Lecture Notes in Computer Science 2452

Edited by G. Goos, J. Hartmanis, and J. van Leeuwen

Springer
Berlin
Heidelberg
New York
Barcelona
Hong Kong
London
Milan
Paris
Tokyo

Roderic Guigó   Dan Gusfield (Eds.)

Algorithms in Bioinformatics

Second International Workshop, WABI 2002
Rome, Italy, September 17-21, 2002
Proceedings

Springer

Series Editors

Gerhard Goos, Karlsruhe University, Germany
Juris Hartmanis, Cornell University, NY, USA
Jan van Leeuwen, Utrecht University, The Netherlands

Volume Editors

Roderic Guigó
IMIM-UPF-CRG, Dr. Aiguader 80, 08003 Barcelona, Spain
E-mail: rguigo@imim.es

Dan Gusfield
Department of Computer Science, University of California
Davis CA 95616, U.S.A.
E-mail: gusfield@cs.ucdavis.edu

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Algorithms in bioinformatics : second international workshop ; proceedings /
WABI 2002, Rome, Italy, September 17 - 21, 2002. Roderic Guigó ; Dan
Gusfield (ed.). - Berlin ; Heidelberg ; New York ; Hong Kong ; London ;
Milan ; Paris ; Tokyo : Springer, 2002
(Lecture notes in computer science ; Vol. 2452)
ISBN 3-540-44211-1

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

ISSN 0302-9743
ISBN 3-540-44211-1 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 2002
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Olgun Computergrafik
Printed on acid-free paper SPIN: 10871128 06/3142 5 4 3 2 1 0

Preface

We are pleased to present the proceedings of the *Second Workshop on Algorithms in Bioinformatics (WABI 2002)*, which took place on September 17-21, 2002 in Rome, Italy. The WABI workshop was part of a three-conference meeting, which, in addition to WABI, included the ESA and APPROX 2002. The three conferences are jointly called ALGO 2002, and were hosted by the Faculty of Engineering, University of Rome "La Sapienza". See `http://www.dis.uniroma1.it/~algo02` for more details.

The Workshop on Algorithms in Bioinformatics covers research in all areas of algorithmic work in bioinformatics and computational biology. The emphasis is on discrete algorithms that address important problems in molecular biology, genomics, and genetics, that are founded on sound models, that are computationally efficient, and that have been implemented and tested in simulations and on real datasets. The goal is to present recent research results, including significant work in progress, and to identify and explore directions of future research.

Original research papers (including significant work in progress) or state-of-the-art surveys were solicited on all aspects of algorithms in bioinformatics, including, but not limited to: exact and approximate algorithms for genomics, genetics, sequence analysis, gene and signal recognition, alignment, molecular evolution, phylogenetics, structure determination or prediction, gene expression and gene networks, proteomics, functional genomics, and drug design.

We received 83 submissions in response to our call for papers, and were able to accept about half of the submissions. In addition, WABI hosted two invited, distinguished lectures, given to the entire ALGO 2002 conference, by Dr. Ehud Shapiro of the Weizmann Institute and Dr. Gene Myers of Celera Genomics. An abstract of Dr. Shapiro's lecture, and a full paper detailing Dr. Myers lecture, are included in these proceedings.

We would like to sincerely thank all the authors of submitted papers, and the participants of the workshop. We also thank the program committee for their hard work in reviewing and selecting the papers for the workshop. We were fortunate to have on the program committee the following distinguished group of researchers:

Pankaj Agarwal (GlaxoSmithKline Pharmaceuticals, King of Prussia)
Alberto Apostolico (Università di Padova and Purdue University, Lafayette)
Craig Benham (University of California, Davis)
Jean-Michel Claverie (CNRS-AVENTIS, Marseille)
Nir Friedman (Hebrew University, Jerusalem)
Olivier Gascuel (Université de Montpellier II and CNRS, Montpellier)
Misha Gelfand (IntegratedGenomics, Moscow)
Raffaele Giancarlo (Università di Palermo)

David Gilbert (University of Glasgow)
Roderic Guigo (Institut Municipal d'Investigacions Mèdiques, Barcelona, co-chair)
Dan Gusfield (University of California, Davis, co-chair)
Jotun Hein (University of Oxford)
Inge Jonassen (Universitetet i Bergen)
Giuseppe Lancia (Università di Padova)
Bernard M.E. Moret (University of New Mexico, Albuquerque)
Gene Myers (Celera Genomics, Rockville)
Christos Ouzonis (European Bioinformatics Institute, Hinxton Hall)
Lior Pachter (University of California, Berkeley)
Knut Reinert (Celera Genomics, Rockville)
Marie-France Sagot (Université Claude Bernard, Lyon)
David Sankoff (Université de Montréal)
Steve Skiena (State University of New York, Stony Brook)
Gary Stormo (Washington University, St. Louis)
Jens Stoye (Universität Bielefeld)
Martin Tompa (University of Washington, Seattle)
Alfonso Valencia (Centro Nacional de Biotecnología, Madrid)
Martin Vingron (Max-Planck-Institut für Molekulare Genetik, Berlin)
Lusheng Wang (City University of Hong Kong)
Tandy Warnow (University of Texas, Austin)

We also would like to thank the WABI steering committee, Olivier Gascuel, Jotun Hein, Raffaele Giancarlo, Erik Meineche-Schmidt, and Bernard Moret, for inviting us to co-chair this program committee, and for their help in carrying out that task.

We are particularly indebted to Terri Knight of the University of California, Davis, Robert Castelo of the Universitat Pompeu Fabra, Barcelona, and Bernard Moret of the University of New Mexico, Albuquerque, for the extensive technical and advisory help they gave us. We could not have managed the reviewing process and the preparation of the proceedings without their help and advice.

Thanks again to everyone who helped to make WABI 2002 a success. We hope to see everyone again at WABI 2003.

July, 2002 Roderic Guigó and Dan Gusfield

Table of Contents

Simultaneous Relevant Feature Identification and Classification in High-Dimensional Spaces 1
L.R. Grate (Lawrence Berkeley National Laboratory), C. Bhattacharyya, M.I. Jordan, and I.S. Mian (University of California Berkeley)

Pooled Genomic Indexing (PGI): Mathematical Analysis and Experiment Design 10
M. Csűrös (Université de Montréal), and A. Milosavljevic (Human Genome Sequencing Center)

Practical Algorithms and Fixed-Parameter Tractability for the Single Individual SNP Haplotyping Problem 29
R. Rizzi (Università di Trento), V. Bafna, S. Istrail (Celera Genomics), and G. Lancia (Università di Padova)

Methods for Inferring Block-Wise Ancestral History from Haploid Sequences 44
R. Schwartz, A.G. Clark (Celera Genomics), and S. Istrail (Celera Genomics)

Finding Signal Peptides in Human Protein Sequences Using Recurrent Neural Networks 60
M. Reczko (Synaptic Ltd.), P. Fiziev, E. Staub (metaGen Pharmaceuticals GmbH), and A. Hatzigeorgiou (University of Pennsylvania)

Generating Peptide Candidates from Amino-Acid Sequence Databases for Protein Identification via Mass Spectrometry 68
N. Edwards and R. Lippert (Celera Genomics)

Improved Approximation Algorithms for NMR Spectral Peak Assignment . 82
Z.-Z. Chen (Tokyo Denki University), T. Jiang (University of California, Riverside), G. Lin (University of Alberta), J. Wen (University of California, Riverside), D. Xu, and Y. Xu (Oak Ridge National Laboratory)

Efficient Methods for Inferring Tandem Duplication History 97
L. Zhang (Nat. University of Singapore), B. Ma (University of Western Ontario), and L. Wang (City University of Hong Kong)

Genome Rearrangement Phylogeny Using Weighbor 112
L.-S. Wang (University of Texas at Austin)

Segment Match Refinement and Applications 126
A.L. Halpern (Celera Genomics), D.H. Huson (Tübingen University), and K. Reinert (Celera Genomics)

Extracting Common Motifs under the Levenshtein Measure: Theory and Experimentation 140
E.F. Adebiyi and M. Kaufmann (Universität Tübingen)

Fast Algorithms for Finding Maximum-Density Segments of a Sequence with Applications to Bioinformatics 157
M.H. Goldwasser (Loyola University Chicago), M.-Y. Kao (Northwestern University), and H.-I. Lu (Academia Sinica)

FAUST: An Algorithm for Extracting Functionally Relevant Templates from Protein Structures 172
M. Milik, S. Szalma, and K.A. Olszewski (Accelrys)

Efficient Unbound Docking of Rigid Molecules 185
D. Duhovny, R. Nussinov, and H.J. Wolfson (Tel Aviv University)

A Method of Consolidating and Combining EST and mRNA Alignments to a Genome to Enumerate Supported Splice Variants 201
R. Wheeler (Affymetrix)

A Method to Improve the Performance of Translation Start Site Detection and Its Application for Gene Finding 210
M. Pertea and S.L. Salzberg (The Institute for Genomic Research)

Comparative Methods for Gene Structure Prediction in Homologous Sequences 220
C.N.S. Pedersen and T. Scharling (University of Aarhus)

MultiProt – A Multiple Protein Structural Alignment Algorithm 235
M. Shatsky, R. Nussinov, and H.J. Wolfson (Tel Aviv University)

A Hybrid Scoring Function for Protein Multiple Alignment 251
E. Rocke (University of Washington)

Functional Consequences in Metabolic Pathways from Phylogenetic Profiles 263
Y. Bilu and M. Linial (Hebrew University)

Finding Founder Sequences from a Set of Recombinants 277
E. Ukkonen (University of Helsinki)

Estimating the Deviation from a Molecular Clock 287
L. Nakhleh, U. Roshan (University of Texas at Austin), L. Vawter (Aventis Pharmaceuticals), and T. Warnow (University of Texas at Austin)

Exploring the Set of All Minimal Sequences of Reversals – An Application to Test the Replication-Directed Reversal Hypothesis 300
Y. Ajana, J.-F. Lefebvre (Université de Montréal), E.R.M. Tillier (University Health Network), and N. El-Mabrouk (Université de Montréal)

Approximating the Expected Number of Inversions Given the Number of Breakpoints 316
N. Eriksen (Royal Institute of Technology)

Invited Lecture – Accelerating Smith-Waterman Searches 331
G. Myers (Celera Genomics) and R. Durbin (Sanger Centre)

Sequence-Length Requirements for Phylogenetic Methods 343
B.M.E. Moret (University of New Mexico), U. Roshan, and T. Warnow (University of Texas at Austin)

Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum-Evolution Principle 357
R. Desper (National Library of Medicine, NIH) and O. Gascuel (LIRMM)

NeighborNet: An Agglomerative Method for the Construction of Planar Phylogenetic Networks 375
D. Bryant (McGill University) and V. Moulton (Uppsala University)

On the Control of Hybridization Noise in DNA Sequencing-by-Hybridization 392
H.-W. Leong (National University of Singapore), F.P. Preparata (Brown University), W.-K. Sung, and H. Willy (National University of Singapore)

Restricting SBH Ambiguity via Restriction Enzymes 404
S. Skiena (SUNY Stony Brook) and S. Snir (Technion)

Invited Lecture – Molecule as Computation: Towards an Abstraction of Biomolecular Systems 418
E. Shapiro (Weizmann Institute)

Fast Optimal Genome Tiling with Applications to Microarray Design and Homology Search 419
P. Berman (Pennsylvania State University), P. Bertone (Yale University), B. DasGupta (University of Illinois at Chicago), M. Gerstein (Yale University), M.-Y. Kao (Northwestern University), and M. Snyder (Yale University)

Rapid Large-Scale Oligonucleotide Selection for Microarrays 434
S. Rahmann (Max-Planck-Institute for Molecular Genetics)

Border Length Minimization in DNA Array Design 435
A.B. Kahng, I.I. Măndoiu, P.A. Pevzner, S. Reda (University of California at San Diego), and A.Z. Zelikovsky (Georgia State University)

The Enhanced Suffix Array and Its Applications to Genome Analysis 449
M.I. Abouelhoda, S. Kurtz, and E. Ohlebusch (University of Bielefeld)

The Algorithmic of Gene Teams 464
A. Bergeron (Université du Québec a Montreal), S. Corteel (CNRS - Université de Versailles), and M. Raffinot (CNRS - Laboratoire Génome et Informatique)

Combinatorial Use of Short Probes
for Differential Gene Expression Profiling 477
L.L. Warren and B.H. Liu (North Carolina State University)

Designing Specific Oligonucleotide Probes
for the Entire *S. cerevisiae* Transcriptome 491
D. Lipson (Technion), P. Webb, and Z. Yakhini (Agilent Laboratories)

K-ary Clustering with Optimal Leaf Ordering for Gene Expression Data . . 506
Z. Bar-Joseph, E.D. Demaine, D.K. Gifford (MIT LCS), A.M. Hamel (Wilfrid Laurier University), T.S. Jaakkola (MIT AI Lab), and N. Srebro (MIT LCS)

Inversion Medians Outperform Breakpoint Medians
in Phylogeny Reconstruction from Gene-Order Data 521
B.M.E. Moret (University of New Mexico), A.C. Siepel (University of California at Santa Cruz), J. Tang, and T. Liu (University of New Mexico)

Modified Mincut Supertrees 537
R.D.M. Page (University of Glasgow)

Author Index 553