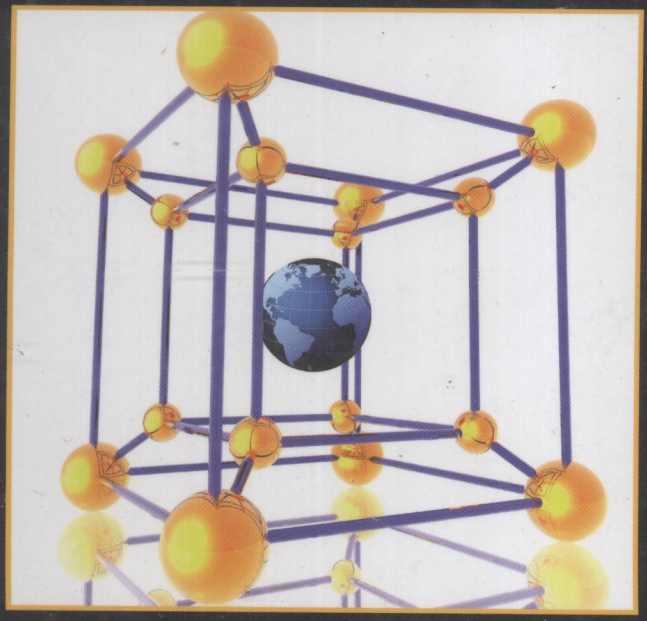
WILEY

Grid Computing for Bioinformatics and Computational Biology



Edited by
EL-GHAZALI TALBI
ALBERT Y. ZOMAYA

Wiley Series on Bioinformatics: Computational Techniques and Engineering
Yi Pan and Albert Y. Zomaya, Series Editors

GRID COMPUTING FOR BIOINFORMATICS AND COMPUTATIONAL BIOLOGY

Edited by

El-Ghazali Talbi
University of Lille, Lille, France

Albert Y. Zomaya
University of Sydney, Sydney, Australia

WILEY-INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Wiley Bicentennial Logo: Richard J. Pacifico

Library of Congress Cataloging-in-Publication Data:
Grid computing for bioinformatics and computational biology / edited by
El-Ghazali Talbi, Albert Y. Zomaya.
p. cm.
Includes index.
ISBN 978-0-471-78409-8 (alk. paper)
1. Bioinformatics. 2. Computational biology. 3. Computational grids (Computer systems)
I. Talbi, El-Ghazali, 1965 - II. Zomaya, Albert Y.
QH324.2.G75 2007
570.285–dc22

2007019075

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

GRID COMPUTING FOR BIOINFORMATICS AND COMPUTATIONAL BIOLOGY

THE WILEY BICENTENNIAL–KNOWLEDGE FOR GENERATIONS

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!

William J. Pesce

WILLIAM J. PESCE
PRESIDENT AND CHIEF EXECUTIVE OFFICER

Peter Booth Wiley

PETER BOOTH WILEY
CHAIRMAN OF THE BOARD

To my daughter Besma, you are my sweet little girl with an impressive intelligence.
To my mother Zehour for her infinite sacrifice.
To my father Ammar who continues to support me in my academic research.
In memory of my uncle Djerabi Hocine. He was like a Big Brother to me.

Prof. E-G. Talbi

To my students for teaching me so many things.

Prof. A.Y. Zomaya

PREFACE

Bioinformatics is fast emerging as an important discipline for academic research and industrial applications. Research and development in bioinformatics and computational biology create and develop advanced information and computational techniques to manage and extract useful information from the DNA/RNA/protein sequence being generated by high-throughput technologies (e.g., DNA microarrays, DNA sequencers). It is the comprehensive application of mathematics (e.g., probability and graph theory), statistics, science (e.g., biochemistry), and computer science (e.g., computer algorithms and machine learning) to the understanding of living systems. These techniques are extremely computationally or data intensive, providing motivation for using grids.

Grids are an enabling technology that permits the transparent coupling of geographically dispersed resources (machines, networks, data storage, visualization devices, and scientific instruments) for large-scale distributed applications. Grids provide several important benefits for users and applications to share: computing and data storage, knowledge, instruments, and so on.

This book not only presents grid algorithms and applications, but also presents software frameworks and libraries that integrate bioinformatics and computational biology applications. Many researchers in this field are not aware of the existence of these frameworks which encourages the reusing of existing code with a high level of transparency in the target grid platform.

The intended audience is mainly in research, development, and some industrial fields (pharmaceutical, biotechnology, etc.). Research and development concerns many domains: bioinformatics, computational biology, grid computing, data mining, and so on. Many biological researchers are also using grids in problem solving.

Many undergraduate courses worldwide on bioinformatics and grid computing would be interested in the contents because of the introductory part of each chapter and the additional information on Internet topical resources. In addition, Ph.D. courses related to grid for bioinformatics are a direct target for this book.

This book's purpose is to serve as a single up-to-date source for grid issues to the bioinformatics and computational biology communities. Its text provides details on modern and ongoing research on grid applications for bioinformatics and computational biology and is organized following two different categories grid platforms and grid algorithms.

Some beginning chapters present a state-of-the-art on data and computational grid platforms devoted to some challenging problems in bioinformatics and computational

biology: docking and conformational analysis (Chapter 8), sequence analysis and phylogenetics (Chapter 4), pairwise sequence alignment (Chapter 5), heterogeneous biomedical data grids (Chapter 7), RNA folding (Chapter 11), and biological sequence comparison (Chapter 13).

Other chapters illustrate grid services such as OpenMolGRID (open computing grid for molecular sciences and engineering, Chapter 1), WISDOM (molecular docking, Chapter 9), interactive visualization and analysis of biomedical data (Chapter 10), and semantic mediation architecture for a clinical data grid (Chapter 12), and biomedical network modeling (Chapter 16).

Grid algorithms for well-known problems in bioinformatics and computational biology are presented: sequence alignment (Chapter 2), multiple sequence alignment for protein sequences (Chapter 3), phylogenetics (Chapter 6), protein threading (Chapter 14), and DNA fragment assembly (Chapter 15).

EL-GHAZALI TALBI
ALBERT Y. ZOMAYA

Lille, France
Sydney, Australia
September 2007

ACKNOWLEDGMENTS

Thanks to all the contributors for their cooperation in bringing this book to completion.

Professor Talbi would like to thank all the members of his research team DOLPHIN: J-C. Boisson, M. Basseur, C. Boutroue, J. Brongniart, F. Clautiaux, C. Dhaenens, G. Even, L. Jourdan, N. Jozefowiez, M. Khabzaoui, J. Lemesre, A. Liefooghe, C. Luit, N. Melab, N. Mezmaz, A. Tantar, and E. Tantar. Also, he would like to dedicate this book to the memory of his Ph.D. student, S. Cahon.

Professor Zomaya would like to acknowledge the support of his research team at the Advanced Networks Research Lab at Sydney University.

Finally, we are grateful to the support and patience of the team from John Wiley & Sons, without their help this book would not have been possible.

CONTRIBUTORS

ENRIQUE ALBA, Department of Computer Language and Science, ETSI Informatica, Campus Teatinos, Malaga, Spain.

RUMEN ANDONOV, IRISA, Campus de Beaulieu, Rennes, France.

RYUZO AZUMA, RIKEN Genomic Sciences Center, Tsurumi, Yokohama, Kanagawa, Japan.

ADAM L. BAZINET, Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland.

EMILIO BENFENATI, Mario Negri Institute of Pharmaceutical Research, Milano, Italy.

SIEGFRIED BENKNER, Institute of Scientific Computing, University of Vienna, Wien, Austria.

AZZEDINE BOUKERCHE, University of Ottawa, Ontario, Canada.

VINCENT BRETON, IN2P3 CNRS, LPC Clermont-Ferrand, Aubière, France.

MARIAN BUBAK, Institute of Computer Science, AGH, Krakow, Poland.

GUILLAUME COLLET, IRISA, Campus de Beaulieu, Rennes, France.

MICHAEL P. CUMMINGS, Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland.

VIPIN CHAUDHARY, Department of Computer Science and Engineering, The State University of New York, Buffalo, New York.

CHUNXI CHEN, School of Computer Engineering, Nanyang Technology University, Singapore.

HANS DE STERCK, Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada.

W. DUBITZKY, School of Biomedical Sciences, University of Ulster, Coleraine, United Kingdom.

G. ENGELBRECHT, Institute of Scientific Computing, University of Vienna, Wien, Austria.

JOCHEN FINGBERG, C&C Research Laboratories, NEC Europe, Ltd., St. Augustin, Germany.

J-F. GIBRAT, Mathematics Information and Genome Unit, INRA, Joue-en-Josas, France.

MICAH HAMADY, Department of Computer Science, University of Colorado, Boulder, Colorado.

NICOLAS JACQ, IN2P3 CNRS, LPC Clermont-Ferrand, Aubière, France.

VINOD KASAM, IN2P3 CNRS, LPC Clermont-Ferrand, Aubière, France.

ROB KNIGHT, Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado.

AKIHIKI KONAGAYA, RIKEN Genomic Sciences Center, Tsurumi, Yokohama, Kanagawa, Japan.

FUMIKAZU KONISHI, RIKEN Genomic Sciences Center, Tsurumi, Yokohama, Kanagawa, Japan.

KAI KUMPF, Department of Bioinformatics, Fraunhofer-Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, St. Augustin, Germany.

JOSEPH LANDMAN, Scalable Informatics LLC, Canton, Michigan.

WEIGUO LIU, School of Computer Engineering, Nanyang Technology University, Singapore.

GABRIAL LUQUE, Department of Computer Language and Science, ETSI Informatica, Campus Teatinos, Malaga, Spain.

ALBA CRISTINA MAGALHAES ALVES DE MELO, University of Brazil, Rio de Janeiro, Brazil.

A. MARIN, IRISA, Campus de Beaulieu, Rennes, France.

KAZUMI MATSUMURA, RIKEN Genomic Sciences Center, Tsurumi, Yokohama, Kanagawa, Japan.

NOUREDINE MELAB, LIFL, University of Lille, INRIA, CNRS, Lille, France.

DANIEL S. MYERS, Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland.

ANTONIO J. NEBRO, Department of Computer Language and Science, ETSI Informatica, Campus Teatinos, Malaga, Spain.

SHINGO OHKI, RIKEN Genomic Sciences Center, Tsurumi, Yokohama, Kanagawa, Japan.

ALEKS PAPO, Department of Applied Mathematics, University of Waterloo, Waterloo, Ontario, Canada.

VINCENT POIRRIEZ, LAMIH, University of Valenciennes, Valenciennes, France.

MATHILDE ROMBERG, School of Biomedical Sciences, University of Ulster, Coleraine, United Kingdom.

JEAN SALZEMANN, IN2P3 CNRS, LPC Clermont-Ferrand, Aubière, France.

BERTIL SCHMIDT, School of Computer Engineering, Nanyang Technology University, Singapore.

PETER M. A. SLOOT, Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands.

EL-GHAZALI TALBI, LIFL, University of Lille, INRIA, CNRS, Lille, France.

ALEXANDRU-ADRIAN TANTAR, LIFL, University of Lille, INRIA, CNRS, Lille, France.

ALFREDO TIRADO-RAMOS, Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands.

DENIS TRYSTRAM, ID-IMAG, Grenoble, France.

RYO UMETSU, RIKEN Genomic Sciences Center, Tsurumi, Yokohama, Kanagawa, Japan.

BHARADWAJ VEERAVALLI, Department of Electrical & Computer Engineering, The National University of Singapore, Singapore.

JOHN PAUL WALTERS, Institute for Scientific Computing, Wayne State University, Detroit, Michigan.

CHEN WANG, School of Information Technology, University of Sydney, Sydney, Australia.

ALEXANDER WÖHRER, Institute for Scientific Computing, University of Vienna, Wien, Austria.

N. YANEV, University of Sofia, Sofia, Bulgaria.

SUMI YOSHIKAWA, RIKEN Genomic Sciences Center, Suehiro, Tsurumi Yokohama, Kanagawa, Japan.

CHEN ZHANG, Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada.

JAROSLAW ZOLA, Institute of Computer & Information Sciences, Czestochowa University of Technology, Poland.

ALBERT Y. ZOMAYA, School of Information Technology, University of Sydney, Sydney, Australia.

BING BING ZHOU, School of Information Technology, University of Sydney, Sydney, Australia.

CONTENTS

1

OPEN COMPUTING GRID FOR MOLECULAR SCIENCES

Mathilde Romberg, Emilio Benfenati, and Werner Dubitzky

All substances are poisons; there is none that is not a poison. The right dose differentiates a poison from a remedy.

Paracelsus (1493–1541)

1.1 INTRODUCTION

The number of chemicals in society is largely increasing, and therewith the risk of being exposed to chemicals increases. Knowledge of possible toxic effects of these chemicals is vital, as are the measurement and assessment of the effects and related risks. Within the European Union, the Registration, Evaluation, and Authorisation of Chemicals (REACH) legislation [1] places responsibility on the chemical industries to properly assess the risks associated with their products. It has been estimated that about 30,000 new chemicals will be put on the European market in the coming years. The assessment of these chemicals would cost billions of euros and involve the use of millions of animals. REACH also aims to ensure that risks from substances of very high concern (SVHC) are properly controlled or that the substances are substituted. To match REACH requirements, fast and reliable methods with reproducible results are crucial, and regulatory bodies would be able to approve results. Property prediction and modeling will play an important role in this case [2].

Toxicology, the study of harmful interactions between chemicals and biological systems [3], uses more and more computer models. These models are based on already available data and help to reduce *in vivo* testing. Toxicity modeling and its data have many applications such as characterizing hazards, assessing environmental risks, and identifying potential lead components in drug discovery. A well-established method for toxicity modeling is quantitative structure–activity relationship (QSAR) or quantitative structure–property relationship (QSPR) [4,5]. On the basis of the available measured and calculated properties or activities and descriptors of compounds, predictive models for a certain property are built, which are then used to predict that

Grid Computing for Bioinformatics and Computational Biology. Edited by E.-G. Talbi and A.Y. Zomaya.

property for new compounds. An example for a property is the lethal dose (LD50), which is the amount of a substance that kills 50% of the population exposed to it. This property is mainly used to compare the toxicity of different compounds and to classify them, for example, for hazard warnings.

Classical QSAR models have been based on a very limited number of parameters, which have been measured (such as simple physicochemical properties) or calculated. The model target has been to find a relationship between these parameters and the property within a very limited congeneric series of chemicals. These chemicals share a common skeleton, and a few fragments are linked to it. In more recent years, there has been a significant change in the QSAR scenario: The interest has shifted from the identification of the relationship between the parameters and the property to a more practical use, the prediction of the properties of new chemicals. This calls attention to the predictive power of the model, since previously a model was not verified but was simply assessed with statistical measurements evaluating the fitting of the calculated values. Meanwhile, the challenge has become to model larger sets of compounds, and in addition the number of calculated chemical descriptors or fragments has drastically increased to several thousands. Finally, new more powerful algorithms are used, and these tools also introduce the possibility to extract new knowledge from the data instead of simply leading the algorithms toward well-known parameters based on *a priori* knowledge or hypotheses.

Classical bioinformatics applications such as data warehousing and data mining are a major part of the model development as a result of the following:

(a) The available data are stored in very different sources such as published journal papers, spreadsheets, and relational databases in different formats and notations with different nomenclature and

(b) Relations between the data are mined and used for building predictive models of various kinds such as multilinear regression (MLR), partial least squares (PLS), or artificial neural networks (ANNs).

Other applications applied within the process of prediction model development belong to the field of molecular modeling. Calculating certain properties of a molecule on the basis of its two- and three-dimensional structures provides the basis for the prediction of an endpoint such as LD50. Currently, the model-building and prediction process includes a variety of steps that a toxicologist or a pharmacologist would perform manually step by step by taking care of data selection, parameter setting, data format conversions and data transfer between each pair of subsequent steps, and so on.

Pharmaceutical industry and regulatory bodies together with environmental agencies are very interested in finding fast, cost-effective, easy, and reliable ways to identify compounds with respect to their toxicity. The process of determining lead compounds for a new drug takes years [6,7] in the laboratory, and in addition about 90% of the potential drugs entering the preclinical phase fail in further process due to their toxicity [8,9]. In recent years, pharmaceutical companies along with research initiatives have investigated modeling and prediction methods together with grid computing to

streamline and speed up processes. The prominent interest of industries lies in cost reduction, for example, reducing failure rate and using in-house PCs' idle time to run modeling tasks [10,11]. Software providers offer matching grid solutions [12,13] for the latter. These approaches exploit the embarrassingly parallel[1] nature of the applications and offer sophisticated scheduling mechanisms. They are deployed as in-house systems, that is, they do not span multiple organizations, mainly for security reasons. Companies do not risk their data and methods being exposed to outsiders.

Publicly funded research projects in bioinformatics investigate data and computational grid methods to integrate huge amounts of data, develop ontologies, model workflows, efficiently integrate application software including legacy codes, define standards, and offer easy-to-use and efficient tools [14–21].

Section 1.2 of this chapter will highlight grid systems in toxicology and drug discovery and their main characteristics. Section 1.3 will give an in-depth overview of the European OpenMolGRID approach, while Section 1.4 will conclude with an outlook for future developments.

1.2 GRIDS FOR TOXICOLOGY AND DRUG DISCOVERY

Toxicology covers important issues in life and environmental sciences. It is essential that the characteristics of a chemical be identified before producing and releasing it into the environment. In drug discovery, one aim is to exclude toxic, chemically unstable and biologically inactive compounds from the drug discovery [22,23] early on in the process. Therefore, models are being developed for predicting which compounds are liable to fail at a later stage of the process. In this context, QSAR models are one of the most popular methods. Another goal is to identify compounds that would bind to a given biological receptor. The area of docking is important to understand biological processes and find cures that succeed by activating or by inhibiting protein actions [24]. The docking studies require the modeling of the enzyme (which has to be known) in addition to the modeling of the small chemical compounds to be studied (ligand). However, these docking studies are more complex and do require a careful tridimensional description of the ligand. This is not always necessary in the case of QSAR models. For this reason, faster and simpler screening based on easier methods is often performed by drug companies, and the detailed docking studies are performed only for a limited number of chemicals. However, grid technologies introduce new possibilities.

The major objectives for using grid technology [25,26] to support biomathematics and bioinformatics approaches in drug discovery and related fields are to shorten the time to solution and reduce its costs. Users of such technology are (computational) biologists, pharmacologists, and chemists, who are usually not computer system experts. To bridge this gap, providing a user-friendly system is crucial. It allows

[1] An application is called embarrassingly parallel if no particular effort is needed to split it into a large number of independent problems that can be executed in parallel, and these processes do not need to communicate with each other.