

# Lecture Notes in Bioinformatics

4360

Edited by S. Istrail, P. Pevzner, and M. Waterman

Editorial Board: A. Apostolico S. Brunak M. Gelfand  
T. Lengauer S. Miyano G. Myers M.-F. Sagot D. Sankoff  
R. Shamir T. Speed M. Vingron W. Wong

Subseries of Lecture Notes in Computer Science

Werner Dubitzky Assaf Schuster  
Peter M.A. Sloot Michael Schroeder  
Mathilde Romberg (Eds.)

# Distributed, High-Performance and Grid Computing in Computational Biology

International Workshop, GCCB 2006  
Eilat, Israel, January 21, 2007  
Proceedings

## Series Editors

Sorin Istrail, Brown University, Providence, RI, USA

Pavel Pevzner, University of California, San Diego, CA, USA

Michael Waterman, University of Southern California, Los Angeles, CA, USA

## Volume Editors

Werner Dubitzky

Mathilde Romberg

University of Ulster, School of Biomedical Sciences, Coleraine, UK

E-mail: {w.dubitzky, me.romberg}@ulster.ac.uk

Assaf Schuster

Technion – Israel Institute of Technology, Haifa, Israel

E-mail: assaf@cs.technion.ac.il

Peter M.A. Sloot

University of Amsterdam, Section of Computational Science, The Netherlands

E-mail: sloot@science.uva.nl

Michael Schroeder

Biotec, Dresden University of Technology, Germany

E-mail: Michael.Schroeder@biotec.tu-dresden.de

Library of Congress Control Number: 2006939917

CR Subject Classification (1998): F.1, F.2, J.3, D.1, D.2, D.4, C.2, C.4, H.4

LNCS Sublibrary: SL 8 – Bioinformatics

ISSN 0302-9743

ISBN-10 3-540-69841-8 Springer Berlin Heidelberg New York

ISBN-13 978-3-540-69841-8 Springer 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. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 2007

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper SPIN: 11976158 06/3142 5 4 3 2 1 0

# Preface

This volume of the Springer *Lecture Notes in Computer Science* series contains the contributions presented at the International Workshop on Distributed, High-Performance and Grid Computing in Computational Biology 2006 (GCCB 2006) held in Eilat, January 21, 2007 in conjunction with the fifth European Conference on Computational Biology (ECCB 2006).

Modern computational biology and bioinformatics are characterized by large and complex-structured data and by applications requiring considerable computing resources, such as processing units, storage elements and software programs. In addition, these disciplines are intrinsically geographically distributed in terms of their instruments, communities and computing resources. Tackling the computational challenges in computational biology and bioinformatics increasingly requires high-end and distributed computing infrastructures, systems and tools. The main objective of this workshop is to bring together researchers and practitioners from these areas to discuss ideas and experiences in developing and applying distributed, high-performance and grid computing technology to problems in computational biology and bioinformatics.

The challenges in distributed, high-performance and grid computing in biology and biotechnology are inherently more complicated than those in such domains as physics, engineering and conventional business areas. Some of the added complexities arise from the:

- Conceptual complexity of biological knowledge and the methodologies used in biology and biotechnology
- Need to understand biological systems and processes at a detailed mechanistic, systemic and quantitative level across several levels of organization (ranging from molecules to cells, populations, and the environment)
- Growing availability of high-throughput data from genomics, transcriptomics, proteomics, metabolomics and other high-throughput methods
- Widespread use of image data in biological research and development (microscopy, NMR, MRI, PET, X-ray, CT, etc.)
- Increasing number of investigations studying the properties and dynamic behavior of biological systems and processes using computational techniques (molecular dynamics, QSAR/QSPR, simulation of gene-regulatory, signaling and metabolic networks, protein folding/unfolding, etc)
- Requirement to combine data, information and compute services (e.g., sequence alignments) residing on systems that are distributed around the world
- Variety of different technologies, instruments, infrastructures and systems used in life science R&D
- Huge variety of information formats and frequent addition of new formats arising from new experimental protocols, instruments and phenomena to be studied

- Large and growing number of investigated biological and biomedical phenomena
- Fact that life science R&D is based heavily on the use of distributed and globally accessible computing resources (databases, knowledge bases, model bases, instruments, text repositories, compute-intensive services)

The GCCB workshop brought together computational biologists, bioinformaticians and life scientists who have researched and applied distributed, high-performance and grid computing technologies in the context of computational biology and bioinformatics. The workshop discussed innovative work in progress and important new directions. By sharing the insights, discussing ongoing work and the results that have been achieved, we hope the workshop participants conveyed a comprehensive view of the state of the art in this area and identified emerging and future research issues. We believe that the GCCB workshop made a valuable contribution in encouraging and shaping future work in the field of distributed, high-performance and grid computing in computational biology.

## Acknowledgements

GCCB 2006 was sponsored and organized by the University of Ulster, Coleraine, UK; the Technion - Israel Institute of Technology, Haifa, Israel; the University of Amsterdam, Amsterdam, The Netherlands; the Dresden University of Technology, Dresden, Germany; and the DataMiningGrid Consortium (EC grant IST-2004-004475, [www.DataMiningGrid.org](http://www.DataMiningGrid.org)). We are indebted to the ECCB 2006 Conference organizers Hershel Safer, Weizmann Institute of Science, and Haim Wolfson, Tel Aviv University. Finally, we would like to extend our gratitude to the members of the GCCB 2006 International Program Committee and to those who helped in the review process. And last but not least we want to thank all authors and workshop participants for their contributions and the valuable discussions.

January 2007

Werner Dubitzky  
Assaf Schuster  
Peter Slood  
Michael Schroeder  
Mathilde Romberg

# Organization

## Program Chairs

Werner Dubitzky, University of Ulster, Coleraine, UK

Assaf Schuster, Technion - Israel Institute of Technology, Haifa, Israel

Peter M.A. Sloot, University of Amsterdam, Faculty of Sciences, Amsterdam, The Netherlands

Michael Schroeder, Dresden University of Technology, Biotechnological Centre, Dresden, Germany

Mathilde Romberg, University of Ulster, Coleraine, UK

## Program Committee

David A. Bader, Georgia Tech, College of Computing, Atlanta, Georgia, USA

Eric Bremer, Children's Memorial Hospital, Northwestern University, Chicago, USA

Rui M. Brito, Universidade de Coimbra, Coimbra, Portugal

Marian Bubak, AGH - University of Science and Technology, Krakow, Poland

Kevin Burrage, The University of Queensland, Australia

Gustavo Deco, Universitat Pompeu Fabra, Barcelona, Spain

Frank Dehne, University of Ottawa, Ottawa, Canada

Guseppe di Fatta, University of Konstanz, Konstanz, Germany

Werner Dubitzky, University of Ulster, Coleraine, UK

Jordi Vill i Freixa, Universitat Pompeu Fabra, Barcelona, Spain

David Gilbert, University of Glasgow, Glasgow, UK

Carol Goble, University of Manchester, Manchester, UK

Danilo Gonzalez, Universidad de Talca, Talca, Chile

Ulrich Hansmann, Forschungszentrum Juelich, NIC, Juelich, Germany

Des Higgins, University College Dublin, Conway Institute, Dublin, Ireland

Alfons Hoekstra, University of Amsterdam, Amsterdam, The Netherlands

Martin Hoffmann, Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Sankt Augustin, Germany

Rod Hose, University of Sheffield, Sheffield, UK

Chun-Hsi (Vincent) Huang, University of Connecticut, Storrs, USA

Akihiko Konagaya, Riken Genomic Sciences Center, Yokohama City, Japan

Miron Livny, University of Wisconsin at Madison, Wisconsin, USA

Uko Maran, University of Tartu, Tartu, Estonia

Hartmut Mix, Dresden University of Technology, Dresden, Germany

Ron Perrot, Queens University, Belfast, UK

Mark Ragan, The University of Queensland, Australia

Stephen Robinson, University of Ulster, Coleraine, UK

Mathilde Romberg, University of Ulster, Coleraine, UK

Michael Schroeder, Dresden University of Technology, Biotechnological Centre,  
Dresden, Germany

Assaf Schuster, Technion - Israel Institute of Technology, Haifa, Israel

Gadi Schuster, Technion - Israel Institute of Technology, Haifa, Israel

Richard Sinnott, University of Glasgow, UK

Peter Sloot, University of Amsterdam, Faculty of Sciences, Amsterdam,  
The Netherlands

Craig A. Stewart, Indiana University, Indiana, USA

Domenico Talia, Università della Calabria, DEIS, Rende, Italy

Albert Zomaya University of Sydney, Sydney, Australia

## External Reviewers

Pinar Alper, University of Manchester, Manchester, UK

Michael Johnston, Universitat Pompeu Fabra, Barcelona, Spain

David Jones, University of Manchester, Manchester, UK

Peter Kral, Fraunhofer Institute for Algorithms and Scientific Computing SCAI,  
Sankt Augustin, Germany

Dean Kuo, University of Manchester, Manchester, UK

Jan Meinke, Forschungszentrum Juelich, NIC, Juelich, Germany

Javier Tamames, Universitat Pompeu Fabra, Barcelona, Spain

# Table of Contents

## Session 1a. “Sequence Analysis”

Combining a High-Throughput Bioinformatics Grid and Bioinformatics Web Services .....	1
<i>Chunyan Wang, Paul M.K. Gordon, Andrei L. Turinsky, Jason Burgess, Terry Dalton, and Christoph W. Sensen</i>	
Using Public Resource Computing and Systematic Pre-calculation for Large Scale Sequence Analysis .....	11
<i>Thomas Rattei, Mathias Walter, Roland Arnold, David P. Anderson, and Werner Mewes</i>	
Accelerated microRNA-Precursor Detection Using the Smith-Waterman Algorithm on FPGAs .....	19
<i>Patrick May, Gunnar W. Klau, Markus Bauer, and Thomas Steinke</i>	

## Session 1b. “Grids for Screening and Property Prediction”

Implementation of a Distributed Architecture for Managing Collection and Dissemination of Data for Fetal Alcohol Spectrum Disorders Research .....	33
<i>Andrew Arenson, Ludmila Bakhireva, Tina Chambers, Christina Deximo, Tatiana Foroud, Joseph Jacobson, Sandra Jacobson, Kenneth Lyons Jones, Sarah Mattson, Philip May, Elizabeth Moore, Kimberly Ogle, Edward Riley, Luther Robinson, Jeffrey Rogers, Ann Streissguth, Michel Tavares, Joseph Urbanski, Helen Yezerets, and Craig A. Stewart</i>	
Grid-Enabled High Throughput Virtual Screening .....	45
<i>Nicolas Jacq, Vincent Breton, Hsin-Yen Chen, Li-Yung Ho, Martin Hofmann, Hurng-Chun Lee, Yannick Legré, Simon C. Lin, Astrid Maaß, Emmanuel Medernach, Ivan Merelli, Luciano Milanesi, Giulio Rastelli, Matthieu Reichstadt, Jean Salzemann, Horst Schwichtenberg, Mahendrakar Sridhar, Vinod Kasam, Ying-Ta Wu, and Marc Zimmermann</i>	
Grid Computing for the Estimation of Toxicity: Acute Toxicity on Fathead Minnow ( <i>Pimephales promelas</i> ) .....	60
<i>Uko Maran, Sulev Sild, Paolo Mazzatorta, Mosé Casalegno, Emilio Benfenati, and Mathilde Romberg</i>	



Peer-to-Peer Experimentation in Protein Structure Prediction: An Architecture, Experiment and Initial Results .....	75
<i>Xueping Quan, Chris Walton, Dietlind L. Gerloff, Joanna L. Sharman, and Dave Robertson</i>	

## **Session 2a. “Data Management”**

Gene Prediction in Metagenomic Libraries Using the Self Organising Map and High Performance Computing Techniques.....	99
<i>Nigel McCoy, Shaun Mahony, and Aaron Golden</i>	
A Distributed System for Genetic Linkage Analysis .....	110
<i>Mark Silberstein, Dan Geiger, and Assaf Schuster</i>	
Enabling Data Sharing and Collaboration in Complex Systems Applications.....	124
<i>Michael A. Johnston and Jordi Villà-Freixa</i>	
Accessing Bio-databases with OGSA-DAI - A Performance Analysis ....	141
<i>Samatha Kottha, Kumar Abhinav, Ralph Müller-Pfefferkorn, and Hartmut Mix</i>	

## **Session 2b. “Collaborative Environments”**

Systems Support for Remote Visualization of Genomics Applications over Wide Area Networks .....	157
<i>Lars Ailo Bongo, Grant Wallace, Tore Larsen, Kai Li, and Olga Troyanskaya</i>	
HVEM DataGrid: Implementation of a Biologic Data Management System for Experiments with High Voltage Electron Microscope .....	175
<i>Im Young Jung, In Soon Cho, Heon Y. Yeom, Hee S. Kweon, and Jysoo Lee</i>	

<b>Author Index</b> .....	191
---------------------------	-----