Lecture Notes in Bioinformatics 2812

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

Springer
Berlin
Heidelberg
New York
Hong Kong
London
Milan
Paris
Tokyo

Gary Benson Roderic Page (Eds.)

Algorithms in Bioinformatics

Third International Workshop, WABI 2003
Budapest, Hungary, September 15-20, 2003
Proceedings

Springer

Series Editors

Sorin Istrail, Celera Genomics, Applied Biosystems, Rockville, MD, USA
Pavel Pevzner, University of California, San Diego, CA, USA
Michael Waterman, University of Southern California, Los Angeles, CA, USA

Volume Editors

Gary Benson
The Mount Sinai School of Medicine
Department of Biomathematical Sciences
Box 1023, One Gustave L. Levy Place, New York, NY 10029-6574, USA
E-mail: benson@camelot.mssm.edu

Roderic Page
University of Glasgow, Institute of Biomedical and Life Sciences
Division of Environmental and Evolutionary Biology
Glasgow G12 8QQ, Scotland
E-mail: r.page@bio.gla.ac.uk

Cataloging-in-Publication Data applied for

A catalog record for this book is available from the Library of Congress.

Bibliographic information published by Die Deutsche Bibliothek.
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.ddb.de>.

CR Subject Classification (1998): F.1, F.2.2, E.1, G.1, G.2, G.3, J.3

ISSN 0302-9743
ISBN 3-540-20076-2 Springer-Verlag 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-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science+Business Media GmbH

http://www.springer.de

© Springer-Verlag Berlin Heidelberg 2003
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Boller Mediendesign
Printed on acid-free paper SPIN: 10949993 06/3142 5 4 3 2 1 0

Preface

We are pleased to present the proceedings of the *Third Workshop on Algorithms in Bioinformatics (WABI 2003)*, which took place on September 15–20, 2003 in Budapest, Hungary. The WABI workshop was part of the four-conference meeting, ALGO 2003, which was locally organized by Dr. János Csirik, Head of the Department of Computer Science, József Attila University, Budapest. See `http://www.conferences.hu/ALGO2003/algo_2003.htm` for more details.

WABI focuses on discrete algorithms that address important problems in molecular biology, genomics, and genetics, that are founded on sound models, that are computationally efficient, that have been implemented and tested in simulations and on real datasets, and that provide new biological results. The workshop goals are to present recent research and identify and explore directions for future research.

We received 78 submissions in response to the call for papers and 36 were accepted. We would like to sincerely thank the authors of all submitted papers and the conference participants. We especially thank a terrific program committee for their diligent and thorough work in reviewing and selecting the papers. We were fortunate to have on the program committee the following distinguished group of researchers:

Amihood Amir (Bar Ilan University, Israel)
Alberto Apostolico (Purdue University)
Pierre Baldi (University of California, Irvine)
Gary Benson (Mount Sinai School of Medicine, New York; Co-chair)
Benny Chor (Tel Aviv University)
Nadia El-Mabrouk (University of Montreal)
Olivier Gascuel (LIRMM-CNRS, Montpellier)
Raffaele Giancarlo (Università di Palermo)
David Gilbert (University of Glasgow)
Jan Gorodkin (The Royal Veterinary and Agricultural University, Denmark)
Roderic Guigó (Institut Municipal d'Investigacions Mèdiques, Barcelona)
Dan Gusfield (University of California, Davis)
Jotun Hein (University of Oxford)
Daniel Huson (Tübingen University)
Simon Kasif (Boston University)
Gregory Kucherov (INRIA-Lorraine/LORIA)
Gad Landau (University of Haifa)
Thierry Lecroq (Université de Rouen)
Bernard M.E. Moret (University of New Mexico, Albuquerque)
Vincent Moulton (Uppsala University, Sweden)
Roderic Page (University of Glasgow; Co-chair)
Sophie Schbath (INRA, Jouy-en-Josas)

Charles Semple (University of Canterbury, New Zealand)
Jens Stoye (Universität Bielefeld)
Fengzhu Sun (University of Southern California)
Alfonso Valencia (Centro Nacional de Biotecnología, Madrid)
Jacques Van Helden (Université Libre de Bruxelles)
Louxin Zhang (National University of Singapore)

The program committee's work was greatly assisted by the helpful reviews provided by Federico Abascal, Ali Al-Shahib, Rumen Andonov, Ora Arbell, Sebastian Böcker, David Bryant, Peter Calabrese, Robert Castelo, Kwok Pui Choi, Miklós Csürös, Minghua Deng, Tobias Dezulian, Nadav Efraty, P.L. Erdos, Revital Eres, Eleazer Eskin, Jose Maria Fernandez, Pierre Flener, Jakob Fredslund, Dan Gieger, Robert Giegerich, Vladimir Grebinskiy, Pawel Herzyk, Mark Hoebeke, Robert Hoffmann, Katharina Huber, Michael Kaufmann, Carmel Kent, Jens Lagergren, Yinglei Lai, Xiaoman Li, Gerton Lunter, Rune Lyngsoe, Laurent Mouchard, Pierre Nicolas, Laurent Noe, Sebastian Oehm, Christian N.S. Pedersen, Johann Pelfrene, Shalom Rackovsky, Mathieu Raffinot, Kim Roland Rasmussen, Christian Rausch, Knut Reinert, Olivier Sand, Klaus-Bernd Schürmann, Steven Skiena, Dina Sokol, Y.S. Song, W. Szpankowski, Helene Touzet, Michael Tress, Juris Viksna, Lusheng Wang, Zohar Yakhini, Kui Zhang, and Michal Ziv-Ukelson.

We also thank the WABI steering committee, Olivier Gascuel, Raffaele Giancarlo, Roderic Guigó, Dan Gusfield, and Bernard Moret, for inviting us to co-chair the conference and for their help in carrying out that task.

We are particularly indebted to Kevin Kelliher of the Mount Sinai School of Medicine, New York, for his conscientious administration of the CyberChair software used to manage the review process, and Robert Castelo at Universitat Pompeu Fabra, Barcelona, for generously sharing the scripts he developed to generate the final copy of last year's WABI. Production of the proceedings was greatly assisted by Richard Koch's wonderful TeXShop software, and large quantities of Peet's coffee.

Thanks again to all who participated to make WABI 2003 a success. It has been, for us, a challenging and rewarding experience.

July 2003 Gary Benson and Roderic Page

Table of Contents

Comparative Genomics

A Local Chaining Algorithm and Its Applications in Comparative Genomics 1
M.I. Abouelhoda (University of Bielefeld) and E. Ohlebusch (University of Ulm)

Common Intervals of Two Sequences 17
G. Didier (Laboratoire Genome et Informatique - CNRS UMR 8116)

Database Searching

A Systematic Statistical Analysis of Ion Trap Tandem Mass Spectra in View of Peptide Scoring 25
J. Colinge, A. Masselot, and J. Magnin (GeneProt Inc.)

Vector Seeds: An Extension to Spaced Seeds Allows Substantial Improvements in Sensitivity and Specificity 39
B. Brejová, D.G. Brown, and T. Vinař (University of Waterloo)

Gene Finding and Expression

A Stochastic Approach to Count RNA Molecules Using DNA Sequencing Methods 55
B. Hollas and R. Schuler (Universität Ulm)

A Method to Detect Gene Structure and Alternative Splice Sites by Agreeing ESTs to a Genomic Sequence 63
P. Bonizzoni (Università degli Studi di Milano - Bicocca), G. Pesole (Università degli Studi di Milano), and R. Rizzi (Università degli Studi di Milano - Bicocca)

Optimal DNA Signal Recognition Models with a Fixed Amount of Intrasignal Dependency 78
B. Brejová, D.G. Brown, and T. Vinař (University of Waterloo)

Genome Mapping

New Algorithm for the Simplified Partial Digest Problem 95
J. Błażewicz and M. Jaroszewski (Poznań University of Technology)

Noisy Data Make the Partial Digest Problem NP-hard 111
M. Cieliebak (ETH Zurich), S. Eidenbenz (Los Alamos National Laboratory), and P. Penna (ETH Zurich)

Pattern and Motif Discovery

Pattern Discovery Allowing Wild-Cards, Substitution Matrices, and Multiple Score Functions 124
A. Mancheron and I. Rusu (Université de Nantes)

A Combinatorial Approach to Automatic Discovery of Cluster-Patterns . . 139
R. Eres, G.M. Landau (Haifa University), and L. Parida (IBM TJ Watson Research Center, Yorktown Heights)

Dynamic Programming Algorithms for Two Statistical Problems in Computational Biology 151
S. Rahmann (Max-Planck-Institute for Molecular Genetics)

Phylogenetic Analysis

Consensus Networks: A Method for Visualising Incompatibilities in Collections of Trees 165
B. Holland (Massey University) and V. Moulton (Uppsala University)

Efficient Generation of Uniform Samples from Phylogenetic Trees 177
P. Kearney (Caprion Pharmaceuticals), J.I. Munro, and D. Phillips (University of Waterloo)

New Efficient Algorithm for Detection of Horizontal Gene Transfer Events 190
A. Boc and V. Makarenkov (Université du Québec)

Ancestral Maximum Likelihood of Evolutionary Trees Is Hard 202
L. Addario-Berry (McGill University), B. Chor (Tel-Aviv University), M. Hallett (McGill University), J. Lagergren (KTH, Stockholm), A. Panconesi (Universitá di Roma "La Sapienza"), and T. Wareham (Memorial University of Newfoundland)

A Linear-Time Majority Tree Algorithm 216
N. Amenta (University of California, Davis), F. Clarke, and K. St. John (City University of New York)

Bayesian Phylogenetic Inference under a Statistical Insertion-Deletion Model 228
G. Lunter, I. Miklós, A. Drummond (University of Oxford), J.L. Jensen (University of Aarhus), and J. Hein (University of Oxford)

Better Hill-Climbing Searches for Parsimony 245
G. Ganapathy, V. Ramachandran, and T. Warnow (University of Texas)

Computing Refined Buneman Trees in Cubic Time 259
G.S. Brodal, R. Fagerberg (University of Aarhus), A. Östlin (University of Copenhagen), C.N.S. Pedersen (University of Aarhus), and S.S. Rao (University of Waterloo)

Distance Corrections on Recombinant Sequences 271
D. Bryant (University Montréal), D. Huson, T. Kloepper, and K. Nieselt-Struwe (University of Tuebingen)

Parsimonious Reconstruction of Sequence Evolution and Haplotype Blocks 287
Y.S. Song and J. Hein (University of Oxford)

Polymorphism

Identifying Blocks and Sub-populations in Noisy SNP Data 303
G. Kimmel (Tel-Aviv University), R. Sharan (International Computer Science Institute, Berkeley), and R. Shamir (Tel-Aviv University)

Designing Optimally Multiplexed SNP Genotyping Assays 320
Y. Aumann, E. Manisterski (Bar-Ilan University), and Z. Yakhini (Agilent Laboratories)

Minimum Recombinant Haplotype Configuration on Tree Pedigrees 339
K. Doi (University of Tokyo), J. Li, and T. Jiang (University of California)

Protein Structure

Efficient Energy Computation for Monte Carlo Simulation of Proteins ... 354
I. Lotan, F. Schwarzer, and J.-C. Latombe (Stanford University)

Speedup LP Approach to Protein Threading via Graph Reduction 374
J. Xu (University of Waterloo)

Homology Modeling of Proteins Using Multiple Models and Consensus Sequence Alignment 389
J.C. Prasad, M. Silberstein, C.J. Camacho, and S. Vajda (Boston University)

Side-Chain Structure Prediction Based on Dead-End Elimination: Single Split DEE-criterion Implementation and Elimination Power 402
J.A. Spriet (KULAK)

Sequence Alignment

A Large Version of the Small Parsimony Problem 417
J. Fredslund (University of Aarhus), J. Hein (University of Oxford), and T. Scharling (University of Aarhus)

Optimal Multiple Parsimony Alignment with Affine Gap Cost Using a Phylogenetic Tree 433
B. Knudsen (University of Florida)

Composition Alignment 447
G. Benson (The Mount Sinai School of Medicine)

String Algorithms

Match Chaining Algorithms for cDNA Mapping 462
T. Shibuya (IBM Tokyo Research Laboratory) and I. Kurochkin (RIKEN Genomic Sciences Center)

Sequencing from Compomers: Using Mass Spectrometry for DNA De-Novo Sequencing of 200+ nt 476
S. Böcker (Technische Fakultät, Universität Bielefeld)

Bounds for Resequencing by Hybridization 498
D. Tsur (Tel Aviv University)

Selecting Degenerate Multiplex PCR Primers 512
R. Souvenir, J. Buhler, G. Stormo, and W. Zhang (Washington University in St. Louis)

Author Index 527