

# Lecture Notes in Bioinformatics

4316

Edited by S. Istrail, P. Pevzner, and M. Waterman

Editorial Board: A. Apostolico S. Brunak M. Gelfand  
T. Lengauer S. Miyano G. Myers M.-F. Sagot D. Sankoff  
R. Shamir T. Speed M. Vingron W. Wong

Subseries of Lecture Notes in Computer Science

Mehmet M. Dalkilic Sun Kim  
Jiong Yang (Eds.)

# Data Mining and Bioinformatics

First International Workshop, VDMB 2006  
Seoul, Korea, September 11, 2006  
Revised Selected Papers

## Series Editors

Sorin Istrail, Brown University, Providence, RI, USA

Pavel Pevzner, University of California, San Diego, CA, USA

Michael Waterman, University of Southern California, Los Angeles, CA, USA

## Volume Editors

Mehmet M. Dalkilic

Sun Kim

Indiana University

Center for Genomics and Bioinformatics

School of Informatics

Bloomington, IN 47408, USA

E-mail: {dalkilic,sunkim2}@indiana.edu

Jiong Yang

Case Western Reserve University

EECS Department, Department of Epidemiology and Biostatistics

Cleveland, OH, 44106, USA

E-mail: jiong.yang@case.edu

Library of Congress Control Number: 2006938545

CR Subject Classification (1998): H.2.8, I.5, J.3, I.2, H.3, F.1-2

LNCS Sublibrary: SL 8 – Bioinformatics

ISSN 0302-9743

ISBN-10 3-540-68970-2 Springer Berlin Heidelberg New York

ISBN-13 978-3-540-68970-6 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 2006

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper SPIN: 11960669 06/3142 5 4 3 2 1 0

# Preface

This volume contains the papers presented at the inaugural workshop on Data Mining and Bioinformatics at the 32nd International Conference on Very Large Data Bases (VLDB). The purpose of this workshop was to begin bringing together researchers from database, data mining, and bioinformatics areas to help leverage respective successes in each to the others. We also hope to expose the richness, complexity, and challenges in this area that involves mining very large complex biological data that will only grow in size and complexity as genome-scale high-throughput techniques become more routine. The problems are sufficiently different enough from traditional data mining problems (outside of life sciences) that novel approaches must be taken to data mine in this area. The workshop was held in Seoul, Korea, on September 11, 2006.

We received 30 submissions in response to the call for papers. Each submission was assigned to at least three members of the Program Committee. The Program Committee discussed the submission electronically, judging them on their importance, originality, clarity, relevance, and appropriateness to the expected audience. The Program Committee selected 15 papers for presentation. These papers are in the areas of microarray data analysis, bioinformatics system and text retrieval, application of gene expression data, and sequence analysis. Because of the format of the workshop and the high number of submissions, many good papers could not be included. Complementing the contributed papers, the program of VDMB 2006 included an invited talk by Simon Mercer, Program Manager for External Research, with an emphasis on life sciences.

We would like to thank the members of the Program Committee for their hard and expert work. We would also like to thank the VLDB organizers, the external reviewers, the authors, and the participants for their contribution to the continuing success of the workshop. Thanks also to Indiana University School of Informatics for the generous financial support.

October 2006

Mehmet Dalkilic  
Sun Kim  
Jiong Yang  
Program Chairs  
VDMB 2006

# VDMB 2006 Organization

## Program Committee Chairs

Mehmet Dalkilic (Indiana University, USA)  
Sun Kim (Indiana University, USA)  
Jiong Yang (Indiana University, USA)

## Program Committee Members

Mark Adams (Case Western University, USA)  
Xue-wen Chen (University of Kansas, USA)  
Jong Bhak (Korea Bioinformatics Center, Korea)  
Dan Fay (Microsoft/Director Technical Computing, North America)  
Hwan-Gue Cho (Busan National University, Korea)  
Jeong-Hyeon Choi (Indiana University, USA)  
Tony Hu (Drexel University, USA)  
Jaewoo Kang (North Carolina State University, USA)  
George Karypis (University of Minnesota, USA)  
Doheon Lee (KAIST, Korea)  
Jing Li (Case Western Reserve University, USA)  
Yanda Li (Tsinghua University, China)  
Birong Liao (Eli Lilly, USA)  
Li Liao (University of Delaware, USA)  
Huiqing Liu (University of Georgia, USA)  
Lei Liu (University of Illinois at Urbana Champaign, USA)  
Xiangjun (Frank) Liu (Tsinghua University, China)  
Qingming Luo (Huazhong University, China)  
Simon Mercer (Microsoft, USA)  
Jian Pei (Simon Fraser University, Canada)  
Meral Ozsoyoglu (Case Western Reserve University, USA)  
Predrag Radivojac (Indiana University, USA)  
Tetsuo Shibuya (University of Tokyo, Japan)  
Keiji Takamoto (Case Western Reserve University, USA)  
Haixu Tang (Indiana University, USA)  
Anthony Tung (National University of Singapore, Singapore)  
Wei Wang (University of North Carolina at Chapel Hill, USA)  
Mohammed Zaki (Rensselaer Polytechnic Institute, USA)  
Aidong Zhang (State University of New York at Buffalo, USA)

# Table of Contents

Bioinformatics at Microsoft Research . . . . .	1
<i>Simon Mercer</i>	
A Novel Approach for Effective Learning of Cluster Structures with Biological Data Applications . . . . .	2
<i>Miyoung Shin</i>	
Subspace Clustering of Microarray Data Based on Domain Transformation . . . . .	14
<i>Jongeun Jun, Seokkyung Chung, and Dennis McLeod</i>	
Bayesian Hierarchical Models for Serial Analysis of Gene Expression . . . .	29
<i>Seungyoon Nam, Seungmook Lee, Sanghyuk Lee, Seokmin Shin, and Taesung Park</i>	
Applying Gaussian Distribution-Dependent Criteria to Decision Trees for High-Dimensional Microarray Data . . . . .	40
<i>Raymond Wan, Ichigaku Takigawa, and Hiroshi Mamitsuka</i>	
A Biological Text Retrieval System Based on Background Knowledge and User Feedback . . . . .	50
<i>Meng Hu and Jiong Yang</i>	
Automatic Annotation of Protein Functional Class from Sparse and Imbalanced Data Sets . . . . .	65
<i>Jaehee Jung and Michael R. Thon</i>	
Bioinformatics Data Source Integration Based on Semantic Relationships Across Species . . . . .	78
<i>Badr Al-Daihani, Alex Gray, and Peter Kille</i>	
An Efficient Storage Model for the SBML Documents Using Object Databases . . . . .	94
<i>Seung-Hyun Jung, Tae-Sung Jung, Tae-Kyung Kim, Kyoung-Ran Kim, Jae-Soo Yoo, and Wan-Sup Cho</i>	
Identification of Phenotype-Defining Gene Signatures Using the Gene-Pair Matrix Based Clustering . . . . .	106
<i>Chung-Wein Lee, Shuyu Dan Li, Eric W. Su, and Birong Liao</i>	
TP+Close: Mining Frequent Closed Patterns in Gene Expression Datasets . . . . .	120
<i>YuQing Miao, GuoLiang Chen, Bin Song, and ZhiHao Wang</i>	

Exploring Essential Attributes for Detecting MicroRNA Precursors from Background Sequences .....	131
<i>Yun Zheng, Wynne Hsu, Mong Li Lee, and Limsoon Wong</i>	
A Gene Structure Prediction Program Using Duration HMM .....	146
<i>Hongseok Tae, Eun-Bae Kong, and Kiejung Park</i>	
An Approximate de Bruijn Graph Approach to Multiple Local Alignment and Motif Discovery in Protein Sequences .....	158
<i>Rupali Patwardhan, Haixu Tang, Sun Kim, and Mehmet Dalkilic</i>	
Discovering Consensus Patterns in Biological Databases.....	170
<i>Mohamed Y. ElTabakh, Walid G. Aref, Mourad Ouzzani, and Mohamed H. Ali</i>	
Comparison of Modularization Methods in Application to Different Biological Networks .....	185
<i>Zhuo Wang, Xin-Guang Zhu, Yazhu Chen, Yixue Li, and Lei Liu</i>	
<b>Author Index</b> .....	197