Approaches in Integrative Bioinformatics

Ming Chen • Ralf Hofestädt
Editors

Approaches in Integrative Bioinformatics

Towards the Virtual Cell

Springer

Editors

Ming Chen
College of Life Sciences
Zhejiang University
Hangzhou
People's Republic of China

Ralf Hofestädt
Department of Bioinformatics
and Medical Informatics
Bielefeld University
Bielefeld, Germany

Additional material to this book can be downloaded from http://extras.springer.com

ISBN 978-3-642-41280-6 ISBN 978-3-642-41281-3 (eBook)
DOI 10.1007/978-3-642-41281-3
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013956814

© Springer-Verlag Berlin Heidelberg 2014
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The unprecedented accumulation of high-throughput data from genomics, transcriptomics, proteomics, metabolomics, phenomics, etc., has resulted not only in new attempts to answer traditional biological questions and solve longstanding issues in biology but also in the formulation of novel hypotheses that arise precisely from this wealth of data. At the present, with thousands of biological data resources and information systems inside the Internet, an unknown number of analysis tools, and exponential growths of molecular data (especially high-throughput data), the storage, processing, description, transmission, connection, and integrative analysis of this data becomes a great challenge for bioinformatics. Thus, the so-called Big Data becomes the new keyword describing the actual situation for which new software tools are needed to analyze this exponentially increasing data.

Important applications of *Big Data* are systems biology and systems medicine. For instance, hospital information systems represent complex patient data. The diagnosis process is now supported by new methods of biotechnology using, for example, high-throughput sequencing approaches. Therefore, we have complex patient data inside the hospital information system which needs to be stored, transported, and analyzed. New software tools are needed to allow the user-specific data access and analysis of this data. Overall, to develop and implement new tools for automatic data integration and analysis will help implement better diagnostic methods in practice. In the future, the entire genomes of patients will be stored within hospital information systems. Furthermore, it will be necessary to share the genome sequences inside the hospital computer network and analyze the genome data to detect, for example, cancer genes. With the availability of Internet, the automatic integration and analysis of data are of the most relevant research topics in computer science. In biology, such tools have become more and more important. Methods like high-throughput sequencing and omics analysis are responsible for the exponential data generation process.

This book will focus on the integration and analysis of omics data. The *Introduction* will present relevant biological background and an overview of these actual methods. When the Internet merged, methods such as data fusion and federated database systems became relevant. The initial tools were implemented and gave birth to a new field of research: Integrative Bioinformatics, which strives to implement user-specific integration and analysis of complex data. The *Introduction* of this book will give a definition and overview of this pertinent field of research. Since then, complex information systems have been developed and implemented. Finally, the data warehouse concept became more relevant. Today the data warehouse concept is still the best construction for the implementation of integrative information systems. The *Information Fusion and Retrieval* section will focus on the said data warehouse concept. Furthermore, this part of the book will give an overview of information retrieval and data mining tools, which allow the user-specific identification and integration of data. Based on the methods described here, we are able to implement user-specific integration tools. The analysis of this data can be done using statistic, visualization, or animation tools. Furthermore, modeling and simulation are important analysis methods. The *Network Visualization, Modeling, and Analysis* section will focus on methods for network prediction, network modeling, and simulation. In the case of network simulation, we prefer the Petri net method, which allows the parallel simulation of complex metabolic pathways. Our application section is divided into two parts. First, we focus on methods of *BioData Mapping*. One interesting aspect is the possibility of molecular disease mapping which allows the pathway prediction of any disease and the semiautomatic mapping of this pathway into a virtual 3D cell. The genotype-phenotype map enables us to uncover the casual networks inside the "black box" that lies between genotypes and phenotypes with advances in high-throughput and high-dimensional genotyping and phenotyping technologies. Another important and actual topic is presented by the *Biocompution* section. After the reconstruction of a biological disease network, the identification of biomarkers or hubs for further analysis is important. To realize such tasks, the implementation of parallel algorithms is fundamental.

Important research topics for the next few years will be Big Data and Systems Medicine. Integrative Bioinformatics will be fundamental in developments for both fields and this book attempts to present an overview of relevant and actual research activities.

We are very grateful to all the authors for sharing their time, wisdom, and expertise. Finally, we want to thank Ms. Na Xu, the editor of Springer Beijing Office, for her continuous advice.

Hangzhou, People's Republic of China — Ming Chen
Bielefeld, Germany — Ralf Hofestädt
June 2013

Contents

Part I Introduction

1 **Integrative Bioinformatics** 3
Ming Chen and Ralf Hofestädt

2 **An Overview of Gene Regulation** 21
Andrew Harrison and Hugh Shanahan

Part II Information Fusion and Retrieval

3 **Information Retrieval in Life Sciences: A Programmatic Survey** 73
Matthias Lange, Ron Henkel, Wolfgang Müller,
Dagmar Waltemath, and Stephan Weise

4 **Data Warehouses in Bioinformatics** 111
Benjamin Kormeier

5 **Molecular Information Fusion in Ondex** 131
Jan Taubert and Jacob Köhler

6 **Text Mining on PubMed** 161
Timofey V. Ivanisenko, Pavel S. Demenkov,
and Vladimir A. Ivanisenko

Part III Network Visualization, Modeling and Analysis

7 **Network Visualization for Integrative Bioinformatics** 173
Andreas Kerren and Falk Schreiber

8 **Biological Network Modeling and Analysis** 203
Sebastian Jan Janowski, Barbara Kaltschmidt,
and Christian Kaltschmidt

9 **Petri Nets for Modeling and Analyzing Biochemical Reaction Networks** 245
Fei Liu and Monika Heiner

Part IV BioData Mapping

10 **Network Analysis and Integration in a Virtual Cell Environment** 275
Björn Sommer

11 **Bridging Genomics and Phenomics** 299
Dijun Chen, Ming Chen, Thomas Altmann,
and Christian Klukas

Part V Biocompution

12 **Parallel Computing for Gene Networks Reverse Engineering** 337
Jaroslaw Zola

13 **Computational Biomarker Discovery** 355
Fan Zhang, Xiaogang Wu, and Jake Y. Chen