Lecture Notes in Computer Science 1278
Edited by G. Goos, J. Hartmanis and J. van Leeuwen

Advisory Board: W. Brauer D. Gries J. Stoer

Ralf Hofestädt Thomas Lengauer
Markus Löffler Dietmar Schomburg (Eds.)

Bioinformatics

German Conference
on Bioinformatics, GCB'96
Leipzig, Germany
September 30 — October 2, 1996
Selected Papers

Springer

Series Editors

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

Volume Editors

Ralf Hofestädt
Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik
Universitätsplatz 2, D-39106 Magdeburg, Germany
E-mail: hofestaedt@iti.cs.uni-magdeburg.de

Thomas Lengauer
GMD-I1, Schloß Birlinghoven, Institut für Methodische Grundlagen
D-53732 Sankt Augustin, Germany
E-mail: lengauer@gmd.de

Markus Löffler
Universität Leipzig, IMISE
Liebigstr. 27, D-04103 Leipzig, Germany
E-mail: loeffler@imise.uni-leipzig.de

Dietmar Schomburg
Universität Köln, Institut für Biochemie
Zülpicher Str. 47, D-50677 Köln, Germany
E-mail: schomburg@uni-koeln.de

Cataloging-in-Publication data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Bioinformatics : selected papers / German Conference on Bioinformatics, GCB '96, Leipzig, Germany, September 30 - October 2, 1996. Ralf Hofestädt ... (ed.). - Berlin ; Heidelberg ; New York ; Barcelona ; Budapest ; Hong Kong ; London ; Milan ; Paris ; Santa Clara ; Singapore ; Tokyo : Springer, 1997
(Lecture notes in computer science ; Vol. 1278)
ISBN 3-540-63370-7

CR Subject Classification (1991): F1-2, F.4.3, I.6, I.5, J.3

ISSN 0302-9743
ISBN 3-540-63370-7 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 1997
Printed in Germany

Typesetting: Camera-ready by author
SPIN 10547826 06/3142 – 5 4 3 2 1 0 Printed on acid-free paper

Preface

Methods and concepts from computer science are gaining increasing importance in the area of biology, especially molecular biology. Methods for storing, accessing, and manipulating biological data – implemented in software and sometimes even in specialized hardware – are essential for such diverse problems as analyzing evolutionary processes, modeling complex molecular structures, simulating aspects of biological processes, and developing drug agents that control critical aspects of diseases.

The main task of this interdisciplinary research field is to develop software tools for the analysis of biological sequences, structures, and systems. The application of methods and concepts from computer science is needed because of the high complexity of the systems to be analyzed and the overwhelming amount of data at hand.

The main source of data comes from the various genome sequencing projects, the most ambitious of which plans to uncover all of the roughly $3 * 10^9$ base pairs in the human genome by the year 2005. Merely obtaining this text is inconceivable without computer science methods to help in cleaning up and assembling the data. The main challenge, however, will be to begin to understand what this text means. Toward this end, we need to identify the proteins manufactured by the organism – their structure as well as their function – and the complex mechanisms of regulation and metabolism that enable the organism to survive. Even today, we have the first eukaryotic genome available, that of yeast (S. cerevisiae) with roughly 6000 genes. Therefore, we are not talking about tasks of the future but demands of here and now.

Computer science can help in various ways to interpret the biological data. Statistics and methods of artificial intelligence, notably neural networks and genetic algorithms, help in structuring and classifying data in the presence of high noise levels. Optimization methods are tools used in the analysis of molecular sequences and structures. Methods of data handling help to navigate through diverse and inhomogeneous data sets and to maintain levels of data consistency. Visualization and computer animation are useful for demonstrating complex relationships, structures, or processes in two or three dimensions.

In Germany connections between universities, research laboratories, and industry have been set up in the past few years in order to work on these interdisciplinary problems. The efforts have been supported with strategic action by the German Federal Ministry of Education, Science, Research, and Technology (BMBF) and, more recently, by the German National Science Foundation (DFG).

In 1992, the German Society of Computer Science (GI) founded a special interest group on "Informatics in the Biological Sciences" (GI-FG 4.0.2). The overall goal of this group is to form a bridge between computer science and biology. Its concrete tasks are: (1) to help to introduce computer science methods into research in molecular biology and biotechnology; (2) to develop new foundations, methods, and tools to solve problems in the field of biology; (3) to increase innovative interactions between biology and computer science. The group organizes a number of workshops and conferences. For details see:

http://wwwiti.cs.uni-magdeburg.de/Veranstaltungen.html

The most recent meeting, the International German Conference on Bioinformatics, took place from September 30th to October 2nd, 1996, in Leipzig (Germany). The meeting was organized in cooperation with the German Society for Chemical Apparatus, Chemical Engineering and Biotechnology (DECHEMA) and the German Society for Medical Informatics, Biometry and Epidemiology (GMDS). The members of the Organizing Committee were Ralf Hofestädt (University of Magdeburg), Thomas Lengauer (University of Bonn, GMD St. Augustin), Markus Löffler (University of Leipzig), and Dietmar Schomburg (University of Köln).

The main topics of this conference included:
- Application of Database Systems to the Human Genome Project (HGP)
- Sequence Analysis
- Modeling and Simulation of Gene Regulation
- Molecular Modeling und Molecular Design
- Formal Languages and DNA
- Metabolic Network Control

Based on these topics the international program committee (Julio Collado-Vides, Antoine Danchin, Andreas Dress, Peter Karp, Heinz Kubinyi, Michael Mavrovouniotis, Hans-Werner Mewes, Jude Shavlik,

Sándor Suhai, Martin Vingron, Edgar Wingender, and Hans Zima) selected 22 talks from more than 120 submissions. In addition to these oral presentations the program committee admitted 69 posters and computer demos. Based on these presentations the organizing committee invited 36 submissions to this volume. All papers received were submitted to the usual refereeing process.

We would like to thank all participants of the workshop – 166 from Germany, Europe, Japan, Canada, and USA – for creating such a good working atmosphere, all who supported the organization, and all others who helped to make GCB'96 a success. Especially, we thank the Ministry of Science and Art (Freistaat Sachsen), the Kurt-Eberhard-Bode-Stiftung im Stifterverband für die Deutsche Wissenschaft, and the company Bode Chemie Hamburg for their support.

Magdeburg, June 1997

Ralf Hofestädt
Thomas Lengauer
Markus Löffler
Dietmar Schomburg

Table of Contents

4 Sequence Analysis

5 Molecular Modeling

6 Visualization

7 Formal Languages and DNA