

# **SpringerBriefs in Computer Science**

## **Series editors**

Stan Zdonik, Brown University, Providence, USA  
Shashi Shekhar, University of Minnesota, Minneapolis, USA  
Jonathan Katz, University of Maryland, College Park, USA  
Xindong Wu, University of Vermont, Burlington, USA  
Lakhmi C. Jain, University of South Australia, Adelaide, Australia  
David Padua, University of Illinois Urbana-Champaign, Urbana, USA  
Xuemin (Sherman) Shen, University of Waterloo, Waterloo, Canada  
Borko Furht, Florida Atlantic University, Boca Raton, USA  
V.S. Subrahmanian, University of Maryland, College Park, USA  
Martial Hebert, Carnegie Mellon University, Pittsburgh, USA  
Katsushi Ikeuchi, University of Tokyo, Tokyo, Japan  
Bruno Siciliano, Università di Napoli Federico II, Napoli, Italy  
Sushil Jajodia, George Mason University, Fairfax, USA  
Newton Lee, Newton Lee Laboratories, LLC, Tujunga, USA

More information about this series at <http://www.springer.com/series/10028>

Jinbo Xu · Sheng Wang  
Jianzhu Ma

# Protein Homology Detection Through Alignment of Markov Random Fields

Using MRAlign

 Springer

Jinbo Xu  
Toyota Technological Institute  
Chicago, IL  
USA

Jianzhu Ma  
Toyota Technological Institute  
Chicago, IL  
USA

Sheng Wang  
Toyota Technological Institute  
Chicago, IL  
USA

ISSN 2191-5768 ISSN 2191-5776 (electronic)  
SpringerBriefs in Computer Science  
ISBN 978-3-319-14913-4 ISBN 978-3-319-14914-1 (eBook)  
DOI 10.1007/978-3-319-14914-1

Library of Congress Control Number: 2014960093

Springer Cham Heidelberg New York Dordrecht London  
© The Author(s) 2015

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.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media  
([www.springer.com](http://www.springer.com))

# Preface

This short book is derived from our paper entitled “MRAlign: Protein Homology Detection Through Alignment of Markov Random Fields,” which won the best paper award at a premier computational biology conference RECOMB2014 and also appeared at PLoS Computational Biology. The intended audience consists of students and researchers involved in developing computational methods for biological sequence analysis and protein structure and functional prediction, those who use sequence analysis tools to study biology problems, and those who would like to have a general idea about protein homology detection and fold recognition. We hope that the new Markov Random Fields (MRF) method described in this book will intrigue further study of protein homology detection and fold recognition. We also hope that the tool described in this book will be helpful for readers with biology backgrounds who need to quantify and analyze protein sequences to answer interesting questions.

This book covers sequence-based protein homology detection, a fundamental and challenging bioinformatics problem with a variety of real-world applications. The book first surveys a few popular homology detection methods such as the position-specific scoring matrix (PSSM) or Hidden Markov Model (HMM)-based methods and then is devoted to a novel MRF-based method which was recently developed by our group. Compared to HMM and PSSM, MRF can model long-range residue–residue interaction and thus, MRF-based methods are much more sensitive than HMM- and PSSM-based methods for remote homolog detection and fold recognition. This book also describes the installation, usage, and result interpretation of our programs implementing the MRF-based method.

The book is organized into four chapters. Chapter 1 describes the background and surveys the existing popular methods of homology detection and fold recognition. Chapter 2 describes a novel MRF-based method for homology detection and fold recognition. In particular, it covers how to build an MRF model for a protein sequence, how to score the similarity of two MRF models, and how to generate an MRF–MRF alignment optimizing the scoring function. Chapter 3 is devoted to the software implementing the ideas presented in Chap. 2, covering installation, usage, and result interpretation of the software. Chapter 4 describes the experimental

results of our MRF-based method for homology detection and fold recognition. Finally, conclusions are drawn in the last part of the book.

We are indebted to a few Ph.D. students in our group such as Dr. Jian Peng (now a faculty member at UIUC Computer Science Department), Dr. Feng Zhao, and Mr. Zhiyong Wang. We are also thankful to Dr. Söding, who developed the popular HHpred program for homology detection. It is their previous excellent work that leads to the MRF-based method described in this book.

# Contents

<b>1</b>	<b>Introduction</b>	1
1.1	Background	1
1.2	Related Work	2
1.3	Alignment-Free Methods for Homology Detection and Fold Recognition	2
1.3.1	Generative and Discriminative Learning for Alignment-Free Homology Detection and Fold Recognition	4
1.3.2	Kernel-Based Learning Methods for Alignment-Free Homology Detection	4
1.4	Alignment-Based Methods for Homology Detection and Fold Recognition	5
1.4.1	Sequence Alignment for Homology Detection and Fold Recognition	7
1.4.2	Profile-Based Alignment for Homology Detection and Fold Recognition	8
1.4.3	Scoring Function for Profile-Based Alignment and Homology Detection	9
1.4.4	Scoring Function for Sequence-Profile Alignment and Comparison	10
1.4.5	Scoring Function for Profile-Profile Alignment and Comparison	11
1.5	Contribution of This Book	12
	References	13
<b>2</b>	<b>Method</b>	17
2.1	Modeling a Protein Family Using Markov Random Fields	17
2.2	Estimating the Parameters of Markov Random Fields	18
2.3	Scoring Similarity of Two Markov Random Fields	21

2.4	Node Alignment Potential of Markov Random Fields . . . . .	22
2.5	Edge Alignment Potential of Markov Random Fields . . . . .	24
2.6	Scoring Similarity of One Markov Random Fields and One Template . . . . .	26
2.7	Algorithms for Aligning Two Markov Random Fields. . . . .	26
	References. . . . .	29
<b>3</b>	<b>Software.</b> . . . .	<b>31</b>
3.1	Overview of Program . . . . .	31
3.2	Software Download . . . . .	31
3.3	Feature Files . . . . .	32
3.4	MRFsearch Ranking File. . . . .	33
3.5	Interpreting P-Value . . . . .	34
3.6	Interpreting a Pairwise Alignment. . . . .	35
	References. . . . .	36
<b>4</b>	<b>Experiments and Results.</b> . . . .	<b>37</b>
4.1	Training and Validation Data . . . . .	37
4.2	Test Data. . . . .	38
4.3	Reference-Dependent Alignment Recall. . . . .	39
4.4	Reference-Dependent Alignment Precision. . . . .	41
4.5	Success Rate of Homology Detection and Fold Recognition . . . .	42
4.6	Contribution of Edge Alignment Potential and Mutual Information. . . . .	44
4.7	Running Time . . . . .	45
4.8	Is Our MRAlign Method Overtrained? . . . . .	45
	References. . . . .	47
	<b>Conclusion</b> . . . . .	<b>49</b>
	<b>Acknowledgments</b> . . . . .	<b>51</b>