



DIMACS

Series in Discrete Mathematics
and Theoretical Computer Science

Volume 37

Mathematical Hierarchies and Biology

DIMACS Workshop
November 13–15, 1996

Boris Mirkin
F. R. McMorris
Fred S. Roberts
Andrey Rzhetsky
Editors



American Mathematical Society

Selected Titles in This Series

- 37 **Boris Mirkin, F. R. McMorris, Fred S. Roberts, and Andrey Rzhetsky, Editors,** Mathematical Hierarchies and Biology
- 36 **Joseph G. Rosenstein, Deborah S. Franzblau, and Fred S. Roberts, Editors,** Discrete Mathematics in the Schools
- 35 **Dingzhu Du, Jun Gu, and Panos M. Pardalos, Editors,** Satisfiability Problem: Theory and Applications
- 34 **Nathaniel Dean, Editor,** African Americans in Mathematics
- 33 **Ravi B. Boppana and James F. Lynch, Editors,** Logic and random structures
- 32 **Jean-Charles Grégoire, Gerard J. Holzmann, and Doron A. Peled, Editors,** The SPIN verification system
- 31 **Neil Immerman and Phokion G. Kolaitis, Editors,** Descriptive complexity and finite models
- 30 **Sandeep N. Bhatt, Editor,** Parallel Algorithms: Third DIMACS Implementation Challenge
- 29 **Doron A. Peled, Vaughan R. Pratt, and Gerard J. Holzmann, Editors,** Partial Order Methods in Verification
- 28 **Larry Finkelstein and William M. Kantor, Editors,** Groups and Computation II
- 27 **Richard J. Lipton and Eric B. Baum, Editors,** DNA Based Computers
- 26 **David S. Johnson and Michael A. Trick, Editors,** Cliques, Coloring, and Satisfiability: Second DIMACS Implementation Challenge
- 25 **Gilbert Baumslag, David Epstein, Robert Gilman, Hamish Short, and Charles Sims, Editors,** Geometric and Computational Perspectives on Infinite Groups
- 24 **Louis J. Billera, Curtis Greene, Rodica Simion, and Richard P. Stanley, Editors,** Formal Power Series and Algebraic Combinatorics/Séries formelles et combinatoire algébrique, 1994
- 23 **Panos M. Pardalos, David I. Shalloway, and Guoliang Xue, Editors,** Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding
- 22 **Panos M. Pardalos, Mauricio G. C. Resende, and K. G. Ramakrishnan, Editors,** Parallel Processing of Discrete Optimization Problems
- 21 **D. Frank Hsu, Arnold L. Rosenberg, and Dominique Sotteau, Editors,** Interconnection Networks and Mapping and Scheduling Parallel Computations
- 20 **William Cook, László Lovász, and Paul Seymour, Editors,** Combinatorial Optimization
- 19 **Ingemar J. Cox, Pierre Hansen, and Bela Julesz, Editors,** Partitioning Data Sets
- 18 **Guy E. Blelloch, K. Mani Chandy, and Suresh Jagannathan, Editors,** Specification of Parallel Algorithms
- 17 **Eric Sven Ristad, Editor,** Language Computations
- 16 **Panos M. Pardalos and Henry Wolkowicz, Editors,** Quadratic Assignment and Related Problems
- 15 **Nathaniel Dean and Gregory E. Shannon, Editors,** Computational Support for Discrete Mathematics
- 14 **Robert Calderbank, G. David Forney, Jr., and Nader Moayeri, Editors,** Coding and Quantization: DIMACS/IEEE Workshop
- 13 **Jin-Yi Cai, Editor,** Advances in Computational Complexity Theory
- 12 **David S. Johnson and Catherine C. McGeoch, Editors,** Network Flows and Matching: First DIMACS Implementation Challenge
- 11 **Larry Finkelstein and William M. Kantor, Editors,** Groups and Computation
- 10 **Joel Friedman, Editor,** Expanding Graphs
- 9 **William T. Trotter, Editor,** Planar Graphs
- 8 **Simon Gindikin, Editor,** Mathematical Methods of Analysis of Biopolymer Sequences
- 7 **Lyle A. McGeoch and Daniel D. Sleator, Editors,** On-Line Algorithms

(Continued in the back of this publication)

This page intentionally left blank

DIMACS

Series in Discrete Mathematics
and Theoretical Computer Science

Volume 37

Mathematical Hierarchies and Biology

DIMACS Workshop
November 13–15, 1996

Boris Mirkin
F. R. McMorris
Fred S. Roberts
Andrey Rzhetsky
Editors

NSF Science and Technology Center
in Discrete Mathematics and Theoretical Computer Science
A consortium of Rutgers University, Princeton University,
AT&T Labs, Bell Labs, and Bellcore



American Mathematical Society

This DIMACS volume contains a set of refereed papers written mostly by participants in the workshop "Mathematical Hierarchies and Biology" held at DIMACS on November 13–15, 1996. The papers provide a contemporary sample of many new results in hierarchy theory with applications in biology, psychology, data analysis, and systems engineering and should be of use to researchers in discrete mathematics, computational biology, and many other areas.

1991 *Mathematics Subject Classification*. Primary 92Bxx; Secondary 05Cxx.

Library of Congress Cataloging-in-Publication Data

Mathematical hierarchies and biology : DIMACS workshop, November 13–15, 1996 / Boris Mirkin... [et al.], editors.

p. cm. — (DIMACS series in discrete mathematics and theoretical computer science, ISSN 1052-1798 ; v. 37)

"NSF Science and Technology Center in Discrete Mathematics and Theoretical Computer Science, a consortium of Rutgers University, Princeton University, AT&T Labs, Bell Labs, and Bellcore."

Includes bibliographical references and index.

ISBN 0-8218-0762-5 (hc : alk. paper)

1. Graph theory—Congresses. 2. Hierarchies—Congresses. 3. Biomathematics—Congresses. I. Mirkin, B. G. (Boris Grigor'evich) II. DIMACS (Group) III. NSF Science and Technology Center in Discrete Mathematics and Theoretical Computer Science. IV. Series.

QA166.M365 1997

511'.5—dc21

97-26706

CIP

Copying and reprinting. Material in this book may be reproduced by any means for educational and scientific purposes without fee or permission with the exception of reproduction by services that collect fees for delivery of documents and provided that the customary acknowledgment of the source is given. This consent does not extend to other kinds of copying for general distribution, for advertising or promotional purposes, or for resale. Requests for permission for commercial use of material should be addressed to the Assistant to the Publisher, American Mathematical Society, P. O. Box 6248, Providence, Rhode Island 02940-6248. Requests can also be made by e-mail to reprint-permission@ams.org.

Excluded from these provisions is material in articles for which the author holds copyright. In such cases, requests for permission to use or reprint should be addressed directly to the author(s). (Copyright ownership is indicated in the notice in the lower right-hand corner of the first page of each article.)

© 1997 by the American Mathematical Society. All rights reserved.

The American Mathematical Society retains all rights
except those granted to the United States Government.

Printed in the United States of America.

⊗ The paper used in this book is acid-free and falls within the guidelines
established to ensure permanence and durability.

Visit the AMS home page at URL: <http://www.ams.org/>

10 9 8 7 6 5 4 3 2 1 02 01 00 99 98 97

Contents

Foreword	ix
Preface	xi
Ancestor-Descendant Relations and Incompatible Data: Motivation for Research in Discrete Mathematics G.F. Estabrook	1
Krohn-Rhodes Theory, Hierarchies & Evolution C.L. Nehaniv and J.L. Rhodes	29
Inferring Evolutionary Trees from Polymorphic Characters, and an Analysis of the Indo-European Family of Languages M. Bonet, C. Phillips, T. Warnow, and S. Yooseph	43
Reconciled Trees and Incongruent Gene and Species Trees R.D.M. Page and M.A. Charleston	57
Comparison of Annotating Duplication, Tree Mapping, and Copying as Methods to Compare Gene Trees with Species Trees O. Eulenstein, B. Mirkin, and M. Vingron	71
Fitting Models of Intron Evolution to Aldehyde Dehydrogenase Data A. Rzhetsky, F.J. Ayala, L.C. Hsu, C. Chang, and A. Yoshida	95
Dissimilarity Maps and Substitution Models: Some New Results V. Moulton, M. Steel, and C. Tuffley	111
The Performance of the NJ Method of Phylogeny Reconstruction K. Atteson	133
Concerning the NJ Algorithm and Its Unweighted Version, UNJ O. Gascuel	149
Order Distances in Tree Reconstruction A. Guénoche	171
Circular Orders of Tree Metrics, and Their Uses for the Reconstruction and Fitting of Phylogenetic Trees V. Makarenkov and B. Leclerc	183

Estimation of Missing Distances in Path-Length Matrices: Problems and Solutions	209
P.-A. Landry and F.-J. Lapointe	
Complexity Issues in Hierarchical Optimization	219
P.M. Pardalos and X. Deng	
Nesticity	225
P. Hansen and D. de Werra	
Phylogeny Graphs of Arbitrary Digraphs	233
F. S. Roberts and L. Sheng	
Agreement Metrics for Trees Revisited	239
E. Kubicka, G. Kubicki, and F.R. McMorris	
Sparse Dynamic Programming for Maximum Agreement Subtree Problem	249
T.M. Przytycka	
The Median Function on Weak Hierarchies	265
F.R. McMorris and R.C. Powers	
Towards a Theory of Holistic Clustering	271
A.W.M. Dress	
On Hierarchies and Hierarchical Classes Models	291
I. Van Mechelen, S. Rosenberg, and P. De Boeck	
The Construction of Globally Optimal Ordered Partitions	299
L. Hubert, P. Arabie, and J. Meulman	
Multiple Trees: Fitting Two or More Tree Structures to Proximity Data	313
J.D. Carroll and G. De Soete	
Linear Embedding of Binary Hierarchies and Its Applications	331
B. Mirkin	
Learning Algorithms Generating Multigranular Hierarchies	357
A. Meystel	
Index	385

Foreword

The Workshop on “Mathematical Hierarchies and Biology”, held in November 1996, was part of DIMACS Special Focus on Mathematical Support for Molecular Biology that began in 1994. We would like to express our appreciation to Boris Mirkin as chair of the committee and to Buck McMorris, Fred Roberts, and Andrey Rzhetsky assisting him for their efforts to organize and plan this successful workshop as well as for editing this volume of papers.

The special year encouraged collaborations among very different research communities, and this volume records one of many workshops in which this was achieved. We also extend our thanks to Fred Roberts and Joachim Messing as chairs of the special year committee, Lawrence Shepp and Michael Waternam as co-chairs, Martin Farach as assistant chair, and Sampath Kannan as publicity chair for their efforts to coordinate the broad range of activities.

DIMACS gratefully acknowledges the generous support that makes these programs possible. The National Science Foundation, through its Science and Technology Center program, the New Jersey Commission on Science and Technology, DIMACS partners at Rutgers, Princeton, AT&T Labs, Bell Labs, and Bellcore generously supported the special year, and we thank NSF for a special supplementary award that helped to support the workshop that led to this volume. Additional support for the special year was obtained from BIOSYM Tech, the Centers for Disease Control, Ciba-Geigy, Merck Research, the National Center for Human Genome Research, the National Institute for Allergy and Infectious Diseases, the National Security Agency, Roche Molecular Systems, Sandia, and SmithKline Beecham.

Fred S. Roberts
Director

Bernard Chazelle
Co-Director for Princeton

Stephen R. Mahaney
Associate Director for Research

This page intentionally left blank

Preface

Viewing hierarchies as combinatorial objects underlies many important areas of current scientific investigation. Examples of areas where this is the case are biology, evolutionary studies and taxonomy, general classification theory, data/knowledge bases, cognitive models and linguistical structures, industrial control, etc. The mathematical approach to the study of hierarchies represents the theoretical underpinning of many important areas of current scientific investigation. Biology has benefited from this research and has also stimulated the mathematical study of hierarchies.

This collection presents a set of refereed papers written mostly by the participants in the Workshop “Mathematical Hierarchies and Biology” held at DIMACS 13-15 November 1996. The papers and some of the results described implement many of the discussions and suggestions for revision made by the other participants and editors. Of two extreme formats of a scientific publication, monographs and journals, the current edition has taken some features of both, combining review papers and papers presenting the latest results.

The papers can roughly be organized into the following four areas:

1. Combinatorial modeling of the evolutionary processes (G. Estabrook; C. Nehaniv and J. Rhodes; M. Bonet, C. Phillips, T. Warnow, and S. Yooseph; R. Page and M. Charleston; O. Eulenstein, B. Mirkin, and M. Vingron; A. Rzhetsky, F. Ayala, L. Hsu, C. Chang, and A. Yoshida; M. Steel, V. Moulton, and C. Tuffley)
2. Reconstructing trees from dissimilarity data (K. Atteson; O. Gascuel; A. Guénoche; B. Leclerc and V. Makarenkov; F.-J. Lapointe and P.-A. Landry; the paper by M. Steel et al in item 1 can be put here, too)
3. Related mathematical issues (P. Pardalos and X. Deng; P. Hansen and D. Werra; F. Roberts and L. Sheng; E. Kubicka, G. Kubicki, and F. McMorris; T. Przytycka; F. McMorris and R. Powers)
4. Clustering and data analysis (A. Dress; I. Van Mechelen, S. Rosenberg, and P. De Boeck; L. Hubert, P. Arabie, and J. Meulman; D. Carrol and G. De Soete; B. Mirkin; A. Meystel)

The reader should be aware that there are many other aspects of mathematical hierarchies covered and many interconnections among the papers beyond the grouping according to the four classes above. The subjects can be identified using the contents and index to the volume. The interconnections can be found based on mathematical problems analyzed, formalisms used, and on the substantive problems involved.

The papers in the volume provide a contemporary sample of many new results in hierarchy theory with applications in biology, psychology, data analysis and

systems engineering, and should be of use to researchers in discrete mathematics, computational biology and many other areas.

B. Mirkin
F.R. McMorris
F.S. Roberts
A. Rzhetsky

Selected Titles in This Series

(Continued from the front of this publication)

- 6 **Jacob E. Goodman, Richard Pollack, and William Steiger, Editors**, Discrete and Computational Geometry: Papers from the DIMACS Special Year
- 5 **Frank Hwang, Fred Roberts, and Clyde Monma, Editors**, Reliability of Computer and Communication Networks
- 4 **Peter Gritzmann and Bernd Sturmfels, Editors**, Applied Geometry and Discrete Mathematics, The Victor Klee Festschrift
- 3 **E. M. Clarke and R. P. Kurshan, Editors**, Computer-Aided Verification '90
- 2 **Joan Feigenbaum and Michael Merritt, Editors**, Distributed Computing and Cryptography
- 1 **William Cook and Paul D. Seymour, Editors**, Polyhedral Combinatorics

ISBN 0-8218-0762-5



9 780821 807620