

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