Shibiao Wan, Man-Wai Mak

Machine Learning for Protein Subcellular Localization Prediction

DE GRUYTER

Authors
Dr. Shibiao Wan
The Hong Kong Polytechnic University
Department of Electronic and Information Engineering
Hung Hom, Kowloon
Hong Kong SAR
shibiao.wan@connect.polyu.hk

Dr. Man-Wai Mak
The Hong Kong Polytechnic University
Department of Electronic and Information Engineering
Hung Hom, Kowloon
Hong Kong SAR
enmwmak@polyu.edu.hk

ISBN 978-1-5015-1048-9
e-ISBN (PDF) 978-1-5015-0150-0
e-ISBN (EPUB) 978-1-5015-0152-4
Set-ISBN (print/ebook) 978-1-5015-0151-7

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://www.dnb.de.

Typesetting: PTP-Berlin, Protago TEX-Production GmbH
Printing and binding: CPI books GmbH, Leck
Cover image: THOMAS DEERINCK, NCMIR/Science Photo Library/gettyimages
♾ Printed on acid-free paper
Printed in Germany

www.degruyter.com

Preface

Proteins, which are essential macromolecules for organisms, need to be located in appropriate physiological contexts within a cell to exhibit tremendous diversity of biological functions. Aberrant protein subcellular localization may lead to a broad range of diseases. Knowing where a protein resides within a cell can give insights into drug target discovery and drug design. This book explores machine-learning approaches to the automatic prediction of protein subcellular localization. The approaches exploit the gene ontology database to extract relevant information. With the ever increasing numbers of new protein sequences in the postgenomic era, machine-learning approaches have become an indispensable tool for assisting the laborious and time-consuming web-lab experiments and for accurate, fast, and large-scale predictions in proteomics research.

Recent years have witnessed an incredibly fast development of molecular biology and computer science, which makes it possible to utilize computational methods to determine the subcellular locations of proteins. It is of paramount significance for wet-lab biologists, bioinformaticians, and computational biologists to be informed of the up-to-date development in this field. Compared to traditional books on protein subcellular localization, this book has the following advantages:

1. This book elaborately presents the latest state-of-the-art machine-learning approaches for protein subcellular localization prediction.
2. This book comprehensively covers many aspects of protein subcellular localization, from single- to multi-label prediction, from prediction of *Homo sapiens* proteins, *Viridiplantae* proteins, *Eukaryota* proteins to prediction of *Virus* proteins.
3. This book systematically introduces three machine-learning approaches to improving predictors' performance, including classification refinement, deeper feature extraction and dimensionality reduction.
4. This book not only proposes several advanced and accurate single- and multi–label predictors but also introduces their easy-to-use online web-servers.

This book is organized into four related parts:

1. Part I – Chapters 1, 2, and 3 – introduces the significance of computationally predicting protein subcellular localization, provides an overview of state-of-the-art approaches, and details the legitimacy of using gene ontology (GO) information for predicting subcellular localization of proteins.
2. Part II – Chapters 4, 5, 6, and 7 – proposes several state-of-the-art predictors for single- and multi-location protein subcellular localization. In Chapter 4, two predictors, namely GOASVM and FusionSVM, both based on GO information, are proposed for single-location protein subcellular localization. Subsequently, multi-location protein subcellular localization is described in Chapter 5. In this chapter, several multi-label predictors, including mGOASVM, AD-SVM, and mPLR-Loc,

which were developed based on different classifiers, are introduced for accurate prediction of subcellular localization of both single- and multi-location proteins. Next, Chapter 6 presents the predictors, namely SS-Loc and HybridGO-Loc, which exploit the deep information embedded in the hierarchical structure of the GO Database. These predictors incorporate the information of semantic similarity over GO terms. For large-scale protein subcellular localization, Chapter 7 introduces ensemble random projection to construct two dimension-reduced multi–label predictors, namely RP-SVM and R3P-Loc. In addition, two compact databases (ProSeq and ProSeq-GO) are proposed to replace the conventional databases (Swiss-Prot and GOA) for fast and efficient feature extraction.

3. Part III – Chapters 8, 9, and 10 – presents the experimental setup and results for all of the proposed predictors and further discusses the properties of the proposed predictors. Chapter 8 details the specific experimental setup, including datasets construction and performance metrics. Extensive experimental results and analyses for all the proposed predictors are detailed in Chapter 9. Further discussions are provided in Chapter 10.
4. Part IV – Chapter 11 – gives a conclusion and possible future directions for further research in this field.

It is confidently believed that this book will provide bioinformaticians and computational biologists with the latest state-of-the-art machine-learning approaches for protein subcellular localization prediction and will enlighten them with a systematic scheme to improve predictors' performance. For wet-lab biologists, this book offers accurate and fast subcellular-localization predictors and easy-to-use online web-servers.

Acknowledgement: This book is an outgrowth of four years of research on the topics of bioinformatics and machine learning. First, the authors would like to express their sincere gratitude and appreciation to Prof. Sun-Yuan Kung from Princeton University, whose insightful comments and invaluable suggestions have facilitated the research.

The authors are also indebted to Prof. Yue Wang from Virginia Tech (VT), USA, and Dr. Zhen Zhang and Dr. Bai Zhang from Johns Hopkins University (JHU), USA. Our gratitude also goes to all of the CBIL members of VT and collaborators at JHU. Deep thanks should also go to Prof. Hong Yan from City University of Hong Kong, Hong Kong SAR, and Dr. Haiying Wang from the University of Ulster, UK. Their critical and constructive suggestions were imperative for the accomplishment of the book.

We are also grateful to senior editorial director Mr. Alexander Greene, project editor Ms. Julia Lauterbach, and project editor Ms. Lara Wysong of the De Gruyter publisher, who have provided professional assistance throughout the project.

Both authors are with the Center for Multimedia Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China. The authors are grateful to the university and the department for their generous and consistent support.

We are pleased to acknowledge that the work presented in this book was in part supported by The Hong Kong Polytechnic University (Grant No. G-YJ86, G-YL78, and G-YN18) and the Research Grant Council of Hong Kong SAR (Grant No. PolyU5264/09E and PolyU 152117/14E).

The authors would also like to thank many collaborators and colleagues, including Wei Wang, Jian Guo, and others.

Particularly, Shibiao Wan would like to give special thanks to his partner Jieqiong Wang for her unreserved love and support. Last but not the least, the authors wish to give their deepest gratitude to their families. Without their generous support and full understanding, this book would not have been so smoothly completed.

Dr. Shibiao Wan is currently a Postdoctoral Fellow of the Department of Electronic and Information Engineering at the Hong Kong Polytechnic University. He obtained his BEng degree in telecommunication engineering from Wuhan University, China in 2010, and his PhD degree in bioinformatics from the Hong Kong Polytechnic University in 2014. He was a visiting scholar in the Virginia Tech and the Johns Hopkins School of Medicine from Spring 2013 to Summer 2013. His current research interests include bioinformatics, computational biology, and machine learning. He has published a number of technical articles on top bioinformatics journals such as *BMC Bioinformatics*, *PLoS ONE*, *Journal of Theoretical Biology*, etc, and key international conferences on signal processing, bioinformatics, and machine learning such as ICASSP, BIBM, MLSP, etc. He serves as a reviewer for a number of journals, such as *IEEE Trans. on Nanobioscience*, *AMC*, *JAM*, *IJBI*, and *IJMLC*.

Dr. Man-Wai Mak is an Associate Professor of the Department of Electronic and Information Engineering at the Hong Kong Polytechnic University. He has authored more than 150 technical articles in speaker recognition, machine learning, and bioinformatics. Dr. Mak is also a coauthor of the postgraduate textbook *Biometric Authentication: A Machine Learning Approach* (Prentice Hall, 2005). He served as a member of the IEEE Machine Learning for Signal Processing Technical Committee from 2005–2007. He has been serving as an associate editor of *IEEE/ACM Trans. on Audio, Speech and Language Processing*, *Journal of Signal Processing Systems*, and *Advances in Artificial Neural Systems*. He has been a technical committee member of a number of international conferences, such as Interspeech, ISCSLP, and IEEE Workshop on MLSP.

Contents

List of Abbreviations

AA	amino-acid compositions
AC	accession number
ACC	overall accuracy
ACR	acrosome
AD	adaptive decision
BLAST	basic local alignment search tool
BR	binary relevance
c-region	C-terminal flanking region
CEL	cell wall
CEN	centrosome
CHL	chloroplast
CM	cell membrane
cTP	chloroplast transit peptide
CYA	cyanelle
CYK	cytoskeleton
CYT	cytoplasm
DLS	distinct label set
EBI	European Bioinformatics Institute
ECC	ensembles of classifier chains
END	endosome
ER	endoplasmic reticulum
EU16	the 16-class eukaryotic dataset
EXP	inferred from experiment
EXT	extracellular
F1	F1-score
GapAA	gapped amino-acid pair compositions
GO	gene ontology
GOA	gene ontology annotation database
GOL	Golgi apparatus
h-region	central hydrophobic region
HCYT	host cytoplasm
HER	host endoplasmic reticulum
HL	Hamming loss
HMMs	hidden Markov models
HNUC	host nucleus
HUM12	the 12-class human dataset
HYD	hydrogenosome
IDA	inferred from direct assay
IEA	inferred from electronic annotation

IMP	inferred from mutant phenotype
IPI	inferred from physical interaction
ISF	inverse sequence-frequency
ISS	inferred from structural and sequence similarity
LC	label cardinality
LCA	lowest common ancestors
LD	label density
LOOCV	leave-one-out cross validation
LP	label powerset
LR	logistic regression
LYS	lysosome
MCC	Mathew's correlation coefficient
MEL	melanosome
MIC	microsome
MIT	mitochondrion
mTP	mitochondrial targeting peptide
n-region	N-terminal flanking region
NE16	the 16-class novel eukaryotic dataset
NNs	neural networks
NUC	nucleus
OAA	overall actual accuracy
OE11	the 11-class old eukaryotic dataset
OET-KNN	optimized evidence-theoretic K-nearest neighbors
OLA	overall locative accuracy
OLS	ordinary least squares
OMCC	overall Mathew's correlation coefficient
PairAA	amino-acid pair compositions
PDLS	proportion of distinct label set
PER	peroxisome
PLA	plastid
PM	plasma membrane
PseAA	pseudo amino-acid compositions
PSFM	position-specific frequency matrix
PSI-BLAST	position-specific iterative BLAST
PSSM	position-specific scoring matrix
RP	random projection
RR	ridge regression
RS	relevance similarity
SEC	secreted
SP	signal peptide
SPI	spindle pole body
SS	semantic similarity

SVMs	support vector machines
SYN	synapse
TF	term-frequency
TF-ISF	term-frequency–inverse sequence-frequency
TLN	total locative number
VAC	vacuole
VC	viral capsid
WAMCC	weighted average Mathew's correlation coefficient