

Thomas Lengauer (Ed.)
Bioinformatics –
From Genomes to Drugs
Vol. I: Basic Technologies
Vol. II: Applications

Methods and Principles in Medicinal Chemistry

Edited by

R. Mannhold

H. Kubinyi

H. Timmerman

Editorial Board

G. Folkers, H.-D. Höltje, J. Vacca,

H. van de Waterbeemd, T. Wieland

**Bioinformatics – From Genomes
to Drugs**

Volume I: Basic Technologies

Volume II: Applications

Edited by Thomas Lengauer

Series Editors

Prof. Dr. Raimund Mannhold

Biomedical Research Center
Molecular Drug Research Group
Heinrich-Heine-Universität
Universitätsstraße 1
D-40225 Düsseldorf
Germany

Prof. Dr. Hugo Kubinyi

BASF AG, Ludwigshafen
c/o Donnersbergstrasse 6
D-67256 Weisenheim am Sand
Germany

Prof. Dr. Gerd Folkers

Department of Applied Biosciences
ETH Zürich
Winterthurer Str. 190
CH-8057 Zürich
Switzerland

Volume Editor:

Prof. Dr. Thomas Lengauer, Ph.D.

Fraunhofer Institute for Algorithms and
Scientific Computing (SCAI)
Schloss Birlinghoven
D-53754 Sankt Augustin
Germany

This book was carefully produced. Nevertheless, editors, authors and publisher do not warrant the information contained therein to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

Library of Congress Card No.: applied for

A catalogue record for this book is available from the British Library. Die Deutsche Bibliothek – CIP Cataloguing-in-Publication-Data A catalogue record for this publication is available from Die Deutsche Bibliothek

© Wiley-VCH Verlag GmbH, Weinheim (Federal Republic of Germany). 2002
All rights reserved (including those of translation in other languages). No part of this book may be reproduced in any form – by photoprinting, microfilm, or any other means – nor transmitted or translated into machine language without written permission from the publishers. Registered names, trademarks, etc. used in this book, even when not specifically marked as such, are not to be considered unprotected by law.

Printed in the Federal Republic of Germany.

Printed on acid-free paper.

Typesetting Asco Typesetters,
Hong Kong

Printing betz-druck GmbH, Darmstadt

Bookbinding J. Schäffer GmbH & Co. KG,
Grünstadt

ISBN 3-527-29988-2

Preface

The present volume of our series “Methods and Principles in Medicinal Chemistry” focuses on a timely topic: Bioinformatics. Bioinformatics is a multidisciplinary field, which encompasses molecular biology, biochemistry and genetics on the one hand, and computer science on the other. Bioinformatics uses methods from various areas of computer science, such as algorithms, combinatorial optimization, integer linear programming, constraint programming, formal language theory, neural nets, machine learning, motif recognition, inductive logic programming, database systems, knowledge discovery and database mining. The exponential growth in biological data, generated from national and international genome projects, offers a remarkable opportunity for the application of modern computer science. The fusion of biomedicine and computer technology offers substantial benefits to all scientists involved in biomedical research in support of their general mission of improving the quality of health by increasing biological knowledge. In this context, we felt that it was time to initiate a volume on bioinformatics with a particular emphasis on aspects of designing new drugs.

The completion of the human genome sequence, published in February 2001, marks a historic event, not only in genomics, but also in biology and medicine in general. We are now able to read the text; but we understand only minor parts of it. “Making sense of the sequence” is the task of the coming years. Bioinformatics will play the leading role in this field, in understanding the regulation of gene expression, in the functional description of the gene products, the metabolic processes, disease, genetic variation and comparative biology. Correspondingly, the publication of this book is “just in time” to jump into the post-genomic era.

Basically, there are two ways of structuring the field of bioinformatics. One is intrinsically by the type of problem that is under consideration. Here, the natural way of structuring is by layers of information that are compiled, starting from the genomic data. The second is extrinsically, by the application scenario in which bioinformatics operates and by the type of molecular biology experiment that it supports. This new contribution to bioinformatics is roughly structured according to this view. The wealth of

information bundled in this volume necessitated a subdivision into two parts.

The intrinsic view is the subject of Part 1: it structures bioinformatics in methodical layers. Lower layers operate directly on the genomic text that is the result of sequencing projects. Higher layers operate on higher-level information derived from this text. Accordingly, Part 1 discusses subproblems of bioinformatics that provide components in a global bioinformatics solution. Each chapter is devoted to one relevant component: after an introductory overview, Chapters follow that are devoted to Sequence Analysis (*written by Martin Vingron*), Structure, Properties and Computer Identification of Eukaryotic Genes (*by Victor Solovyev*), Analyzing Regulatory Regions in Genomes (*by Thomas Werner*), Homology-Based Protein Modeling in Biology and Medicine (*by Roland Dunbrack*), Protein Structure Prediction and Applications in Structural Genomics, Protein Function Assignment and Drug Target Finding (*by Ralf Zimmer and Thomas Lengauer*), Protein–Ligand Docking and Drug Design (*by Matthias Rarey*) and Protein–Protein and Protein–DNA Docking (*by Mike Sternberg and Gidon Moont*). An appendix by *Thomas Lengauer*, sketching the algorithmic methods that are used in bioinformatics, concludes this first Part.

The extrinsic view is the focus of the second Part: Chapters concentrate on several important application scenarios that can only be supported effectively by combining components discussed in Part 1. These Chapters cover Integrating and Accessing Molecular Biology Resources (*by David Hansen and Thure Etzold*), Bioinformatics Support of Genome Sequencing Projects (*by Xiaoqiu Huang*), Analysis of Sequence Variations (*by Christopher Carlson et al.*), Proteome Analysis (*by Pierre-Alan Binz et al.*), Target Finding in Genomes and Proteomes (*by Stefanie Fuhrman et al.*) as well as Screening of Drug Databases (*by Martin Stahl et al.*). In a concluding Chapter, *Thomas Lengauer* highlights the Future Trends in the field of bioinformatics.

The series editors are grateful to Thomas Lengauer, who accepted the challenging task to organize this volume on bioinformatics, to convince authors to participate in the project and to finish their chapters in time, despite the fact that research runs hot these days. We are sure that the result of his coordinative work constitutes another highlight in our series on Methods and Principles in Medicinal Chemistry. In addition, we want to thank Gudrun Walter and Frank Weinreich, Wiley-VCH, for their effective collaboration.

September 2001

Raimund Mannhold
Hugo Kubinyi
Henk Timmerman

Düsseldorf
Ludwigshafen
Amsterdam

Contents

Part I: Basic Technologies

List of Contributors *xvii*

Foreword *xix*

1	From Genomes to Drugs with Bioinformatics	3
1.1	The molecular basis of disease	3
1.2	The molecular approach to curing diseases	8
1.3	Finding protein targets	10
1.3.1	Genomics vs proteomics	12
1.3.2	Extent of information available on the genes/proteins	12
1.4	Developing drugs	14
1.5	A bioinformatics landscape	15
1.5.1	The intrinsic view	16
1.6	The extrinsic view	21
1.6.1	Basic contributions: molecular biology database and genome comparison	22
1.6.2	Scenario 1: Gene and protein expression data	22
1.6.3	Scenario 2: Drug screening	23
1.6.4	Scenario 3: Genetic variability	24
2	Sequence Analysis	27
2.1	Introduction	27
2.2	Analysis of individual sequences	28
2.2.1	Secondary structure prediction	31
2.3	Pairwise sequence comparison	32
2.3.1	Dot plots	33
2.3.2	Sequence alignment	34
2.4	Database searching I: single sequence heuristic algorithms	39
2.5	Alignment and search statistics	42
2.6	Multiple sequence alignment	45

2.7	Multiple alignments and database searching	47
2.8	Protein families and protein domains	49
2.9	Conclusion	50
3	Structure, Properties and Computer Identification of Eukaryotic Genes	59
3.1	Structural characteristics of eukaryotic genes	59
3.2	Classification of splice sites in mammalian genomes	62
3.3	Methods for the recognition of functional signals	66
3.3.1	Search for nonrandom similarity with consensus sequences	66
3.3.2	Position-specific sensors	69
3.3.3	Content-specific measures	71
3.3.4	Frame-specific measures for recognition of protein coding regions	71
3.3.5	Accuracy measures	72
3.3.6	Application of linear discriminant analysis	73
3.3.7	Prediction of donor and acceptor splice junctions	74
3.3.8	Recognition of promoter regions in human DNA	78
3.3.9	Prediction of poly-A sites	81
3.4	Gene identification approaches	84
3.5	Discriminative and probabilistic approaches for multiple gene prediction	85
3.5.1	Multiple gene prediction using HMM approach	85
3.5.2	Pattern based multiple gene prediction approach	88
3.5.3	Accuracy of gene identification programs	93
3.5.4	Using protein or EST similarity information to improve gene prediction	95
3.6	Annotation of sequences from genome sequencing projects	97
3.7	InfoGene: database of known and predicted genes	99
3.7.1	Annotation of human genome draft	101
3.8	Functional analysis and verification of predicted genes	101
3.9	Internet sites for gene finding and functional site prediction	104
4	Analyzing Regulatory Regions in Genomes	113
4.1	General features of regulatory regions in eukaryotic genomes	113
4.2	General functions of regulatory regions	113
4.2.1	Transcription factor binding sites (TF-sites)	114
4.2.2	Sequence features	114
4.2.3	Structural elements	115
4.2.4	Organizational principles of regulatory regions	115
4.2.5	Bioinformatics models for the analysis and detection of regulatory regions	129
4.3	Methods for element detection	122
4.3.1	Detection of transcription factor binding sites	122

4.3.2	Detection of structural elements	123
4.3.3	Assessment of other elements	123
4.4	Analysis of regulatory regions	125
4.4.1	Training set selection	125
4.4.2	Statistical and biological significance	126
4.4.3	Context dependency	126
4.5	Methods for detection of regulatory regions	126
4.5.1	Types of regulatory regions	128
4.5.2	Programs for recognition of regulatory sequences	129
4.6	Annotation of large genomic sequences	136
4.6.1	The balance between sensitivity and specificity	136
4.6.2	The larger context	137
4.6.3	Aspects of comparative genomics	137
4.6.4	Analysis of data sets from high throughput methods	138
4.7	Conclusions	138

5 Homology Modeling in Biology and Medicine 145

5.1	Introduction	145
5.1.1	The concept of homology modeling	145
5.1.2	How do homologous protein arise?	146
5.1.3	The purposes of homology modeling	147
5.1.4	The effect of the genome projects	149
5.2	Input data	151
5.3	Methods	153
5.3.1	Modeling at different levels of complexity	153
5.3.2	Loop modeling	155
5.3.3	Side-chain modeling	171
5.3.4	Methods for complete modeling	184
5.4	Results	188
5.4.1	Range of targets	188
5.4.2	Example: amyloid precursor protein β -secretase	189
5.5	Strengths and limitations	194
5.6	Validation	195
5.6.1	Side-chain prediction accuracy	196
5.6.2	The CASP meetings	196
5.6.3	Protein health	198
5.7	Availability	199
5.8	Appendix	199
5.8.1	Backbone conformations	199
5.8.2	Side-chain conformational analysis	208

6 Protein Structure Prediction 237

6.1	Overview	238
6.1.1	Definition of terms	241
6.1.2	What is covered in this chapter	243

6.2	Data	245
6.2.1	Input data	245
6.2.2	Output data	246
6.2.3	Additional input data	246
6.2.4	Structure comparison and classification	247
6.2.5	Scoring functions and (empirical) energy potentials	249
6.3	Methods	254
6.3.1	Secondary structure prediction	255
6.3.2	Knowledge-based 3D structure prediction	256
6.4	Results	273
6.4.1	Remote homology detection	275
6.4.2	Structural genomics	282
6.4.3	Selecting targets for structural genomics	283
6.4.4	Genome annotation	284
6.4.5	Sequence-to-structure-to-function paradigm	284
6.5	Validation of predictions	287
6.5.1	Benchmark set tests	287
6.5.2	Blind prediction experiments (CASP)	289
6.6	Conclusion: strengths and limitations	292
6.6.1	Threading	292
6.6.2	Strengths	293
6.6.3	Limitations	294
6.7	Accessibility	295
7	Protein–Ligand Docking in Drug Design	315
7.1	Introduction	315
7.1.1	A taxonomy of docking problems	316
7.1.2	Application scenarios in structure-based drug design	318
7.2	Methods for protein–ligand docking	319
7.2.1	Rigid-body docking algorithms	320
7.2.2	Flexible ligand docking algorithms	324
7.2.3	Docking by simulation	332
7.2.4	Docking of combinatorial libraries	336
7.2.5	Scoring protein–ligand complexes	338
7.3	Validation studies and applications	340
7.3.1	Reproducing X-ray structures	340
7.3.2	Validated blind predictions	342
7.3.3	Screening small molecule databases	342
7.3.4	Docking of combinatorial libraries	344
7.4	Molecular docking in practice	344
7.4.1	Preparing input data	345
7.4.2	Analyzing docking results	345
7.4.3	Choosing the right docking tool	346
7.5	Concluding remarks	347
7.6	Software accessibility	348

8	Modelling Protein–Protein and Protein–DNA Docking	361
8.1	Introduction	361
8.1.1	The need for protein–protein and protein–DNA docking	361
8.1.2	Overview of the computational approach	362
8.1.3	Scope of this chapter	364
8.2	Structural studies of protein complexes	364
8.3	Methodology of a protein–protein docking strategy	366
8.3.1	Rigid body docking by Fourier correlation theory	366
8.3.2	Use of residue pair potentials to re-rank docked complexes	373
8.3.3	Use of distance constraints	376
8.3.4	Refinement and additional screening of complexes	377
8.3.5	Implementation of the docking suite	379
8.4	Results from the protein–protein docking strategy	380
8.5	Modelling protein–DNA complexes	384
8.5.1	Method	385
8.5.2	Results	387
8.6	Strategies for protein–protein docking	389
8.6.1	Evaluation of the results of docking simulations	389
8.6.2	Fourier correlation methods	390
8.6.3	Other rigid-body docking approaches	391
8.6.4	Flexible protein–protein docking	392
8.6.5	Rigid-body treatment to re-rank putative docked complexes	393
8.6.6	Introduction of flexibility to re-rank putative docked complexes	394
8.7	Blind trials of protein–protein docking	395
8.8	Energy landscape for protein docking	397
8.9	Conclusions	398
	<i>Appendix</i>	405
	Glossary of Algorithmic Terms in Bioinformatics	405
	<i>Subject Index</i>	425
	<i>Name Index</i>	439

Part II: Applications

List of Contributors *xvii*

Foreword *xix*

1	Integrating and Accessing Molecular Biology Resources	3
1.1	Introduction	3
1.2	Molecular biology resources	4
1.2.1	Databases	4
1.2.2	Applications	9
1.2.3	Databases and applications world-wide	9
1.3	An overview of SRS	10
1.3.1	The meta-definition layer	11
1.3.2	The SRS core	12
1.3.3	Wrappers	12
1.3.4	Clients	13
1.4	Integrating molecular biology resources	13
1.4.1	The SRS token server	14
1.4.2	Meta definition of molecular biology resources	14
1.4.3	Indexing databases	16
1.4.4	Querying and linking databases	16
1.4.5	Views and object loader	16
1.4.6	Applications – analyzing data	17
1.5	The SRS data warehouse	18
1.6	Accessing integrated data	18
1.6.1	The web interface	19
1.6.2	The application programmers' interfaces (APIs)	21
1.6.3	Other interfaces	22
1.7	Other approaches	22
1.8	Conclusions	23
2	Bioinformatics Support of Genome Sequencing Projects	25
2.1	Introduction	25
2.2	Methods	30

2.2.1	Fast identification of pairs of similar reads	31
2.2.2	Clipping of poor end regions	34
2.2.3	Computation and evaluation of overlaps	36
2.2.4	Construction of contigs	36
2.2.5	Construction of consensus sequences	38
2.3	An example	39
2.4	Other assembly programs	44
2.5	Conclusion	46
3	Analysis of Sequence Variations	49
3.1	Introduction	49
3.2	Sequence variation	49
3.3	Linkage analysis	50
3.4	Association analysis	53
3.5	Why is genetic analysis shifting to single nucleotide polymorphisms (SNPs)?	54
3.6	SNP Discovery	55
3.7	Genotyping technologies	56
3.8	SNPs are frequent in the human genome and their organization complex	59
3.9	Pooling strategies	61
3.10	Conclusions	62
4	Proteome analysis	69
4.1	Introduction and principles	69
4.2	Protein separation	74
4.2.1	Experimental aspects: description of the 2-DE technology, a wet-lab technique, from a wet-lab point of view	74
4.2.2	The use of 2-DE as a tool towards diagnostics and disease description	76
4.3	Computer analysis of 2-DE gel images	77
4.3.1	2-DE analysis software	82
4.4	Identification and characterization of proteins after separation	92
4.4.1	Introduction	92
4.4.2	The tools	93
4.5	Proteome databases	98
4.5.1	Introduction	98
4.5.2	Protein sequence databases	98
4.5.3	Nucleotide sequence databases	102
4.5.4	Databases for protein families, domains and functional sites: InterPro	103
4.5.5	2-DE databases	104
4.5.6	Post-translational modification databases	105
4.5.7	Conclusion	106

4.6	Automation in proteome analysis	106
4.6.1	Introduction	106
4.6.2	Robotized protein identification using peptide mass fingerprinting	107
4.6.3	The molecular scanner	109
4.6.4	Other techniques	112
4.7	Conclusion	113
5	Target Finding in Genomes and Proteomes	119
5.1	Introduction	119
5.2	Experimental design for large-scale gene expression studies and drug target identification	120
5.3	Computational analyses in drug target discovery	123
5.3.1	Shannon entropy	123
5.3.2	Clustering	126
5.3.3	Combining analytical methods in the development of experimental therapies	129
5.3.4	How these methods were selected	130
5.3.5	Welcome to the future: reverse engineering of genetic networks	130
5.3.6	Genomes and proteomes	133
5.4	Concluding remarks	133
6	Screening of Drug Databases	137
6.1	Introduction	137
6.2	Methods for virtual screening	140
6.2.1	Ligand similarity-based virtual screening	140
6.2.2	Structure-based virtual screening	145
6.3	Practical virtual screening	147
6.4	First test scenario: application of fast similarity searching algorithms	148
6.4.1	Library generation	148
6.4.2	Computational details	150
6.4.3	Discussion of screening results	150
6.5	Second test scenario: docking as a tool for virtual screening tool	158
6.5.1	Library generation	159
6.5.2	Docking procedure	159
6.5.3	Discussion of docking results	160
6.6	Conclusions and outlook	165
7	Future Trends	171
7.1	How does the progress in genomics and bioinformatics change our view on biology and medicine?	171
7.2	What are the main coming challenges for bioinformatics?	173

7.2.1	New experimental data	174
7.2.2	New analysis methods	176
7.2.3	Integrated views on biology	179
7.3	What are well founded expectations and intrinsic limitations of bioinformatics?	183
	References	185
	<i>Subject Index</i>	189
	<i>Name Index</i>	203

List of Contributors

Part I: Basic Technologies

Prof. Ron D. Appel
Swiss Institute of Bioinformatics
Proteome Informatics Group
CMU-1, rue Michel Servet
1211 Geneva 4
Switzerland
ron.appel@isb-sib.ch

Dr. Amos Bairoch
Swiss Institute of Bioinformatics
SWISS-PROT group
CMU-1, rue Michel Servet
1211 Geneva 4
Switzerland
Amos.Bairoch@isb-sib.ch

Dr. Pierre-Alain Binz
Swiss Institute of Bioinformatics
Proteome Informatics group
CMU-1, rue Michel Servet
1211 Geneva 4
Switzerland
Pierre-Alain.Binz@isb-sib.ch

Dr. Christopher S. Carlson
University of Washington
Department of Molecular Biotechnology
Box 357730
Seattle, WA 98195
USA
peterpan@mbt.washington.edu

Prof. Roland L. Dunbrack, Jr.
Institute for Cancer Research
Fox Chase Cancer Center
7701 Burholme Avenue
Philadelphia, PA 19111
USA
rl_dunbrack@fccc.edu

Dr. Thure Etzold
Lion Bioscience Ltd.
Sheraton House, Castle Business Park
Cambridge CB3 0AX

United Kingdom
etzold@lionbio.uk.com

Dr. Stefanie Fuhrman
Incyte Pharmaceuticals, Inc.
3174 Porter Dr.
Palo Alto, CA 94304
USA
sfuhrman@incyte.com

Dr. Elisabeth Gasteiger
Swiss Institute of Bioinformatics
SWISS-PROT group
CMU-1, rue Michel Servet
1211 Geneva 4
Switzerland
Elisabeth.Gasteiger@isb-sib.ch

Dr. David P. Hansen
Lion Bioscience Ltd.
Sheraton House, Castle Business Park
Cambridge CB3 0AX
United Kingdom
David.Hansen@uk.lionbioscience.com

Prof. Dr. Denis F. Hochstrasser
Laboratoire Central de Chimie
Clinique
Hôpital Cantonal Universitaire
24, rue Micheli-du-Crest
1211 Genève 14
Switzerland
Denis.Hochstrasser@dim.hcuge.ch

Prof. Xiaoqi Huang
Department of Computer Science
Iowa State University
226 Atanasoff Hall
Ames, IA 50011
USA
xghuang@cs.iastate.edu

Prof. Gerhard Klebe
Philipps-Universität Marburg

Institut für Pharmazeutische Chemie
Marbacher Weg 6
35032 Marburg
Germany
klebe@mail.uni-marburg.de

Prof. Thomas Lengauer
Fraunhofer Institute for Algorithms
and Scientific Computing
Schloss Birlinghoven
53754 Sankt Augustin
Germany
(present address:
Max-Planck-Institute for Computer
Science
Stuhlsatzenhausweg 85
66123 Saarbrücken
Germany
lengauer@mpi-sb.mpg.de)

Dr. Shoudan Liang
Incyte Pharmaceuticals, Inc.
3174 Porter Dr.
Palo Alto, CA 94304
USA
sliang@incyte.com

Dr. Gidon Moont
Imperial Cancer Research Fund
Biomolecular Modelling Laboratory
44 Lincoln's Inn Fields
London WC2A 3PX
United Kingdom
moont@icrf.icnet.uk

Prof. Deborah Nickerson
University of Washington
Department of Molecular Biotechnology
Box 357730
Seattle, WA 98195
USA
debnick@u.washington.edu

Dr. Matthias Rarey
Fraunhofer Institute for Algorithms
and Scientific Computing
Schloss Birlinghoven
53754 Sankt Augustin
Germany
matthias.rarey@gmd.de

Dr. Mark J. Rieder
University of Washington
Department of Molecular Biotechnology
Box 357730
Seattle, WA 98195
USA
mrieder@uwashington.edu

Dr. Jean-Charles Sanchez
Laboratoire Central de Chimie Clinique
Hôpital Cantonal Universitaire
24, rue Micheli-du-Crest
1211 Genève 14
Switzerland
Jean-Charles.Sanchez@dim.hcuge.ch

Victor Solovye
EOS Biotechnology
225A Gateway Boulevard
South San Francisco, CA 94080
USA
solovye@eosbiotech.com

Dr. Roland Somogyi
Incyte Pharmaceuticals, Inc.
3174 Porter Dr.
Palo Alto, CA 94304
USA
rsomogyi@incyte.com

Dr. Martin Stahl
Pharmaceuticals Division
F. Hoffmann-La Roche AG
4070 Basel
Switzerland

Dr. Michael J. E. Sternberg
Imperial Cancer Research Fund
Biomolecular Modelling Laboratory
44 Lincoln's Inn Fields
London WC2A 3PX
United Kingdom
m.sternberg@icrf.icnet.uk

Dr. Martin Vingron
Max-Planck-Institute of Molecular
Genetics
Innstraße 73
14195 Berlin
Germany
vingron@molgen.mpg.deGermany

Dr. Xiling Wen
Incyte Pharmaceuticals, Inc.
3174 Porter Dr.
Palo Alto, CA 94304
USA
xwen@incyte.com

Prof. Dr. Ralf Zimmer
Ludwig-Maximilians-Universität München
Institut für Informatik
Theresienstraße 39
80333 München
Germany
zimmer@bio.informatik.uni-muenchen.de

Part II: Applications

Prof. Ron D. Appel
Swiss Institute of Bioinformatics
Proteome Informatics Group
CMU – 1, rue Michel Servet
1211 Geneva 4
Switzerland
ron.appel@isb-sib.ch

Prof. Roland L. Dunbrack, Jr
Institute for Cancer Research
Fox Chase Cancer Center
7701 Burholme Avenue
Philadelphia, PA 19111
USA
rl_dunbrack@fccc.edu

Dr. Thure Etzold
Lion Bioscience Ltd.
Sheraton House, Castle Business Park
Cambridge CB3 0AX
UK
etzold@lionbio.uk.com

Prof. Xiaoqiu Huang
Department of Computer Science
Iowa State University
226 Atanasoff Hall
Ames, IA 50011
USA
xghuang@cs.iastate.edu

Prof. Gerhard Klebe
Philipps-Universität Marburg
Institut für Pharmazeutische Chemie
Marbacher Weg 6
35032 Marburg
Germany
klebe@mail.uni-marburg.de

Prof. Thomas Lengauer
Fraunhofer Institute for Algorithms
and Scientific Computing

Schloss Birlinghoven
53754 Sankt Augustin
Germany
thomas.lengauer@gmd.de
(present address:
Max-Planck-Institute
for Computer Science
Stuhlsatzenhausweg 85
66123 Saarbrücken
Germany)

Prof. Deborah Nickerson
University of Washington
Department of Molecular
Biotechnology
Box 357730
Seattle, WA 98195
USA
debnick@u.washington.edu

Dr. Matthias Rarey
Fraunhofer Institute for Algorithms
and Scientific Computing
Schloss Birlinghoven
53754 Sankt Augustin
Germany
matthias.rarey@gmd.de

Victor Solovyev
EOS Biotechnology
225A Gateway Boulevard
South San Francisco, CA 94080
USA
solovyev@eosbiotech.com

Dr. Roland Somogyi
Incyte Pharmaceuticals, Inc.
3174 Porter Dr.
Palo Alto, CA 94304
USA
rsomogyi@incyte.com

Dr. Michael J. E. Sternberg
Imperial Cancer Research Fund
Biomolecular Modelling Laboratory
44 Lincoln's Inn Fields
London WC2A 3PX
UK
m.sternberg@icrf.icnet.uk

Dr. Martin Vingron
German Cancer Research Centre
(DKFZ)
Theoretical Bioinformatics Division
Im Neuenheimer Feld 280
69120 Heidelberg
Germany
m.vingron@dkfz-heidelberg.de

Dr. Thomas Werner
GSF-National Research Center for
Environmental and Health
Institute of Experimental Genetics
Ingolstädter Landstrasse 1
85764 Neuherberg
Germany
werner@gsf.de

Dr. Ralf Zimmer
Fraunhofer Institute for Algorithms
and Scientific Computing
Schloss Birlinghoven
53757 Sankt Augustin
Germany
ralf.zimmer@gmd.de

For my son Nico

Foreword

Computational biology and *bioinformatics* are terms for an interdisciplinary field joining information technology and biology that has skyrocketed in recent years. The field is located at the interface between the two scientific and technological disciplines that can be argued to drive a significant if not the dominating part of contemporary scientific innovation. In the English language, computational biology refers mostly to the scientific part of the field, whereas bioinformatics addresses mainly the infrastructure part. In some other languages (e.g. German) bioinformatics covers both aspects of the field.

The goal of this field is to provide computer-based methods for coping with and interpreting the genomic data that are being uncovered in large volumes through the diverse genome sequencing projects and other new experimental technology in molecular biology. The field presents one of the grand challenges of our times. It has a large basic research aspect, since we cannot claim to be close to understanding biological systems on an organism or even cellular level. At the same time, the field is faced with a strong demand for immediate solutions, because the genomic data that are being uncovered encode many biological insights whose deciphering can be the basis for dramatic scientific and economical success. At the end of the pre-genomic era that was characterized by the effort to sequence the human genome we are entering the postgenomic era that concentrates on harvesting the fruits hidden in the genomic text. In contrast to the pregenomic era which, from the announcement of the quest to sequence the human genome to its completion, has lasted less than 15 years, the postgenomic era can be expected to last much longer, probably extending over several generations.