions-of-interest could be displayed on the plate view.

The visualization software should support different types of plates of various sizes. Moreover, since there are plates available having several drop positions per well, the visualization software may need to consider those plates as well. Figure 9.5 shows Visual-X2 interface showing 3 drop positions per well.

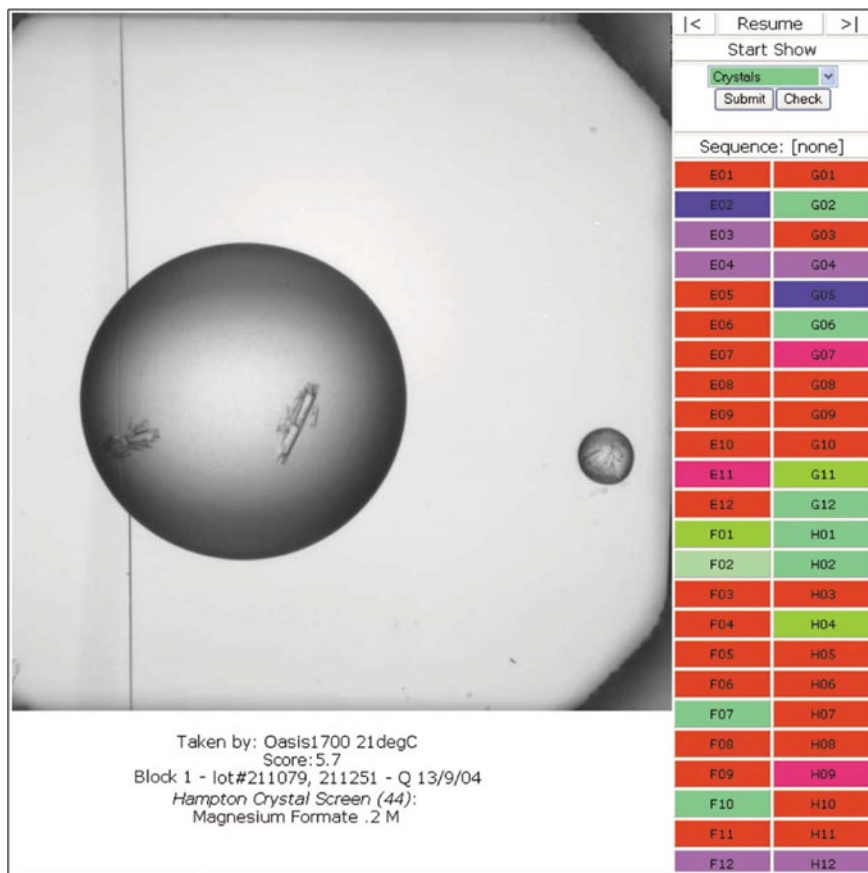


Fig. 9.6 Well view with well-ids score-colored [14]

9.3 Well View

While plate view provides insight about overall success of experiments for a specific plate, only well view may provide detailed information and high-resolution image captured for a specific cocktail. In general, experts focus on a single well rather than the complete well view. As mentioned earlier, the plate view may serve as an index, where the user can choose or click a specific well on the plate interface to see the image at a high resolution. Since the complete scene is dedicated for a single cocktail, additional information such as the screen used for the experiment, cocktail, the imaging system, temperature, date and time captured, and its score could be displayed (Figs. 9.6 and 9.7). Moreover, while viewing a well at a high resolution, the color coded wells of that plate (Figs. 9.6 and 9.7) are helpful to quickly access other wells.

Visualization interface should provide a detailed view of specific regions on the image. For example, RoCKS [1] allows users to drag a rectangle on the images to get a more detailed view of the selected region of interest. Images can be zoomed to investigate regions in an image [4]. The interface should support selecting points on the image and provide the length of chosen segments (Fig. 9.7). If electron microscopy is used for crystallization trial analysis, image acquisition can be achieved at low magnification (60x), medium magnification (1350x), high magnification (21,000x) by the microscope computer [3]. The user may choose the number of images per grid, the number of images per region of interest (ROI), and the frequency of focus measurement for image processing. Hu et al. [8] proposed a system that associates images collected from the electron microscopy with the screen conditions. Their

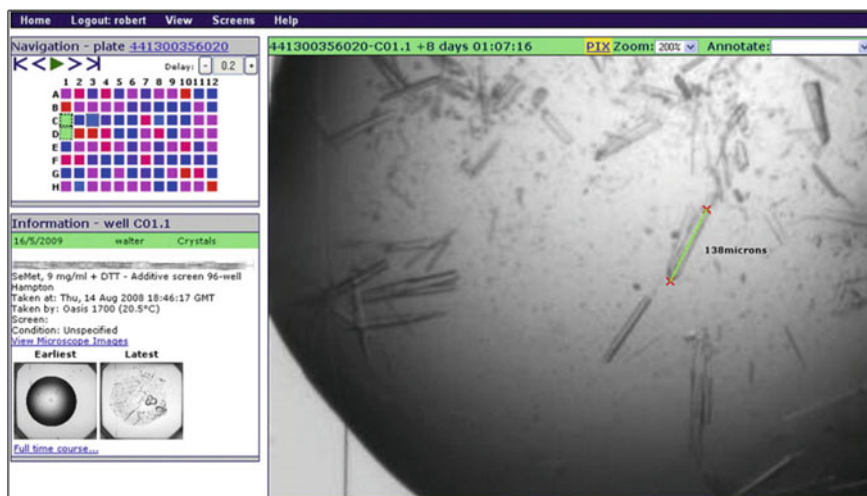


Fig. 9.7 The crystallization trial viewer interface in xtalPIMS

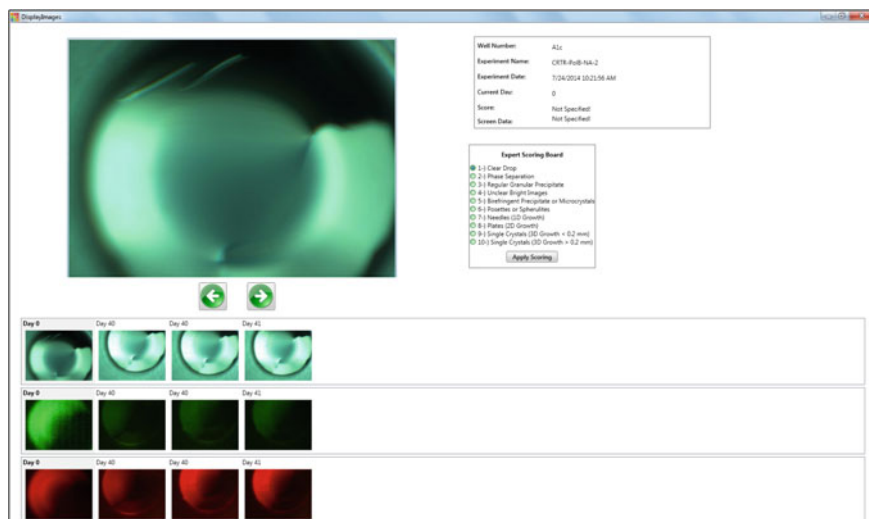


Fig. 9.8 Scoring window of Visual-X2

interface maintains crystallization screen information such as crystallization trials, target protein sequence, score, and conditions for expression and purification.

9.4 Scoring Crystallization Trials

Scoring crystallization trials could be automatic as well as manual. Ideally, since there could be many trials that do not yield crystalline conditions, automation of scoring is desired. Nevertheless, these automated scoring tools are not foolproof. The plate visualization tool should enable to update scores by an expert. Once the scores have been updated, the new scores should be stored in the database. The system should maintain the automated score as well as the expert score to distinguish whether the expert has changed the scoring or not. If the automated scoring is not correct, the corresponding image may be used in training a new classifier. These systems do not need to limit experts to assign a single score. MacroScope [13] allows experts to assign multiple categories to images except for the clear category. RoCKS [1] allows experts to choose a score as well as mark wells as salt, hit, or for future attention. Figure 9.8 shows the scoring window in Visual-X2.

9.5 Multiple Crystallization Trial Analysis

Rather than analyzing a single well, observing outputs of multiple experiments simultaneously may help experts analyze crystallization conditions. Multiple images may be displayed with respect to time course, their scores, positions in wells, or various light sources used for capturing images.

9.5.1 Time Course Analysis

Protein crystallization is not an instant process once a plate is set up. Depending on the type of protein, the crystal growth may take from several hours to months. During this process, the plate may be analyzed regularly to investigate crystal growth. The time course [14] of a well could be displayed as a sequence (Figs. 9.9 and 9.10), slideshow or movie to the user. xtalPIMS displays the earliest and latest images in their crystallization trial viewer as a summary while supporting displaying of the full time course (Fig. 9.7).

9.5.2 Support for Sequential View

Sequential view of a plate may provide the image, the cocktail, and the score simultaneously. The images may be sorted based on their position in the well as well on their score [5, 14]. The sequential view may also sort the images based on the scores so that the expert may focus only crystallization trials with high scores. The results can be provided as a slideshow, movie, or list to the user. xtalPIMS allows a movie mode for watching well images of a plate. xtalPIMS provides a color coded grid for displaying scores of plate wells (Fig. 9.7).

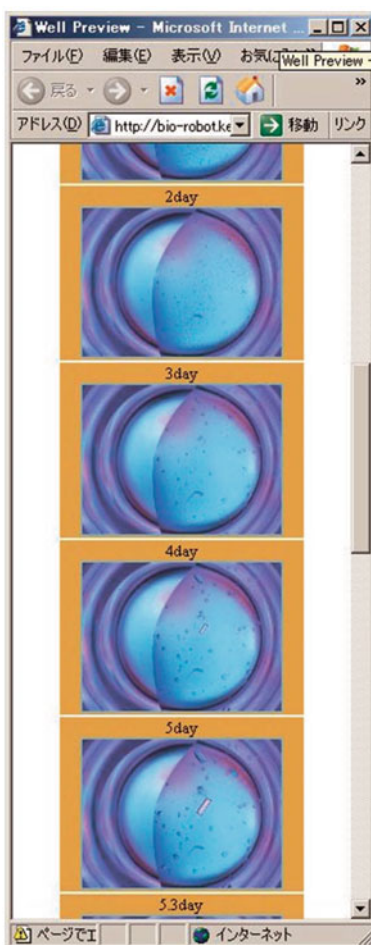
9.5.3 Multiple Light Source Support

Previously, the protein crystallization trials were usually captured under white light. Recently, trace fluorescent labeling has started to become popular. Depending on the fluorescence dyes used, the plate wells can be captured using different light sources. Trace fluorescent labeling may highlight crystals that would be difficult to detect under white light. Figure 9.8 shows multiple light support interface for Visual-X2.

9.6 Chemical Space Mapping

Autosherlock [12] is based on Microsoft Excel sheets. Its main goal is to help experts develop new successful crystals by analyzing experiment results where each cell has a numeric score and color. In Autosherlock the columns are first sorted based on cations and each cation is sorted based on anions vertically. In other words, there are vertical groups of cations. Horizontally, the groups are formed based on the molecular weight of PEG and each PEG group is sorted based on pH values. Such a representation provides an overview of conducted experiments as well as unsampled experiments. The visualization of unsampled experiments along with neighboring successful experiments may help crystallographer optimize conditions.

Fig. 9.9 The timecourse view [7]



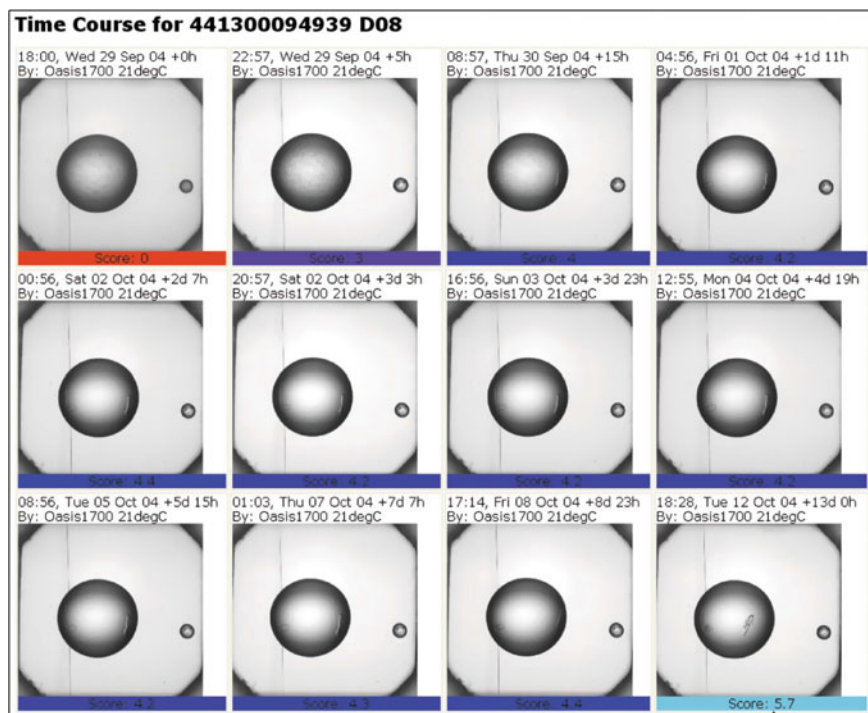


Fig. 9.10 The timecourse view [14]

AutoSherlock generates four worksheets: (1) outputs of the incomplete factorial cocktails, (2) outputs of commercial screens, (3) global view of outputs, and (4) a listing of scores, image names, and conditions. AutoSherlock interface [12] can be used to optimize conditions. For example, identifying the range of PEG concentration or plotting the chemical space PEG vs. pH by keeping salt constant helps identify crystalline conditions. Such plots give hints to experts about the outcomes of unsampled conditions (Fig. 9.11). They provide examples of how their interface could be used for apoferritin protein.

The web interface for PlateDB crystallization database [11] also lists occurrences of common conditions for crystals.

9.7 Summary

Visualization of protein crystallization trial images may help crystallographers identify crystalline conditions and determine conditions to be tried in newer experiments. The amount of information to be displayed on a regular display may not provide the details an expert might be looking for. Moving away from microscope to a computer

- and Symbolic Visualization Tool for Analysis of Protein Crystallization Trial Images (Poster)* Huntsville, AL.
7. Hiraki, M., Kato, R., Nagai, M., Satoh, T., Hirano, S., Ihara, K., et al. (2006). Development of an automated large-scale protein-crystallization and monitoring system for high-throughput protein-structure analyses. *Acta Crystallographica Section D: Biological Crystallography*, 62(9), 1058–1065.
 8. Hu, M., vink, M., Kim, C., Derr, K., Koss, J., DAmico, K., et al. (2010). Automated electron microscopy for evaluating two-dimensional crystallization of membrane proteins. *Journal of Structural Biology*, 171(1), 102–110.
 9. Luft, J. R., Collins, R. J., Fehrman, N. A., Lauricella, A. M., Veatch, C. K., & DeTitta, G. T. (2003). A deliberate approach to screening for initial crystallization conditions of biological macromolecules. *Journal of Structural Biology*, 142(1), 170–179.
 10. Luft, J. R., Wolfley, J. R., & Snell, E. H. (2011). Whats in a drop? correlating observations and outcomes to guide macromolecular crystallization experiments. *Crystal Growth and Design*, 11(3), 651–663.
 11. Mayo, C. J., Diprose, J. M., Walter, T. S., Berry, I. M., Wilson, J., Owens, R. J., et al. (2005). Benefits of automated crystallization plate tracking, imaging, and analysis. *Structure*, 13(2), 175–182.
 12. Nagel, R. M., Luft, J. R., & Snell, E. H. (2008). AutoSherlock: a program for effective crystallization data analysis. *Journal of Applied Crystallography*, 41(6), 1173–1176.
 13. Snell, E. H., Nagel, R. M., Wojtaszczyk, A., O'Neill, H., Wolfley, J. L., & Luft, J. R. (2008). The application and use of chemical space mapping to interpret crystallization screening results. *Acta Crystallographica Section D: Biological Crystallography*, 64(Pt 12), 1240–1249.
 14. Walter, T. S., Diprose, J. M., Mayo, C. J., Siebold, C., Pickford, M. G., Carter, L., et al. (2005). A procedure for setting up high-throughput nanolitre crystallization experiments. Crystallization workflow for initial screening, automated storage, imaging and optimization. *Acta Crystallographica Section D: Biological Crystallography*, 61(6), 651–657.

Chapter 10

Other Structure Determination Methods

Abstract There are more ways of gaining insight into macromolecular structure than X-ray diffraction. Like X-ray diffraction, some of these are based on the generation of ordered arrays of the molecule to be studied. For many reasons, based on either the protein or its function, this is not always possible. Others, some of which are currently enjoying a marked increase in popularity, do not require crystals. Many of these come with the added advantage that they can be used to capture reaction intermediates and/or enable the experimenter to observe changes in specific amino acids, which is often not possible with X-ray diffraction methods. This chapter divides into two sections; those methods that can be used to obtain a 3D structure (neutron diffraction, cryogenic electron microscopy, nuclear magnetic resonance, and X-ray free electron laser diffraction) and those that are suitable for more general structural information (chemical cross linking, fluorescence resonance energy transfer, circular dichroism). Virtually all of the methods discussed below can be expanded for the study of other aspects of macromolecular structure-function relationships, and some, such as fluorescence and chemical cross linking, are a subset of a rich methodology for the study of macromolecules.

10.1 Introduction

Merely having a pure protein and an understanding of how it behaves in solution does not guarantee that it can be crystallized. Just as there are some proteins that seemingly crystallize just because one wants them to, there are others that seem to resist all efforts at getting them into an ordered array. Several programs have been put forth where one inputs the proteins' primary sequence, either as the genetic codons or as the corresponding amino acids, and the program puts forth an assessment of the likelihood of obtaining crystals (for example, [17, 28, 29]). This assessment often comes with an output that indicates the location of the likely problematic regions. This information can be used to either delete those regions at the genetic level or, if the structural interest is centered on a specific interacting domain, to guide the experimenter in the specific sequence to be expressed for crystallization of that domain.

The process of purifying a protein itself can provide basic structural information useful for guiding the crystallization process. For example, sodium dodecylsulfate (SDS) gel electrophoresis can give the monomeric molecular weight(s), while passage down a calibrated size exclusion column can yield information about the oligomerization state. Native electrophoresis can be used to examine binding to added solution components using a gel-shift assay [26], useful for crystallization of protein-nucleic acid or protein-protein complexes. Concentration by ultrafiltration can also give size information; a hexameric protein having a monomeric molecular weight of 20 kDa should be retained by a membrane having a MW cut-off of 50 or 100 kDa, considerations that may speed up the purification process to enable getting the crystallization trials started with the most freshly purified protein. Light scattering studies can be an indication of the polydispersity of the protein preparation; while not direct structural information per se, monodispersity is generally a pre-requirement for successful crystallization. There are many more methods, not covered below, that can be employed to gain some insight into a macromolecules structure. The presentations are not meant to be exhaustive, but each includes references to one or more comprehensive reviews as a starting point for readers who wish to know about them in greater depth.

10.2 Neutron Diffraction (ND)

X-ray diffraction intensity is a function of the electron cloud around a diffracting atom, and as a result the diffraction intensity is proportional to the size of that cloud. This makes the diffraction from hydrogen, with just one electron, very weak. Neutron diffraction is from the nuclei of atoms, and the diffraction intensity will depend upon the isotope of that atom. Neutron scattering intensity does not vary linearly with atomic number. Of particular interest to biomolecular structure, both hydrogen and deuterium are strong neutron scatterers, which results in the hydrogen positions being much easier to determine by ND. As the scattering lengths for hydrogen and deuterium have opposite sign, and hydrogen scattering has a large inelastic component which contributes to the background noise, it is common for biomolecule ND data to be acquired from samples after exchanging hydrogens for deuterium. This exchange can be carried out by deuteration where the more labile protons are exchanged by dialysis of the protein solution or soaking of the crystals in D₂O mother liquor. Less labile hydrogens are deuterated by recombinant expression of the macromolecule in a deuterated growth media

ND requires a neutron source, which is provided by a nuclear reactor or a spallation source. The type of source used dictates the associated components of the instrument. Neutrons from a reactor source are produced continuously, and monochromators and filters are employed with a reactor source for selecting the wavelength. Neutrons from spallation sources are produced in discrete bursts. As higher energy (lower wavelength) neutrons are faster spallation sources use a time of flight approach to

separate the different wavelengths. Currently, the most intense spallation neutron source is that at Oak Ridge National Laboratory.

Neutrons only interact weakly with matter, and as a result a major drawback of ND is the requirement for larger crystals, typically having a minimum size of $\sim 1 \text{ mm}^3$. For macromolecules obtaining crystals in this size range can be a daunting task, requiring careful control of the growth conditions and environment, as well as a large supply of protein, which must be kept stable during the growth of the crystal. A second negative aspect of the weak interaction with matter, coupled with the relatively low intensity (compared to X-rays) sources of neutrons, is the need for long data collection times. Times of one week to acquire a data set are typical [4], and depending upon the crystal symmetry data collection times exceeding 2–3 weeks are not uncommon. As with X-ray crystallography technological advances in producing higher flux neutron sources, as well as improved detectors, are expected to reduce these data collection times.

Half of the atoms of a protein are hydrogen and as a result, a neutron structure as twice the number of atomic positions to be determined as from the corresponding X-ray data set. While ND is presented here as an alternative method for macromolecule structure determination, success is in fact highly dependent upon having X-ray data in hand. Recent software advances have taken this into account, enabling joint X-ray/neutron refinement strategies [2, 12].

Resolving hydrogen atoms by X-ray crystallography typically requires data collected to a diffraction resolution of $\leq 1 \text{ \AA}$. In contrast, after deuteration, these locations are visible by ND at resolutions in the 2.5 \AA range. Neutrons do not cause damage to the biomolecule, and as a result data collection can be carried out at room temperature. The ability to locate hydrogen (deuterium) by ND is important for determining the protonation states of active site residues. Suggested recent reviews for further reading are: Blakeley et al. [3] and O'Dell et al. [24].

10.3 Nuclear Magnetic Resonance (NMR)

NMR is a non-crystal dependent method for determining macromolecule structure. In practice, an aqueous solution of the sample is placed inside a strong magnet and subjected to radio frequency signals. The nuclei absorb at frequencies that are dependent upon their environment, which is affected by other atoms in close proximity. Analysis of the signals is used to build a proximity map, and thus a model of the protein structure. A strength of NMR is that, as a protein molecule may have many conformational forms, structural models based on NMR may, in fact, be a better representation of the protein than a “fixed” structure as obtained by X-ray crystallography. This also makes NMR a useful technique for looking at subtle changes in a macromolecule’s structure due to factors such as environmental changes or binding to another molecule.

Isotopes having an odd number of protons and/or neutrons have a nonzero spin, which spin is aligned in a magnetic field and can be perturbed by a radio frequency

pulse. NMR signals are enhanced by isotopes such as ^1H , ^{13}C , ^{15}N , and ^{31}P . These all have a magnetic dipole, with an associated orientation energy, which can be set in a magnetic field. Except for ^1H , used for 1D NMR, the samples must be enriched with the other isotopes due to their low natural abundance. By inputting electromagnetic radiation, i.e., radio waves, at specific frequencies these dipoles can be transitioned to a new orientation, which is the measured NMR signal. The utility of NMR for structure determination comes from the field strength of the observed resonance for a given signal source (nucleus) being highly sensitive to its electronic environment. Interactions with its neighboring nuclei cause the resonance band of a particular atom to be split into a group that is determined by the number, distance, and symmetry of those neighbors.

NMR for macromolecular structure studies requires NMR-active, spin-1/2, atoms. While the natural abundance of ^1H is sufficient, samples must be enriched with ^{13}C and/or ^{15}N for NMR studies. This enrichment process can be carried out using recombinant protein produced using *E. coli* grown in minimal growth media having isotopically labeled carbon and nitrogen sources. While this approach is relatively cheap, costs can increase significantly if protein expression requires a eukaryotic expression system. These higher costs are somewhat ameliorated by more recent instrumentation which requires sample volumes of 0.1 mL at concentrations of ~ 0.05 mM. For a 20 kDa protein, this translates to ~ 0.1 mg of protein, making NMR a very attractive approach for the parsimonious structural biologist.

NMR spectrometers are characterized by their proton resonant frequency in MHz, which is a function of the instrument's magnetic field strength. Higher strength magnets result in higher resonant frequencies. The resolution of a NMR instrument, the ability to separate resonances, increases with the magnetic field strength. Complex molecules have more crowded spectra, and as a result higher field strength magnets are needed to resolve the chemical shifts in the spectra. Some useful starting reviews of NMR and its capabilities for those interested would be Kwan et al. [18] and Ziarek et al. [32]. NMR is particularly useful for determining the structure of membrane proteins, particularly for their study in native lipid environments, as reviewed by Opella and Marassi [25].

10.4 Cryogenic Electron Microscopy (Cryo-EM)

Cryo-EM, not surprisingly, has its origins in electron microscopy, where the sample is placed on a grid in a high vacuum for imaging. An electron beam is emitted from a filament and accelerated toward an anode, with lenses and apertures used to control the shape and size of the beam. The electrons are deflected by interactions with the sample, then the beam is passed through additional lenses and aperture(s) to reduce scattering and magnify the image on a detector. While electron microscopes were capable of achieving resolutions to the 5–10 Å range, in practice “seeing” the contrast from different protein domains was generally not possible. This led to methods for enhancing contrast to provide the necessary intensity differences, using methods

such as negative or positive staining and shadow casting. While these help, the harsh treatment of the macromolecule led to a practical resolution limit of ~ 15 Å. This is usually sufficient resolution to determining molecular shape and the subunit structure of multimeric proteins, but not the needed atomic resolution for detailed structural information.

The advance to cryo-EM started with the use of liquid ethane to flash cool the proteins to immobilize them on a grid in vitreous buffer solution for imaging [1]. By flash cooling the experimenter avoids problems associated with dehydration (“classical” EM) or from ice crystals due to freezing. Flash cooling, aka vitrification, is where the solution is cooled at a sufficiently fast rate that structured ice crystals are not formed, but rather an unstructured glass-like state of the aqueous solution. Dehydration of the sample in the vacuum of the electron microscope is prevented by virtue of being embedded in a layer of vitreous solvent, which also reduces radiation-induced damage to the protein. The next advance was in detector development, away from film-based methods [23]. The detector advances, coupled with improved image processing methods, enabled cryo-EM to move from large complexes and low resolution to levels that are now approaching those of X-ray crystallography.

The image from a single particle is a 2D projection onto the imaging plane. By rotation of the particle in the imaging beam one can obtain a series of 2D image projections that can be processed to obtain a 3D structure. This rotation is virtually provided by each image typically having multiple randomly oriented particles present. Each projected 2D image is defined by three rotational Euler angles that define the molecular orientation, two positional angles, and the position of the molecule in the z-axis (the focus or beam direction). Thus, the image is a field of randomly oriented particles, the orientation of each of which is defined by the six orientational parameters. Image processing in software is used to align each of the separate molecular images to each other. Use of more images improves the signal to noise level and provides an enhanced range of orientations to provide an improved 3D structure reconstruction. The advances in cryo-EM to X-ray crystallographic level resolutions have been driven by those for the detectors and image processing software.

The major advantage of cryo-EM is that one does not have to have crystals of the molecule under analysis, enabling structures to be determined from macromolecules that have heretofore resisted crystallization. A major stumbling block for cryo-EM is the high cost of the instrumentation, currently \sim \$7 million or greater. The field is rapidly evolving, and ever higher resolution structures are now being obtained by cryo-EM. Recent reviews in this rapidly evolving field are Elmlund et al. [9], Takizawa et al. [30], Jonic [16].

10.5 X-Ray Free Electron Laser (XFEL)

Crystallization screening experiments often result in the growth of microcrystals having sizes in the ≤ 1 μm size range. While this may constitute proof that the protein can be crystallized, crystals in this size range are too small for use with conventional

synchrotron X-ray sources. Use of smaller crystals is desirable as they are likely to have fewer defects, to diffract to a higher resolution. However, smaller crystals have fewer diffraction centers, and thus require a higher X-ray intensity to produce a measurable diffraction signal, resulting in radiation damage which considerably shortens the crystal lifetime in the beam. Thus “classical” X-ray crystallography requires growth of crystals having sufficient size for use.

The recent development of femtosecond X-ray lasers having intensities $\sim 10^9$ greater than synchrotron sources provides a means past this small size limitation. The experimental approach taken was to collect a single diffraction image from a crystal using the laser beam. While the intense beam does destroy the crystal, the single image of diffraction data is collected before this destruction occurs. A steady stream of small crystals is run through the beam as it is pulsed. The probability of collecting a diffraction image from each beam pulse is proportional to the pulse rate and the crystal density in the stream. A complete data set results from the collection of thousands of diffraction images, and in actuality is derived from only a very small percentage of the crystals that are passed through the beam path. As the crystals are randomly oriented when in the beam a complete diffraction data set can be assembled. Diffraction has been observed from crystals ≤ 10 unit cells on a side [7]. The data is collected from crystals in solution at room temperature, which enables time resolved functional studies (references).

The immediate advantage of XFELs is that small crystals are often more readily obtained than large. The exception to this of course being in those cases where one only wants small crystals. A second advantage comes from being able to follow enzyme kinetics. While optical triggering of photoactive proteins can be carried out to high time resolution, following non-optical enzymatic action on a substrate requires that the protein molecules simultaneously “see” the substrate at the same time. Due to the diffusion rate of small molecules through the solvent channels this is not feasible with larger crystals. However, with a very small crystal the diffusion is very quick, and by varying the time interval between when a substrate solution is mixed with a nanocrystalline stream and when it is diffracted by the FEL beam one can build up a sequence of structural models as the reaction proceeds [6].

XFELs found immediate application in the field of membrane protein structure, where small crystals, particularly as a result of LCP growth, are common. One of the earliest structures solved using XFELs was for cyanobacterial Photosystem 1 [10]. The experiment consumed 14 mg of protein, with the data being collected over a 24 h period. It was estimated that at a beam pulse rate of 30-Hz diffraction data was collected for only one out of every 25,000 nanocrystals that were passed through the beam path. Faster laser pulse rates and a higher crystal density in the flow will reduce the time to collect a data set.

The small crystal size of XFELs is also a limitation. Crystals in the nanometer size range are not easily identified by the usual microscopy instrumentation employed in protein crystallization. Also, the amount of protein required leads to larger volumes than usually set up in a crystallization experiment. Similar to ND, the available facilities for conducting XFELs diffraction experiments is a limiting factor. Rapid

advances in the field are increasing the utility of XFELs, resulting an increasing demand for access to the few facilities extant. Some recent reviews for XFELs are Johansson et al. [15], Martin-Garcia et al. [20], and Jain and Techert [14].

10.6 Other Approaches

Several of the above approaches, such as NMR and EM, can be used to determine the 3D structure of a macromolecule without having to resort to its crystallization. There are times, particularly at the start of a project, when the experimenter wants to have some basic structural information without having to resort to a full-on structure determination. This information might be, for example, which domains of two molecules interact with each other to form a complex, what is the overall shape of the macromolecule, or which surface interacts with a membrane. This type of information can be used to inform the selection of just those specific domains for subsequent structural determination, or for the inclusion or not of a domain for deletion for a structural study. Or they can be used to gain some understanding of how different domains within a protein interact. An advantage of those presented here is that they are not dependent upon the presence of major expensive pieces of capital equipment, but can generally be carried out by resources that are likely to be available to many laboratories or groups.

10.6.1 *Chemical Cross linking*

Chemical cross linking is often employed to determine the proximity of one molecule to another. Crosslinkers are available having a range of sizes and chemistries. Crosslinkers as one might imagine typically have two reactive groups separated by a spacer group of a defined length which determines the reach of the agent, thus the distance that can be probed. Thus, for a crosslinker specifically bound to a given amino acid on the surface of a protein, its ability or not to react at the distal end with another amino acid gives an indication of the distance between those two amino acids. An obvious limitation to this approach is now does one keep the distal reactive group from reacting to the same macromolecule, which is where reagent selection and experimental design comes into play.

Cross linking studies are dependent upon knowing where the donor and acceptor are located on the macromolecule, making the chemistry of their placement a critical component of the measurement process. This is not trivial experimentally, unless one has a reactive group that can be specifically targeted. Examples of such would be active site residues, the N-terminal amine which has a different pKa than side chain amines, and sulfhydryl groups. Another approach would be to form the desired complex, reversibly block the exposed reactive groups, separate the complex, modify the previously occluded reactive group(s), unblock the previously blocked groups,

then proceed with the binding and cross linking experiment. Alternatively, a simple approach where the complex is formed and random cross linking is carried out can be taken.

Crosslinkers can be commercially obtained having a range of reactive chemistries, as well as a range of spacer arm lengths between the reactive groups. Those having the same reactive group on both ends are known as homobifunctional, while those having different reactivities are known as heterobifunctional. Trifunctional crosslinkers are also available. The reactivities may be specific to amine, sulfhydryl, carboxyl, hydroxyl, or aldehyde and ketone groups. Photoreactive groups are also available. An excellent guide to cross linking, as well as other methods involving the modification of macromolecules, can be had with Hermanson [13].

10.6.2 Fluorescence Resonance Energy Transfer

Fluorescence resonance energy transfer (FRET) is a means for determining the distance between a donor and acceptor fluorescent probe. The probe pairs are selected such that the donor probe fluorescence emission spectrum overlaps the absorption spectrum of the acceptor species. Energy transfer is by nonradiative dipole-dipole coupling; emission and reabsorption of a photon is not involved. FRET efficiency is a function of the spectral overlaps, the distance between the pair, their relative orientation their emission and excitation dipole moments, and the quantum yield of the donor in the absence of the acceptor. The distance range that can be measured is a function of these factors, but the typical maximum range for a well-matched donor-acceptor pair is ~ 10 nm.

When FRET occurs the donor fluorescence intensity decreases, with the energy decrease showing up as an increase in the acceptor fluorescence intensity. In practice, the acceptor can be just an absorbing species, as one typically measures the decrease in the donor fluorescence intensity. The transfer efficiency is inversely proportional to the sixth power of the distance between the donor and acceptor, and with a well-chosen FRET pair very sensitive changes in distance between the two molecules, and thus the parts of the macromolecule(s) to which they are attached, can be made.

Accurate FRET measurements are dependent upon knowing where the donor and acceptor are located. Approaches that can be taken for this are similar to those outlined above for cross linking. However, unlike cross linking, one does not have to be concerned about the reactivity of a distal end while setting up an experiment.

FRET is only one of a broad range of fluorescence-based approaches to the studies of macromolecules. For proteins having one or more fluorescent amino acids one can often track changes in their conformational state by following the accessibility of those amino acids to the bulk solvent, or deliberately added quenchers in the solvent, as a function of imposed solution conditions such as pH, temperature, composition, etc. Thermal stability studies are carried out on proteins to determine the optimum solution conditions for stability by adding a hydrophobic-binding dye to the protein, then raising the temperature while tracking the fluorescence intensity [5, 8]. The

added dye is typically one that has little to no fluorescence in aqueous solution, with the fluorescence increasing as it binds to hydrophobic surfaces exposed as the protein unfolds. An upward shift in the melting temperature indicates solution conditions that stabilize the protein. A very useful starting source for fluorescence methods is Lakowitz [19].

10.6.3 *Circular Dichroism (CD)*

Circular dichroism (CD) is a measure of the unequal absorptivity for left and right circularly polarized light. A molecule that differentially absorbs left- and right-handed circularly polarized light is optically active, or chiral. A chiral molecule is one where one cannot superimpose a mirror image of its structures, a standard example being ones left and right hands. Chirality arises from the presence of one or more atoms capable of having an asymmetric distribution of bound groups. For a tetrahedral carbon this requires that all 4 substituent groups be different. All amino acids, with the exception of glycine, have a chiral alpha carbon atom, with the L-form being that found in proteins.

CD only occurs at wavelengths that can be absorbed by a chiral molecule. In a CD measurement, the linearly polarized light at the wavelength of an optically active transition energy comes from the sample elliptically polarized because of unequal absorption of the left and right-handed components. CD spectra may exhibit both positive and negative peaks. The CD spectra of a biological macromolecule is affected by the 3D structure of that molecule, and is not the sum of the spectra of the constituent residues. More, the secondary structural elements of proteins, alpha-helix, beta-sheets, random coil, have specific CD absorption bands, all below 260 nm, which makes CD spectroscopy a very useful tool for gaining a de novo understanding of their structure. Not surprisingly changes in these structural elements as a function of added solution parameters can also be followed.

CD instrumentation is generally available in most larger chemistry or biochemistry departments. Desktop instrumentation can typically measure down to ~190 nm wavelength. As with all measurement techniques the performance limits are determined by the signal to noise characteristics. Other considerations for CD measurements are the path length and the buffer employed; many common buffers and other solution components absorb in the 185–250 nm region, and their use can impose limitations on the spectral range that can be studied. Protein concentrations inversely vary with the cell path length, but as a rule of thumb the maximum absorbance at any of the wavelengths to be investigated should be ≤ 0.9 OD. For performance beyond desktop instrumentation it is possible to carry out CD measurements at synchrotron facilities, which have higher light intensities extending down to much lower wavelengths. However, at wavelengths < 190 nm water absorption becomes a critical

consideration, as well as that of the method used to secure the sample. CD spectra have been collected for single amino acids down to 120 nm, using samples in the solid state immobilized on MgF₂ windows [21]. Some reviews for further exploration of CD spectroscopy are Johnson [31], Gottarelli et al. [11], Ranjbar and Gill [27], and Micsonai et al. [22].

10.7 Summary

There are a number of methods that can be used to obtain structural information about a macromolecule. Some can be used early on in the study to obtain basic solution data, such as the oligomeric state, gross molecular shape, and an indication of what it may be bound to. Other methods can be used to study distances, between domains on a single molecule or between regions on two molecules, and to better determine the surfaces that are interacting. Most of these methods can be carried out with bench-top equipment likely to be present in a laboratory or life sciences or biochemistry department. Detailed structural studies can be carried out with methods other than X-ray crystallography, and many of these methods can be used to gain functional information that is not readily accessible to “standard” X-ray methods. The negative aspect of this is that the instrumentation for these methods is generally rather expensive, and not likely to be available in most departments but rather as a central, often national laboratory, facility.

References

1. Adrian, M., Dubochet, J., Lepault, J., & McDowell, A. W. (1984). Cryo-electron microscopy of viruses. *Nature*, 308(5954), 32–36.
2. Afonine, P. V., Mustyakimov, M., Grosse-Kunstleve, R. W., Moriarty, N. W., Langan, P., & Adams, P. D. (2010). Joint X-ray and neutron refinement with phenix.refine. *Acta Crystallographica Section D: Biological Crystallography*, 66(11), 1153–1163.
3. Blakeley, M. P., Hasnain, S. S., & Antonyuk, S. V. (2015). Sub-atomic resolution X-ray crystallography and neutron crystallography: promise, challenges and potential. *IUCrJ*, 2(4), 464–474.
4. Blakeley, M. P., Langan, P., Niimura, N., & Podjarny, A. (2008). Neutron crystallography: opportunities, challenges, and limitations. *Current Opinion in Structural Biology*, 18(5), 593–600.
5. Boivin, S., Kozak, S., & Meijers, R. (2013). Optimization of protein purification and characterization using ThermoFluor screens. *Protein Expression and Purification*, 91(2), 192–206.
6. Calvey, G. D., Katz, A. M., Schaffer, C. B., & Pollack, L. (2016). Mixing injector enables time-resolved crystallography with high hit rate at X-ray free electron lasers. *Structural Dynamics*, 3(5), 054301.
7. Chapman, H. N., Fromme, P., Barty, A., White, T. A., Kirian, R. A., Aquila, A., et al. (2011). Femtosecond X-ray protein nanocrystallography. *Nature*, 470(7332), 73–77.
8. Dupeux, F., Rwer, M., Seroul, G., Blot, D., & Mrquez, J. A. (2011). A thermal stability assay can help to estimate the crystallization likelihood of biological samples. *Acta Crystallographica Section D: Biological Crystallography*, 67(11), 915–919.

9. Elmlund, D., Le, S. N., & Elmlund, H. (2017). High-resolution cryo-EM: the nuts and bolts. *Current Opinion in Structural Biology*, *46*, 1–6.
10. Fromme, P., & Spence, J. C. (2011). Femtosecond nanocrystallography using X-ray lasers for membrane protein structure determination. *Current Opinion in Structural Biology*, *21*(4), 509–516.
11. Gottarelli, G., Lena, S., Masiero, S., Pieraccini, S., & Spada, G. P. (2008). The use of circular dichroism spectroscopy for studying the chiral molecular self-assembly: an overview. *Chirality*, *20*(3–4), 471–485.
12. Gruene, T., Hahn, H. W., Luebben, A. V., Meilleur, F., & Sheldrick, G. M. (2014). Refinement of macromolecular structures against neutron data with SHELXL2013. *Journal of Applied Crystallography*, *47*(1), 462–466.
13. Hermanson, G. T. (2013). *Bioconjugate techniques* (3rd ed.): Academic Press.
14. Jain, R., & Techert, S. (2016). Time-resolved and in-situ X-ray scattering methods beyond photoactivation: utilizing high-flux X-ray sources for the study of ubiquitous non-photoactive proteins. *Protein and Peptide Letters*, *23*(3), 242–254.
15. Johansson, L. C., Stauch, B., Ishchenko, A., and Cherezov, V. (2017). A bright future for serial femtosecond crystallography with XFELs. *Trends in Biochemical Sciences*.
16. Joni, S. (2016). Cryo-electron microscopy analysis of structurally heterogeneous macromolecular complexes. *Computational and Structural Biotechnology Journal*, *14*, 385–390.
17. Kurgan, L., Razib, A. A., Aghakhani, S., Dick, S., Mizianty, M., & Jahandideh, S. (2009). CRYSTALP2: sequence-based protein crystallization propensity prediction. *BMC Structural Biology*, *9*, 50.
18. Kwan, A. H., Mobli, M., Gooley, P. R., King, G. F., & Mackay, J. P. (2011). Macromolecular NMR spectroscopy for the non-spectroscopist. *FEBS Journal*, *278*(5), 687–703.
19. Lakowicz, J. R. (2006). In J. R. Lakowicz (Ed.), *Principles of fluorescence spectroscopy* (3rd ed.). US: Springer.
20. Martin-Garcia, J. M., Conrad, C. E., Coe, J., Roy-Chowdhury, S., & Fromme, P. (2016). Serial femtosecond crystallography: a revolution in structural biology. *Archives of Biochemistry and Biophysics*, *602*, 32–47.
21. Meierhenrich, U. J., Filippi, J.-J., Meinert, C., Bredehft, J. H., Takahashi, J.-I., Nahon, L., et al. (2010). Circular dichroism of amino acids in the vacuum-ultraviolet region. *Angewandte Chemie International Edition*, *49*(42), 7799–7802.
22. Micsonai, A., Wien, F., Kernya, L., Lee, Y.-H., Goto, Y., Rfregiers, M., et al. (2015). Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. *Proceedings of the National Academy of Sciences*, *112*(24), E3095–E3103.
23. Milazzo, A.-C., Leblanc, P., Duttweiler, F., Jin, L., Bouwer, J. C., Peltier, S., et al. (2005). Active pixel sensor array as a detector for electron microscopy. *Ultramicroscopy*, *104*(2), 152–159.
24. O'Dell, W. B., Bodenheimer, A. M., & Meilleur, F. (2016). Neutron protein crystallography: a complementary tool for locating hydrogens in proteins. *Archives of Biochemistry and Biophysics*, *602*, 48–60.
25. Opella, S. J., & Marassi, F. M. (2017). Applications of NMR to membrane proteins. *Archives of Biochemistry and Biophysics*, *628*, 92–101.
26. Park, S.-H., & Raines, R. T. (2004). Fluorescence gel retardation assay to detect Protein-Protein Interactions. In *Protein-Protein Interactions, Methods in molecular biology* (pp. 155–159): Humana Press. <https://doi.org/10.1385/1-59259-762-9:155>.
27. Ranjbar, B., & Gill, P. (2009). Circular dichroism techniques: biomolecular and nanostructural analyses- a review. *Chemical Biology & Drug Design*, *74*(2), 101–120.
28. Slabinski, L., Jaroszewski, L., Rychlewski, L., Wilson, I. A., Lesley, S. A., & Godzik, A. (2007). XtalPred: a web server for prediction of protein crystallizability. *Bioinformatics*, *23*(24), 3403–3405.
29. Snm, M. M., & Kurgan, L. A. (2012). CRYSpred: accurate sequence-based protein crystallization propensity prediction using sequence-derived structural characteristics. *Protein and Peptide Letters*, *19*(1), 40–49.

30. Takizawa, Y., Binshtein, E., Erwin, A. L., Pyburn, T. M., Mittendorf, K. F., & Ohi, M. D. (2017). While the revolution will not be crystallized, biochemistry reigns supreme. *Protein Science*, 26(1), 69–81.
31. Johnson, W. C., Jr. (1988). Secondary structure of proteins through circular dichroism spectroscopy. *Annual Review of Biophysics and Biophysical Chemistry*, 17(1), 145–166.
32. Ziarek, J. J., Baptista, D., & Wagner, G. (2017). Recent developments in solution nuclear magnetic resonance (NMR)-based molecular biology. *Journal of Molecular Medicine*, 1–8.

Chapter 11

Future of Computational Protein Crystallization

Abstract This book provides the lifecycle of data analytics for protein crystallization. A wide range of topics starting from setting up screens to identifying macromolecular structure has been covered. In earlier chapters, the status-of-art and effective low-cost and real-time techniques for protein crystallization analysis have been provided. This chapter provides some of the challenges and future directions for protein crystallization.

11.1 Introduction

This book presents a diverse set of coherent methods related to data analytics for protein crystallization. The book includes almost complete lifecycle of protein crystallization analysis starting from developing low-cost real-time microscope for image acquisition, analyzing screening results, feature analysis, crystal growth analysis, focusing, segmenting crystallization images, and the visualization of results to identifying macromolecular structure. Some methods such as spatio-temporal crystal growth analysis are still in their infancy. Screening methods such as associative experimental design should be evaluated in other research laboratories. The success of these methods relies on their successful outcomes at different labs.

This book projected light on some of the problems crystallographers may face:

1. how to build a low-cost and real-time microscope that could capture crystallization trial images,
2. screens to be tested for a protein based on the outcomes of prior experiments,
3. useful feature sets for classifying crystallization phases,
4. how to perform spatio-temporal crystal growth analysis in terms of the number of crystals and size of crystals,
5. critical factors for focusing for protein crystallization microscopy,
6. how to benefit from existing thresholding or segmentation methods while working on protein crystal images, and
7. ways of presenting and visualizing results of crystallization experiments.

11.2 Challenges and Future Directions

Large chemical space Analysis of previous protein crystallization trials is essential for effective setting up new screens. Identifying screens from the large chemical space is challenging. Once chemicals to be tested are chosen, trying varying concentrations for each chemical as in full-factorial design may generate myriad cocktails. Effective ranking methods are needed for sorting recommended cocktails.

Workflow In terms of computational methods, there are not many screening methods that analyze prior experiments. The outputs of computational screening methods could be directly fed into a high-throughput system to simplify the process.

Accuracy and Trust The automated systems should gain the trust of protein crystallization experts. It may not be possible to develop a system that has 100% accuracy on classifying protein crystallization trials. Crystallographers do not want the automated systems miss crystals while they may accept misclassification of non-crystals as crystals since those misclassifications have different impacts.

Feature sets and classifiers The feature sets used in classification affect the accuracy as well as overall running time. Our experience shows that using all available features does not necessarily increase the accuracy, and there were cases where the accuracy decreased when all features are fed to the classifier. The use of deep learning for protein crystallization classification analysis may determine features automatically by the deep learning system. Although high accuracy was obtained using deep learning, the likelihood of missing crystals was a down point for the existing proposed CrystalNet architecture.

Necessity for low-cost real-time systems High-throughput systems become essential to conduct large number of experiments. Real-time and low-cost systems with reliable analysis are needed for diverse and wide use of these systems. The transition from systems that are hard to fit in large laboratories to small portable systems is needed for just-in-time analysis of experimental results. Simple user interfaces with any reasonable device should be possible. Universal access for analysis of experimental results can be adopted in this context. The data should be able to analyzed any time, anywhere and by any crystallographer.

Segmenting crystallization images While image thresholding or binarizing images may help determine regions of interest, incorrect binarization yields incorrect analysis. Segmentation is a challenging problem if the images that are captured are so diverse as in protein crystallization domain. Nevertheless, it plays critical factors in feature extraction, classification crystallization phases and crystal growth. A single thresholding technique is expected to fail for some images. Using multiple thresholding techniques to overcome the limitations of a single thresholding technique introduces unnecessary and sometimes noisy results for methods relying on the output of thresholding. Hence, this may increase analysis time by several factors and may not be preferable for real-time analysis. Super-thresholding method selects the best thresholding method for an image and it is likely to provide better results than single

or multiple thresholding methods. This method can benefit from existing available thresholding techniques. It is desired that whatever thresholding technique is used, its completeness and soundness are important. In other words, it should generate acceptable results for all images in the dataset. With a limited number of features and thresholding techniques, super-thresholding provided satisfactory results. By exploring additional features along with other thresholding techniques, its performance could be improved further.

Focusing for trace fluorescence microscopy Focusing is an important problem for the analysis of microscopically obtained protein crystallization images. There may be several crystals in a liquid solution and they may appear at different depths with respect to the microscope's focal length. Moreover, large 3D crystals may have different regions at different depths of field as well. For proper analysis of these images, it is important to generate in focus images. Nevertheless, two assumptions for finding in focus regions may not be true always in this domain: (a) high contrast regions belong to in focus regions and (b) high intensity regions belong to in focus regions. Especially, for trace fluorescence microscopy, since the background is black, out-of-focus high intensity regions form highlighted regions around the boundaries of objects. Moreover, changing the depth of the field also affects the size of visible objects in images. In our experiments, FocusALL method has delivered satisfactory results for both protein crystallization images and other biological images.

Naming inconsistency When analyzing screens from different companies, it was observed that the same chemical appeared with different names. Such naming inconsistency is a handicap for multiple screen analysis for screen kits coming from different companies. It would be great if these companies develop a standard for naming chemicals and digitally storing screen kits.

Visualization displays Display technologies are behind capturing technologies. While 4 K displays are state-of-the-art, images captured at 4 K could be considered at low resolution considering maximum resolution capability of cameras. Moreover, displaying multiple images on the same display screen may not present all details in these images. To overcome these limitations, large research laboratories may consider building display walls where each image can be displayed at their proper resolution. It is unfortunate that images are captured at a very high resolution but cannot be visualized at their native resolution on a display screen without zooming.

We would have a few suggestions for readers as well:

1. Avoid using binary classification as crystals versus non-crystals. There is no trustable classification model that can achieve 100% accuracy for complex data sets without overfitting. Such classification models are not trustable as they may miss crystals. Having additional categories where doubtful categories may fall into would be helpful for experts.
2. Identify the crystallization phases needed for the analysis. Having too many categories could complicate the classification model. Hierarchical classification models may be opted, however, an error at the higher levels of the hierarchy propagate to the lower levels of the hierarchical classification model.

3. Do not use all possible features for the analysis. Some features may act like noise. Moreover, using too many features increase the processing time. Rather identify the best set of features and use those features. Deep learning for protein crystallization is promising, but is still at its early stages.
4. Repeat experiments with a family conditions. Evaluating a screen just once may not yield a successful outcome although it may have the right set of chemicals and concentrations. Crystal nucleation is a stochastic process, and just because there are no crystals in one experiment does not mean that the conditions are not good for crystallization. Too, subtle differences in protein preparations can lead to different outcomes in crystallization screening trials.
5. Do not rely on a single thresholding technique. All thresholding techniques are likely to fail for some image. Rather consider using composite techniques such as super-thresholding that can benefit from multiple techniques.

11.3 Summary

The goal of data analytics for protein crystallization is to reduce the overhead on crystallographers, help them make critical decisions for successful protein crystallization, develop new ideas for crystallization, and assist them by providing proper analysis of crystallization experiments. There were two major reasons for the preparation of the chapters in this book. Firstly, we show that it is not necessary to acquire significantly high cost systems to perform research in this area. Secondly, by providing simple and low-cost techniques, we provide the methodology of building automated systems for small research groups so that motivated researchers in protein crystallization have fair competition as large research groups. The goal of this book was to show the current achievements, current limitations, and possible future work to be considered. In this sense, this book should serve as a motivation for necessary work to be achieved in the future by providing current accomplishments.

The trace fluorescent labeling for protein crystallization analysis may enhance automated high-throughput systems by reducing their time to analyze results as well as increasing their accuracy of the analysis. It is very likely that trace fluorescent labeling will be an industry standard for protein crystallization analysis.

Index

A

Accuracy, 25, 58, 79, 106, 114, 118, 120, 135, 144, 147, 195, 224
Accuracy measures, 106
Amino acid, 2, 8, 211, 217, 218, 220
Artificial neural networks, 35, 36, 91, 95, 185, 188
Associative Experimental Design (AED), 34, 38, 39, 53, 223
Auto-focusing, 151, 152, 155
Auto-focusing, active, 155
Auto-focusing, passive, 155
Automated image scoring, 58, 83
AutoSherlock, 208

B

Batch method, 15
Bayesian classifier, 60, 87, 91, 95, 185, 189, 191, 192
Binding, 2, 54, 212, 213, 218
Binding analysis, 53
Binding reaction, 12
Bin – Recall, 51
Birefringent precipitate, 23
Boundary uniformity, 72
Bray–Curtis dissimilarity measure, 44
Bright spots, 23, 26

C

Categories of classification, 225
Categories of protein crystallization, 28, 58, 86, 105, 110, 126, 200, 205
Chemical cross linking, 217
Chemical distance, 43

Chemical space, 9–11, 33–35, 37, 43, 145, 200, 208, 224
Chemical space mapping, 208
Circular Dichroism (CD), 219
Classification, 29, 35, 58, 60, 75, 77, 86, 87, 94, 95, 103, 105, 126, 135, 146, 152, 175, 177, 185, 199, 224, 225
Classification accuracy, 75, 126, 189
Classification, 3-class (non-crystals, likely-leads, and crystals), 105
Classification, crystals, 69, 111
Classification, likely-leads, 110
Classification, non-crystals, 110
Clear drop, 28, 36, 58, 84, 86, 110
Cocktail distance coefficient (CD_{coeff}), 43
Convolutional neural networks, 115
Co-occurrence matrix, 97
Cryogenic Electron Microscopy (Cryo-EM), 214, 215
Crystal growth, 3, 27, 125, 127, 131, 135, 136, 140, 141, 144, 152, 178, 206, 223
Crystallization of complexes, 17, 145
Crystallization pathway, 3–6, 18, 57
Crystallization phase, 21, 84, 140, 202, 223
Crystallization screening, 2, 4, 5, 9, 11, 12, 14, 33, 46
Crystallization space, 10, 36
Crystal miss, 29, 75, 77, 79, 224
CrystalNet, 115, 116, 224
Crystal nucleation, 3, 5, 9, 26, 53, 226
Crystal size, 54, 126, 131, 132, 146, 153, 216
Crystal size growth, 139, 144, 147
Crystal symmetry, 213
Crystal X2, 46, 61, 63, 64, 68, 91, 103, 113, 117, 119, 131, 188, 193
CrystPro, 130, 144, 146

D

3D crystals, 26, 27
 3D crystals, large, 23, 28, 225
 3D crystals, small, 23, 28
 Decision tree, 60, 87, 91, 92, 118, 146, 182, 185, 188
 Deep convolutional neural network, 60, 87, 88
 Deep learning, 115, 224, 226
 Depth of field, 151, 153, 155
 Dialysis, 13
 Diffraction resolution, 18, 53, 145, 213
 Dimensionality reduction, 92, 118
 Diviner, 128
 Droplet boundary, 58
 Dye-binding assay, 128

E

Edge detection, Canny, 59, 93, 132–134, 179, 182
 Edge detection, Sobel, 59, 86, 93, 126, 169, 170, 172, 173, 175
 Emission filter, 62
 Emission spectrum, 218
 Emission wavelength, 63
 Euler angle, 215
 Euler number, 59, 86
 Excitation, 62, 68, 78
 Excitation source, 131
 Excitation wavelength, 63

F

Family of conditions, 40, 47, 49, 50, 53, 226
 Feature analysis, 88
 Feature extraction, 58, 59, 61, 64–66, 69, 70, 77, 78, 86–88, 91, 96, 103, 112–114, 119, 120, 126, 138, 152, 175, 178, 182, 185–189, 193, 195, 224
 Feature normalization, 92
 Feature reduction, 87, 91, 105, 106, 118
 Feature selection, 84, 92, 93, 107, 118, 119, 185, 194
 Features, graph, 101
 Features, histogram, 78, 83, 86, 88, 91, 96, 98, 106, 109–111, 114, 118, 119, 126, 185, 189, 194
 Features, histogram autocorrelation, 119
 Features, intensity, 69, 96, 118, 194
 Features, regions, 71, 78, 99, 106, 111, 119
 Features, shape adaptive, 101
 Features, spatiotemporal, 139

Features, texture, 59, 86, 88, 98, 112, 126, 185
 Femtosecond X-ray laser, 216
 Fluorescence microscopy, 61, 225
 Fluorescence Resonance Energy Transfer (FRET), 218
 Fluorescent probe, 12, 63, 126, 145, 218
 Focal stacking, 151, 156, 164, 165, 168
 Focal Stacking, Neighborhood Based (NBFS), 158
 Focal Stacking, Pixel Based (PBFS), 158
 Focal stacking, transformation based, 158
 FocusALL, 161, 164, 166, 172, 175, 225
 Full factorial design, 33, 34

G

Gabor wavelet, filter, 59, 86
 Genetic algorithm, 36, 59, 60, 87, 88
 Gray-level co-occurrence matrix, 59, 86, 126, 183
 Grayscale conversion, 67
 Green level co-occurrence matrix, 97
 Grid screening, 10, 11, 34, 35

H

Hampton's score, 91
 Harris Corner Response Measure (HCRM), 159, 160
 Heavy precipitate, 22, 26
 Hierarchical classification, 28, 105, 114, 116, 225
 High-throughput rate, 59, 87
 High-throughput system, 57, 58, 83, 125, 224, 226
 Histogram, 59, 96
 Hough transform, 59, 86, 88, 126
 Hydrophobic-binding dye, 218

I

Image acquisition, 62, 64, 78, 79, 113, 151, 154, 155, 204, 223
 Image down-sampling, 65
 Image processing, 16, 58, 59, 61, 64, 66, 79, 86, 88, 91, 93, 113, 126, 129, 132, 152, 202, 204, 215
 Image registration, 138
 Image segmentation, 64, 67, 177, 179, 224
 Image thresholding, 64, 68, 93, 99, 132, 147, 178, 179, 224, 226
 Incomplete factorial cocktails, 208
 Incomplete factorial design, 125

Incomplete factorial experiment, 10, 34, 36
Incomplete factorial screen, 11, 12, 35
INFAC, 11
Integral histogram, 59, 86
Integral membrane proteins, 17
Intensity histogram, 96, 97, 167, 179, 183
Intensity statistics, 59, 70, 86, 88, 100, 194
IXpressGenes, Inc., 46, 61, 64, 91, 103, 131, 188, 193, 200

L

Labile zone, 3–5, 7
Laplacian, 158, 194
Laplacian filter, 156
Laplacian pyramid filter, 59, 86, 126, 179
Light Emitting Diode (LED), 63, 131
Light precipitate, 24, 26
Light scattering, 5, 6, 212
Linear discriminant analysis, 60, 87, 120
Liquid–liquid diffusion screening, 13

M

MacroScope, 201
Max-Class ensemble method, 76
Mean Decrease in Accuracy (*MDA* – *RF*), 92, 93, 107
Measure of symmetry, 73
Metastable zone, 3, 4
Microcrystal, 28, 36, 58, 84, 86, 110, 126, 215
minimal-Redundancy-Maximal-Relevance (mRMR) criterion, 185
Misclassification, 29, 75, 76, 84, 88, 109, 119, 224
Multilayer perceptron neural network, 75, 76, 88

N

Naming inconsistency, 225
Needle crystals, 26, 27
Neural networks, 35
Neutron diffraction, 28, 54, 212
Noise removal, 66
Non-faceted crystals, 4, 23, 26, 28, 50
Nuclear Magnetic Resonance (NMR), 213, 214
Nucleation, 3, 6, 145
Nucleation event, 31
Nucleation rate, 9

O

Optimization, 18, 26, 36, 42, 47, 49, 53, 54, 57, 58, 63
Optimization crystallization conditions, 31
Optimization experiments, 16, 30
Optimization methods, 35, 40
Optimization of cocktails, 34, 35, 41
Optimization of conditions, 37, 39, 50, 126, 208
Optimization problems, constrained and unconstrained, 36
Optimization screen, 43, 45, 47, 49, 50
Optimization screening, 54
Optimization strategies, 22
Optimization tests, 9
Optimization trials, 131
Optimizing concentration values, 42, 45, 53
Optimizing conditions, 126

P

Phase diagram, 3
Phase separation, 22, 26, 29, 36, 40, 42, 84, 103, 110
Phase transition, 24, 26
Plate, 16, 21, 138, 152
Plate analysis, 145
Plate, crystallization, 24, 63, 128
Plate, crystallization screening, 12
Plate crystals, 27, 47, 111, 179
PlateDB, 208
Plate imaging system, 145
Plate, screening, 16
Plate view, 200, 207
Plate visualization, 200
Plate well, 200
Plate, 96-well, 64, 201
Plate, 1536-well, 201
Precipitate, 3–5, 23, 26, 28, 29, 36, 42, 58, 69, 77, 86, 126, 130, 154
Principle Components Analysis (PCA), 83, 91, 92, 106
Prioritization of reagents, 42
Prioritized cocktails, 43
Prohibited combinations, 41, 42
Protein crystallization, 1, 8, 18, 226
Protein molecule, 2, 8, 12, 126, 128, 213, 216

Q

Quenching, 145

R

Radon-Laplacian, 59, 86
 Random forest, 60, 87, 95, 112, 116, 118, 185, 188
 Random screen, 10, 11
 Ranked category, 110
 Ranking cocktails, 224
 Ranking conditions, 50
 Ranking features, 93, 118, 119
 Ranking methods, 50
 Ranking prioritized cocktails, 42, 43, 53
 Real-time analysis, 64, 84, 88, 129
 Real-time application, 113
 Real-time autofocusing, 156
 Real-time classification, 195
 Real-time classifier, 88
 Real-time microscope, 223
 Real-time system, 57, 126, 224
 Region segmentation, 94, 129, 137
 Robotic image acquisition, 57
 Robotic methods, 14
 Robotic microscopy, 60, 135
 Robotic setup, 9, 16, 57, 61, 125
 Robotic system, 57, 61, 83, 126
 RoCKS, 200, 204, 205
 Running time, 58, 59, 224

S

Scoring, 21, 22, 24, 26, 30, 31, 45, 51, 57, 131, 199, 200, 205
 Scoring, expert, 85
 Scoring, image, 78
 Scoring, levels, 58
 Scoring methods, 58
 Scoring outcome, 39, 42, 46
 Scoring procedure, 22
 Scoring, real-time, 117, 119
 Scoring scale, 21, 22
 Scoring system, 36, 75, 91
 Screen designing, 43
 Screening analysis, 11, 22, 39, 47, 53, 223
 Screening cocktails, 63
 Screening conditions, 11
 Screening experiments, 5, 8, 9, 14, 31, 34, 215, 226
 Screening factors, 39
 Screening, genetic algorithm, 12, 36
 Screening, incomplete factorial design, 34
 Screening kits, 9
 Screening methods, 9, 10, 12, 34, 223, 224
 Screening, neural networks, 35
 Screening process, 8, 9

Screening solution, 21
 Screening solution volume, 9
 Screens, ranking, 50
 Second virial coefficient, 5–7
 Seed crystals, 50
 Self-organizing neural networks, 60, 87, 88
 Shape-adaptive Discrete Cosine Transform (SA-DCT), 101
 Significance ratio, 43
 Solubility line, 3, 7, 22
 Soluble protein zone, 3
 Soundness and completeness, 191, 195
 Sparse matrix sampling, 10, 35
 Sparse matrix screen, 10
 Spatio-temporal analysis, 125, 130, 132, 135, 146, 223
 Stochastic multiparameter optimization, 12
 Super-thresholding, 178, 185, 188, 192, 195, 226
 Super-thresholding, posteriori, 187, 189, 193
 Super-thresholding, priori, 186, 189, 194
 Supervised thresholding, 180
 Support Vector Machines (SVM), 60, 87, 91, 95, 118
 Symmetry, 8

T

Temporal analysis, 128, 129
 Thresholding, global, 178
 Thresholding, green percentile, 67, 68, 94, 179
 Thresholding, local, 178
 Thresholding, morphological, 94
 Thresholding, Otsu's, 67, 76, 77, 93, 114, 132, 136, 177, 185, 189
 Time analysis, 103, 106, 193
 Timing analysis, 75, 112, 116
 Trace fluorescent labeling, 12, 17, 26, 39, 61, 69, 91, 113, 119, 125, 131, 145, 207, 208, 226

V

Vapor diffusion, 14
 Vector quantization, 60, 87
 Visualization, 199
 Visualization display, 225
 Visualization interface, 199, 204
 Visualization software, 199, 200, 203
 Visualization, plate well, 203
 Visualization, screening results, 16
 Visualizaton, color coding, 201, 204, 207

Visualizaton, glyphs, [202](#)
Visualizaton, multiple light sources, [207](#)
Visualizaton, region-of-interest, [202](#)
Visualizaton, sequential view, [207](#)
Visualizaton, thumbnails, [200](#)
Visualizaton, time course, [206](#)
Visual-X2, [200](#), [203](#), [205](#)
Vollath-F4, [156](#), [169](#), [172](#)
Vollath-F5, [156](#)

W
Well-plate, [61](#), [125](#)

X
X-ray beam, [127](#)
X-ray crystallography, [1](#), [126](#), [213](#)
X-ray diffraction, [10](#), [57](#), [153](#), [211](#)
X-ray Free Electron Laser (XFEL), [215](#), [216](#)
XtalPIMS, [200](#), [206](#), [207](#)