

# Lecture Notes in Bioinformatics

3692

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

Rita Casadio Gene Myers (Eds.)

# Algorithms in Bioinformatics

5th International Workshop, WABI 2005  
Mallorca, Spain, October 3-6, 2005  
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

Rita Casadio  
University of Bologna, Department of Biology/CIRB  
Via Irnerio 42, 40126 Bologna, Italy  
E-mail: casadio@alma.unibo.it

Gene Myers  
Howard Hughes Medical Institute  
4000 Jones Bridge Road, Chavy Chase, MD 20815-6789, USA  
E-mail: gene@eecs.berkeley.edu

Library of Congress Control Number: 2005932938

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

ISSN 0302-9743  
ISBN-10 3-540-29008-7 Springer Berlin Heidelberg New York  
ISBN-13 978-3-540-29008-7 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

[springeronline.com](http://springeronline.com)

© Springer-Verlag Berlin Heidelberg 2005  
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India  
Printed on acid-free paper SPIN: 11557067 06/3142 5 4 3 2 1 0

# Preface

We are pleased to present the proceedings of the 5th Workshop on Algorithms in Bioinformatics (WABI 2005) which took place in Mallorca, Spain, October 3–6, 2005. The WABI 2005 workshop was part of the five ALGO 2005 conference meetings, which, in addition to WABI, included ESA, WAOA, IWPEC, and ATMOS. WABI 2005 was sponsored by EATCS (the European Association for Theoretical Computer Science), the ISCB (the International Society for Computational Biology), the Universitat Politècnica de Catalunya, the Universitat de les Illes Balears, and the Ministerio de Educación y Ciencia. See <http://www.lsi.upc.edu/~wabi05/> for more details.

The Workshop on Algorithms in Bioinformatics highlights research work specifically developed to address algorithmic problems in biosequence analysis. The emphasis is therefore on statistical and probabilistic algorithms that address important problems in the field of molecular and structural biology. At present, given the enormous scientific and technical efforts in functional and structural genomics, the relevance of the problem is therefore constrained by the need for sound, efficient and specialized algorithms, capable of achieving solutions that can be tested by the biological community. Indeed the ultimate goal is to implement algorithms capable of extracting real features from real biological data sets. Therefore the workshop aims 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 94 submissions in response to our call for papers, and were able to accept 35 of these. In addition, WABI 2005 hosted a distinguished lecture by Dr. Marino Zerial of the Max Planck Institute for Molecular Cell Biology and Genetics in Dresden, given to the entire ALGO 2005 conference.

We would like to sincerely thank all the authors of submitted papers, and the participants of the workshop. We also thank the Program Committee and their sub-referees for their hard work in reviewing and selecting the papers for the workshop. The Program Committee consisted of the following 40 distinguished researchers:

Pankaj Kumar Agarwal (Duke University)  
Tatsuya Akutsu (Kyoto University)  
Amir Amihood (Bar-Ilan University)

Alberto Apostolico (Purdue University)  
Craig Benham (University of California, Davis)  
Gary Benson (MSSN, New York)  
Mathieu Blanchette (McGill University)  
Nadia El-Mabrouk (University of Montreal)  
Olivier Gascuel (LIRMM, Montpellier)  
Raffaele Giancarlo (University of Palermo)  
Roderic Guigo (IMIM, Barcelona)  
Michael Hallet (McGill University)  
Daniel Huson (University of Tuebingen)  
Gregory Kucherov (INRIA Nancy)  
Michelle Lacey (Tulane University)  
Jens Lagergren (KTH Stockholm)  
Giuseppe Lancia (Univeristy of Udine)  
Gad M. Landau (University of Haifa)  
Thierry Lecroq (Université de Rouen)  
Bernard Moret (University of New Mexico)  
Shinichi Morishita (University of Tokyo)  
Elchanan Mossel (Univeristy of California, Berkeley)  
Vincent Moulton (University of Uppsala)  
Lior Pachter (University of California, Berkeley)  
Knut Reinert (Free University of Berlin)  
Isidore Rigoutsos (IBM Watson)  
Marie-France Sagot (INRIA Rhône-Alpes)  
David Sankoff (University of Ottawa)  
Sophie Schbath (INRIA Jouv-en-Josas)  
Eran Segal (Rockefeller University)  
Charles Semple (University of Canterbury)  
Joao Carlos Setubal (Virginia Polytechnic Institute)  
Roded Sharan (Tel Aviv Univeristy)  
Steven Skiena (University of New York, Stony Brook)  
Jens Stoye (University of Bielefeld)  
Esko Ukkonen (University of Helsinki)  
Lisa Vawter (Aventis Inc., USA)  
Alfonso Valencia (CNB-CSIC, Spain)  
Tandy Warnow (University of Texas)  
Lusheng Wang (City Univeristy of Hong Kong)

Finally we would like to especially thank Bernard Moret, the de facto steering committee, for answering questions on history and precedence, for his advice on difficult protocol issues, and for setting up and hosting the EasyChair refereeing system used by the Program Committee.

# Table of Contents

## Expression

### 1. Hybrid Methods

Spectral Clustering Gene Ontology Terms to Group Genes by Function <i>Nora Speer, Christian Spieth, Andreas Zell</i> .....	1
Dynamic De-Novo Prediction of microRNAs Associated with Cell Conditions: A Search Pruned by Expression <i>Chaya Ben-Zaken Zilberstein, Michal Ziv-Ukelson</i> .....	13

### 2. Time Patterns

Clustering Gene Expression Series with Prior Knowledge <i>Laurent Bréhélin</i> .....	27
A Linear Time Biclustering Algorithm for Time Series Gene Expression Data <i>Sara C. Madeira, Arlindo L. Oliveira</i> .....	39
Time-Window Analysis of Developmental Gene Expression Data with Multiple Genetic Backgrounds <i>Tamir Tuller, Efrat Oron, Erez Makavy, Daniel A. Chamovitz, Benny Chor</i> .....	53

## Phylogeny

### 1. Quartets

A Lookahead Branch-and-Bound Algorithm for the Maximum Quartet Consistency Problem <i>Gang Wu, Jia-Huai You, Guohui Lin</i> .....	65
Computing the Quartet Distance Between Trees of Arbitrary Degree <i>Chris Christiansen, Thomas Mailund, Christian N.S. Pedersen, Martin Randers</i> .....	77

## 2. Tree Reconciliation

Using Semi-definite Programming to Enhance Supertree Resolvability <i>Shlomo Moran, Satish Rao, Sagi Snir</i> .....	89
An Efficient Reduction from Constrained to Unconstrained Maximum Agreement Subtree <i>Z.S. Peng, H.F. Ting</i> .....	104

## 3. Clades and Haplotypes

Pattern Identification in Biogeography <i>Ganeshkumar Ganapathy, Barbara Goodson, Robert Jansen, Vijaya Ramachandran, Tandy Warnow</i> .....	116
On the Complexity of Several Haplotyping Problems <i>Rudi Cilibrasi, Leo van Iersel, Steven Kelk, John Tromp</i> .....	128
A Hidden Markov Technique for Haplotype Reconstruction <i>Pasi Rastas, Mikko Koivisto, Heikki Mannila, Esko Ukkonen</i> .....	140
Algorithms for Imperfect Phylogeny Haplotyping (IPPH) with a Single Homoplasy or Recombination Event <i>Yun S. Song, Yufeng Wu, Dan Gusfield</i> .....	152

## Networks

A Faster Algorithm for Detecting Network Motifs <i>Sebastian Wernicke</i> .....	165
Reaction Motifs in Metabolic Networks <i>Vincent Lacroix, Cristina G. Fernandes, Marie-France Sagot</i> .....	178
Reconstructing Metabolic Networks Using Interval Analysis <i>Warwick Tucker, Vincent Moulton</i> .....	192

## Genome Rearrangements

### 1. Trasposition Model

A 1.375-Approximation Algorithm for Sorting by Transpositions <i>Isaac Elias, Tzvika Hartman</i> .....	204
---	-----

A New Tight Upper Bound on the Transposition Distance <i>Anthony Labarre</i> .....	216
---	-----

## 2. Other Models

Perfect Sorting by Reversals Is Not Always Difficult <i>S�everine B�erard, Anne Bergeron, Cedric Chauve, Christophe Paul</i> .....	228
Minimum Recombination Histories by Branch and Bound <i>Rune B. Lyngs�, Yun S. Song, Jotun Hein</i> .....	239

## Sequences

### 1. Strings

A Unifying Framework for Seed Sensitivity and Its Application to Subset Seeds <i>Gregory Kucherov, Laurent No�, Mikhail Roytberg</i> .....	251
Generalized Planted $(l,d)$ -Motif Problem with Negative Set <i>Henry C.M. Leung, Francis Y.L. Chin</i> .....	264
Alignment of Tandem Repeats with Excision, Duplication, Substitution and Indels (EDSI) <i>Michael Sammeth, Thomas Weniger, Dag Harmsen, Jens Stoye</i> .....	276
The Peres-Shields Order Estimator for Fixed and Variable Length Markov Models with Applications to DNA Sequence Similarity <i>Daniel Dalevi, Devdatt Dubhashi</i> .....	291

### 2. Multi-alignment and Clustering

Multiple Structural RNA Alignment with Lagrangian Relaxation <i>Markus Bauer, Gunnar W. Klau, Knut Reinert</i> .....	303
Faster Algorithms for Optimal Multiple Sequence Alignment Based on Pairwise Comparisons <i>Pankaj K. Agarwal, Yonatan Bilu, Rachel Kolodny</i> .....	315
Ortholog Clustering on a Multipartite Graph <i>Akshay Vashist, Casimir Kulikowski, Ilya Muchnik</i> .....	328



### 3. Clustering and Representation

Linear Time Algorithm for Parsing RNA Secondary Structure  
*Baharak Rastegari, Anne Condon* ..... 341

A Compressed Format for Collections of Phylogenetic Trees and  
 Improved Consensus Performance  
*Robert S. Boyer, Warren A. Hunt Jr, Serita M. Nelesen* ..... 353

## Structure

### 1. Threading

Optimal Protein Threading by Cost-Splitting  
*Philippe Veber, Nicola Yanev, Rumen Andonov, Vincent Poirriez* .... 365

Efficient Parameterized Algorithm for Biopolymer Structure-Sequence  
 Alignment  
*Yinglei Song, Chunmei Liu, Xiuzhen Huang, Russell L. Malmberg,  
 Ying Xu, Liming Cai* ..... 376

### 2. Folding

Rotamer-Pair Energy Calculations Using a Trie Data Structure  
*Andrew Leaver-Fay, Brian Kuhlman, Jack Snoeyink* ..... 389

Improved Maintenance of Molecular Surfaces Using Dynamic Graph  
 Connectivity  
*Eran Eyal, Dan Halperin* ..... 401

The Main Structural Regularities of the Sandwich Proteins  
*Alexander Kister* ..... 414

Discovery of Protein Substructures in EM Maps  
*Keren Lasker, Oranit Dror, Ruth Nussinov, Haim Wolfson* ..... 423

**Author Index** ..... 435