

Advanced Information and Knowledge Processing

Also in this series

Gregoris Mentzas, Dimitris Apostolou, Andreas Abecker and Ron Young
Knowledge Asset Management
1-85233-583-1

Michalis Vazirgiannis, Maria Halkidi and Dimitrios Gunopulos
Uncertainty Handling and Quality Assessment in Data Mining
1-85233-655-2

Asunción Gómez-Pérez, Mariano Fernández-López and Oscar Corcho
Ontological Engineering
1-85233-551-3

Amo Scharil (Ed.)
Environmental Online Communication
1-85233-783-4

Shichao Zhang, Chengqi Zhang and Xindong Wu
Knowledge Discovery in Multiple Databases
1-85233-703-6

Jason T.L. Wang, Mohammed J. Zaki,
Hannu T.T. Toivonen and Dennis Shasha (Eds)

Data Mining in Bioinformatics

With 110 Figures



Springer

Jason T.L. Wang, PhD
New Jersey Institute of Technology, USA

Mohammed J. Zaki, PhD
Computer Science Department, Rensselaer Polytechnic Institute, USA

Hannu T.T. Toivonen, PhD
University of Helsinki and Nokia Research Center

Dennis Shasha, PhD
New York University, USA

Series Editors

Xindong Wu
Lakhmi Jain

British Library Cataloguing in Publication Data
Data mining in bioinformatics. — (Advanced information and
knowledge processing)
1. Data mining 2. Bioinformatics — Data processing
I. Wang, Jason T. L.
006.3'12
ISBN 1852336714

Library of Congress Cataloging-in-Publication Data
A catalogue record for this book is available from the American Library of Congress.

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

AI&KP ISSN 1610-3947
ISBN 1-85233-671-4 Springer London Berlin Heidelberg
Springer Science+Business Media
springeronline.com

© Springer-Verlag London Limited 2005

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

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

Typesetting: Electronic text files prepared by authors
Printed and bound in the United States of America
34/3830-543210 Printed on acid-free paper SPIN 10886107

Contents

Contributors	ix
Part I. Overview	1
1. Introduction to Data Mining in Bioinformatics	3
1.1 Background.....	3
1.2 Organization of the Book.....	4
1.3 Support on the Web	8
2. Survey of Biodata Analysis from a Data Mining Perspective	9
2.1 Introduction	9
2.2 Data Cleaning, Data Preprocessing, and Data Integration ...	12
2.3 Exploration of Data Mining Tools for Biodata Analysis.....	16
2.4 Discovery of Frequent Sequential and Structured Patterns ...	21
2.5 Classification Methods	24
2.6 Cluster Analysis Methods	25
2.7 Computational Modeling of Biological Networks	28
2.8 Data Visualization and Visual Data Mining.....	31
2.9 Emerging Frontiers	35
2.10 Conclusions.....	38
Part II. Sequence and Structure Alignment	41
3. AntiClustAl: Multiple Sequence Alignment by Antipole Clustering	43
3.1 Introduction	43
3.2 Related Work	45
3.3 Antipole Tree Data Structure for Clustering	47
3.4 AntiClustAl: Multiple Sequence Alignment via Antipoles	48
3.5 Comparing ClustalW and AntiClustAl	51
3.6 Case Study	53
3.7 Conclusions.....	54
3.8 Future Developments and Research Problems	56

4. RNA Structure Comparison and Alignment	59
4.1 Introduction	59
4.2 RNA Structure Comparison and Alignment Models	60
4.3 Hardness Results	67
4.4 Algorithms for RNA Secondary Structure Comparison	67
4.5 Algorithms for RNA Structure Alignment	71
4.6 Some Experimental Results	76
Part III. Biological Data Mining	83
5. Piecewise Constant Modeling of Sequential Data Using Reversible Jump Markov Chain Monte Carlo	85
5.1 Introduction	85
5.2 Bayesian Approach and MCMC Methods	88
5.3 Examples	94
5.4 Concluding Remarks	102
6. Gene Mapping by Pattern Discovery	105
6.1 Introduction	105
6.2 Gene Mapping	106
6.3 Haplotype Patterns as a Basis for Gene Mapping	110
6.4 Instances of the Generalized Algorithm	117
6.5 Related Work	124
6.6 Discussion	124
7. Predicting Protein Folding Pathways	127
7.1 Introduction	127
7.2 Preliminaries	129
7.3 Predicting Folding Pathways	132
7.4 Pathways for Other Proteins	137
7.5 Conclusions	141
8. Data Mining Methods for a Systematics of Protein Subcellular Location	143
8.1 Introduction	144
8.2 Methods	147
8.3 Conclusion	186
9. Mining Chemical Compounds	189
9.1 Introduction	189
9.2 Background	191
9.3 Related Research	193
9.4 Classification Based on Frequent Subgraphs	196
9.5 Experimental Evaluation	204
9.6 Conclusions and Directions for Future Research	213

Part IV. Biological Data Management	217
10. Phyloinformatics: Toward a Phylogenetic Database	219
10.1 Introduction	219
10.2 What Is a Phylogenetic Database For?	222
10.3 Taxonomy	224
10.4 Tree Space	229
10.5 Synthesizing Bigger Trees	230
10.6 Visualizing Large Trees	234
10.7 Phylogenetic Queries	234
10.8 Implementation	239
10.9 Prospects and Research Problems	240
11. Declarative and Efficient Querying on Protein Secondary Structures	243
11.1 Introduction	243
11.2 Protein Format	246
11.3 Query Language and Sample Queries	246
11.4 Query Evaluation Techniques	248
11.5 Query Optimizer and Estimation	252
11.6 Experimental Evaluation and Application of Periscope/PS ² ..	267
11.7 Conclusions and Future Work	271
12. Scalable Index Structures for Biological Data	275
12.1 Introduction	275
12.2 Index Structure for Sequences	277
12.3 Indexing Protein Structures	280
12.4 Comparative and Integrative Analysis of Pathways	283
12.5 Conclusion	295
Glossary	297
References	303
Biographies	327
Index	337

Contributors

Peter Bajcsy

Center for Supercomputing
Applications

University of Illinois at
Urbana-Champaign
USA

Deb Bardhan

Department of Computer Science
Rensselaer Polytechnic Institute
USA

Chris Bystroff

Department of Biology
Rensselaer Polytechnic Institute
USA

Mukund Deshpande

Oracle Corporation
USA

Cinzia Di Pietro

School of Medicine
University of Catania
Italy

Alfredo Ferro

Department of Mathematics and
Computer Science
University of Catania
Italy

Laurie Jane Hammel

Department of Defense
USA

Jiawei Han

Department of Computer Science
University of Illinois at
Urbana-Champaign
USA

Kai Huang

Department of Biological Sciences
Carnegie Mellon University
USA

Donald P. Huddler

Biophysics Research Division
University of Michigan
USA

George Karypis

Department of Computer Science
and Engineering
University of Minnesota
USA

Michihiro Kuramochi

Department of Computer Science
and Engineering
University of Minnesota
USA

Lei Liu

Center for Comparative
and Functional Genomics
University of Illinois at
Urbana-Champaign
USA

Heikki Mannila

Department of Computer Science
Helsinki University of Technology
Finland

Robert F. Murphy

Departments of Biological Sciences
and Biomedical Engineering
Carnegie Mellon University
USA

Vinay Nadimpally

Department of Computer Science
Rensselaer Polytechnic Institute
USA

Päivi Onkamo

Department of Computer Science
University of Helsinki
Finland

Roderic D. M. Page

Division of Environmental
and Evolutionary Biology
Institute of Biomedical and
Life Sciences
University of Glasgow
United Kingdom

Jignesh M. Patel

Electrical Engineering and
Computer Science Department
University of Michigan
USA

Giuseppe Pigola

Department of Mathematics and
Computer Science
University of Catania
Italy

Alfredo Pulvirenti

Department of Mathematics and
Computer Science
University of Catania
Italy

Michele Purrello

School of Medicine
University of Catania
Italy

Marco Ragusa

School of Medicine
University of Catania
Italy

Marko Salmenkivi

Department of Computer Science
University of Helsinki
Finland

Petteri Sevon

Department of Computer Science
University of Helsinki
Finland

Dennis Shasha

Courant Institute of Mathematical
Sciences
New York University
USA

Ambuj K. Singh

Department of Computer Science
University of California at
Santa Barbara
USA

Hannu T. T. Toivonen

Department of Computer Science
University of Helsinki
Finland

Mohammed J. Zaki

Department of Computer Science
Rensselaer Polytechnic Institute
USA

Jason T. L. Wang

Department of Computer Science
New Jersey Institute of Technology
USA

Kaizhong Zhang

Department of Computer Science
University of Western Ontario
Canada

Jiong Yang

Department of Computer Science
University of Illinois at
Urbana-Champaign
USA

Part I

Overview