Lecture Notes in Bioinformatics 4146

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

Subseries of Lecture Notes in Computer Science

Jagath C. Rajapakse Limsoon Wong
Raj Acharya (Eds.)

Pattern Recognition in Bioinformatics

International Workshop, PRIB 2006
Hong Kong, China, August 20, 2006
Proceedings

Springer

Volume Editors

Jagath C. Rajapakse
Nanyang Technological University
BioInformatics Research Centre, Singapore
E-mail: asjagath@ntu.edu.sg

Limsoon Wong
National University of Singapore
School of Computing and Graduate School for Integrated Sciences and Engineering
3 Science Drive 2, 117543, Singapore
E-mail: wongls@nus.edu.sg

Raj Acharya
Penn. State University
Computer Science and Engineering
220 Pond Lab., University Park, Pennsylvania 16802-6106, USA
E-mail: acharya@cse.psu.edu

Library of Congress Control Number: 2006930615

CR Subject Classification (1998): H.2.8, I.5, I.4, J.3, I.2, H.3, F.1-2

LNCS Sublibrary: SL 8 – Bioinformatics

ISSN 0302-9743
ISBN-10 3-540-37446-9 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-37446-6 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11818564 06/3142 5 4 3 2 1 0

Preface

The field of bioinformatics has two main objectives: the creation and maintenance of biological databases, and the discovery of knowledge from life sciences data in order to unravel the mysteries of biological function, leading to new drugs and therapies for human disease. Life sciences data come in the form of biological sequences, structures, pathways, or literature. One major aspect of discovering biological knowledge is to search, predict, or model specific patterns of a given dataset, which have some relevance to an important biological phenomenon or another dataset. To date, many pattern recognition algorithms have been applied or catered to address a wide range of bioinformatics problems. The 2006 Workshop of Bioinformatics in Pattern Recognition (PRIB 2006) marks the beginning of a series of workshops that is aimed at gathering researchers applying pattern recognition algorithms in an attempt to resolve problems in computational biology and bioinformatics.

This volume presents the proceedings of Workshop PRIB 2006 held in Hong Kong, China, on August 20, 2006. It includes 19 technical contributions that were selected by the Program Committee from 43 submissions. We give a brief introduction to pattern recognition in bioinformatics in the first paper. The rest of the volume consists of three parts. Part 1: signal and motif detection, and gene selection. Part 2: models of DNA, RNA, and protein structures. Part 3: biological databases and imaging.

Part 1 of the proceedings contains eight chapters that deal with detection of signals, motifs, and gene structure of genomic sequences and gene selection from microarray data. Ryo et al. suggest an approach to derive rules for alphabet indexing to predict the position of N-myristoylation signal by using decision trees. Stepanova, Lin, and Lin present an approach to recognize steroid hormone regulation elements within promoters of vertebrate genomes, based on a hidden Markov model (HMM). Ho and Rajapakse present a novel graphical approach for weak motif detection in noisy datasets. They examine the robustness of the approach on synthetic datasets and illustrate its applicability to find the motifs in eukaryotes.

Hsieh et al. propose a program, GeneAlign, that predicts genes on one genome by incorporating annotated genes on another genome. This approach achieves higher accuracies of gene prediction by employing the conservation of gene structures and sequence homologies between protein coding regions of genomes. Logeswaran, Ambikairajah, and Epps propose a method for predicting short initial exons, based on the weight arrays and CpG islands.

Chua, Ivshina, and Kuznetsov propose a mixture probability model for microarray signals. The noise term due to non-specific mRNA hybridization was modeled by a lognormal distribution; and the true signal was described by the generalized Pareto-gamma function. The model, applied to expression data of 251 human breast cancer tumors on the Affymetrix microarray platform, yields accurate fits for all tumor samples. Using the degree of differential prioritization between relevance and anti-redundancy on microarray data, Ooi, Chetty, and Teng propose a feature selection technique for tumor classification. Kim and Gao propose an enhanced Max-Relevance criterion for gene selection, which combines the collective impact of the most expressive features in emerging patterns (EPs) and independent criteria such as *t*-test or symmetrical uncertainty. By capturing the joint effect of features with EPs algorithm, the method finds the most discriminative features in a broader scope.

Part 2 of the proceedings focuses on the prediction of different models of DNA, RNA, and amino acids to predict protein secondary structure, protein subcellular localization, RNA structure, phylogeny, and nucleosome formation. Loong and Mishra investigate the topological properties of synthetic RNAs by applying a spectral graph partitioning technique. Their analysis shows that the majority of synthetic RNAs possess two to six vertices, in contrast to natural RNA structures that mostly have nine or ten vertices, and are less compact with the second eigenvalue below unity. Gassend et al. propose a biophysically-motivated energy model through the use of hidden Markov support vector machines (HM-SVMs) for protein secondary structure prediction from amino acid sequences.

Shi et al. construct three types of moment descriptors to obtain sequence order information in a protein sequence to predict the subcellular localization of proteins, without needing the information of physicochemical properties of amino acids. Karim, Parida, and Lakhotia explore the use of permutation patterns from genome rearrangement data as a content similarity measure to infer phylogenies, in polynomial time.

Part 3 of the proceedings deals with biological databases and images. Sette et al. announce the availability of the Immune Epitope Database and Analysis Resource (IEDB) to facilitate the exploration of immunity to infectious diseases, allergies, autoimmune diseases, and cancer. The utility of the IEDB was recently demonstrated through a comprehensive analysis of all current information regarding antibody and T cell epitopes derived from influenza A and determining possible cross-reactivity among H5N1 avian flu and human flu viruses. Zhang, Ng, and Bajic combine information of protein functional domains and gene ontology descriptions for highly accurate identification of transcription factor entries in Swiss-Prot and Entrez gene databases. Lam et al. propose a novel method to support automatic incremental updating of specialist biological databases by using association rule mining.

Wang et al. report a blind source separation method, based on non-negative least-correlated component analysis (nLCA), for quantitative dissection of mixed yet correlated biomarker patterns in cellular images. Two approaches for handling large-scale biological data were proposed by Havukkala et al. and illustrated in the contexts of molecular image processing for chemoinformatics and fractal visualization methods for genome analyses. Smolinski et al. investigate hybridization of the multi-objective evolutionary algorithms (MOEA) and rough sets (RS) for the classificatory decomposition of signals recorded from the surface of the cerebral cortex. By using independent component analysis (ICA) to initialize the MOEA, reconstruction errors are significantly improved.

We would like to sincerely thank all authors who have spent time and effort to make important contributions to this book. Our gratitude also goes to the LNBI editors, Sorin Istrail, Pavel Pevzner, and Michael Waterman, for their most kind support and help in editing this book.

Jagath C. Rajapakse
Limsoon Wong
Raj Acharya

Acknowledgement

We would like to thank all individuals and institutions who contributed to the success of the workshop, especially the authors for submitting papers and the sponsors for generously providing financial support. We are very grateful to the IAPR Technical Committee (TC-20) on Pattern Recognition for BioInformatics for their invaluable guidance and advice. In addition, we would like to express our gratitude to all PRIB 2006 Program Committee members for their thoughtful and rigorous reviews of the submitted papers. We fully appreciate the Organizing Committee for their enormous and excellent work.

We are also grateful to the ICPR 2006 General Chairs, Yuan Yan Tang, Patrick Wang, G. Lorette, and Daniel So Yeung, for their willingness to coordinate with PRIB 2006, and, especially to ICPR 2006 Workshop Chairs, James Kwok and Nanning Zheng, for their effort in the local arrangements. Many thanks go to PRIB 2006 secretary, Norhana Ahmad, for coordinating all the logistics of the workshop. Last but not least, we wish to convey our sincere thanks to Springer for providing excellent support in preparing this volume.

Raj Acharya
PRIB 2006 General Chair

Jagath C. Rajapakse
Limsoon Wong
PRIB 2006 Program Co-chairs

Organization

IAPR Technical Committee on Pattern Recognition on Bioinformatics

Raj Acharya (Chair)	Pennsylvania State Univ., USA
Fransisco Azuaje	Univ. of Ulster, UK
Vladimir Brusic	Univ. of Queensland, Australia
Phoebe Chen	Deakin University, Australia
David Corne	Heriot-Watt Univ., UK
Elena Marchiori	Vrije Univ. of Amsterdam, The Netherlands
Mariofanna Milanova	Univ. of Arkansas at Little Rock, USA
Gary B. Fogel	Natural Selection, Inc., USA
Saman K. Halgamuge	Univ. of Melbourne, Australia
Visakan Kadirkamanathan	Univ. of Sheffield, UK
Nik Kasabov	Auckland Univ. of Technology, New Zealand
Irwin King	Chinese Univ. of Hong Kong, Hong Kong
Alex V. Kochetov	Russian Academy of Sciences, Russia
Graham Leedham	Nanyang Tech. Univ., Singapore
Ajit Narayanan	Univ. of Exeter, UK
Nikhil R. Pal	Indian Statistical Inst., India
Marimuthu Palaniswami	Univ. of Melbourne, Australia
Jagath C. Rajapakse (Vice-chair)	Nanyang Tech. Univ., Singapore
Gwenn Volkert	Kent State Univ., USA
Roy E. Welsch	Massachusetts Inst. of Technology, USA
Kay C. Wiese	Simon Fraser Univ., Canada
Limsoon Wong	National Univ. of Singapore, Singapore
Jiahua (Jerry) Wu	Wellcome Trust Sanger Inst., UK
Yanqing Zhang	Georgia State Univ., USA
Qiang Yang	Hong Kong Univ. of Science and Technology, Hong Kong

PRIB 2006 Organization

General Chair

Raj Acharya	Pennsylvania State Univ., USA

Program Co-chairs

Jagath C. Rajapakse (Co-chair)	Nanyang Tech. Univ., Singapore
Limsoon Wong (Co-chair)	National Univ. of Singapore, Singapore

Publicity

Phoebe Chen	Deakin University, Australia
Elena Marchiori	Vrije Univ. of Amsterdam, The Netherlands
Mariofanna Milanova	Univ. of Arkansas at Little Rock, USA

Publication

Loi Sy Ho	Nanyang Tech. Univ., Singapore

Local Arrangement Chair

Irwin King	Chinese Univ. of Hong Kong, Hong Kong

Secretariat

Norhana Binte Ahmad	Nanyang Tech. Univ., Singapore

System Administration

Linda Ang Ah Giat	Nanyang Tech. Univ., Singapore

Program Committee

Shandar Ahmad	Kyushu Inst. of Technology, Japan
Tatsuya Akutsu	Kyoto Univ., Japan
Ron Appel	Swiss Inst. of Bioinformatics, Switzerland
Vladimir Brusic	Univ. of Queensland, Australia
Madhu Chetty	Monash Univ., Australia
Francis Y.L. Chin	Univ. of Hong Kong, Hong Kong
Koon Kau Byron Choi	Nanyang Tech. Univ., Singapore
Ching Ming Maxey Chung	National Univ. of Singapore, Singapore
Carlos Cotta	Univ. of Malaga, Spain
David Corne	Heriot-Watt Univ., UK
Alexandru Floares	Inst. of Oncology, Romania
Gary B. Fogel	Natural Selection, Inc., USA
Vivekanand Gopalkrishnan	Nanyang Tech. Univ., Singapore

Saman K. Halgamuge	Univ. of Melbourne, Australia
Dongsoo Han	Information and Communications Univ., Korea
Yulan He	Nanyang Tech. Univ., Singapore
Hsuan-Cheng Huang	National Yang-Ming Univ., Taiwan
Ming-Jing Hwang	Academia Sinica, Taiwan
Visakan Kadirkamanathan	Univ. of Sheffield, UK
Nik Kasabov	Auckland Univ. of Technology, New Zealand
Alex V. Kochetov	Russian Academy of Sciences, Russia
Natalio Krasnogor	Univ. of Nottingham, UK
Chee Keong Kwoh	Nanyang Tech. Univ., Singapore
Tak-Wah Lam	Univ. of Hong Kong, Hong Kong
Jinyan Li	Inst. of Infocomm Research, Singapore
Alan Wee-Chung Liew	Chinese Univ. of Hong Kong, Hong Kong
Feng Lin	Nanyang Tech. Univ., Singapore
Gary F. Marcus	New York Univ., USA
Hiroshi Matsuno	Yamaguchi Univ., Japan
Satoru Miyano	Univ. of Tokyo, Japan
Jason H. Moore	Dartmouth Medical School, USA
Kenta Nakai	Univ. of Tokyo, Japan
Ajit Narayanan	Univ. of Exeter, UK
Zoran Obradovic	Temple Univ., USA
Marimuthu Palaniswami	Univ. of Melbourne, Australia
Laxmi Parida	IBM T.J. Watson Research Center, USA
Mihail Popescu	Univ. of Missouri, USA
Predrag Radivojac	Indiana Univ., USA
Jem Rowland	Univ. of Wales Aberystwyth, UK
Alexander Schliep	Max Planck Inst. for Mol. Genetics, Germany
Bertil Schmidt	Nanyang Tech. Univ., Singapore
Alessandro Sette	La Jolla Inst. for Allergy & Immunology, USA
Roberto Tagliaferri	Universita di Salerno, Italy
Gwenn Volkert	Kent State Univ., USA
Michael Wagner	Cincinnati Children's Hospital Research Foundation, USA
Haiying Wang	Univ. of Ulster at Jordanstown, N. Ireland
Lusheng Wang	City Univ. of Hong Kong, Hong Kong
Wei Wang	Fudan Univ., China
Banzhaf Wolfgang	Memorial Univ. of Newfoundland, Canada
Jiahua (Jerry) Wu	Wellcome Trust Sanger Inst., UK
Ying Xu	Univ. of Georgia, USA
Hong Yan	City Univ. of Hong Kong, Hong Kong
Yanqing Zhang	Georgia State Univ., USA
Jun Zhang	Nanyang Tech. Univ., Singapore

Table of Contents

Part 1: Signal and Motif Detection; Gene Selection

Part 2: Models of DNA, RNA, and Protein Structures

Part 3: Biological Databases and Imaging