Computational Biology

Volume 25

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

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 as 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

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