

Lecture Notes in Bioinformatics

5109

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

Amos Bairoch Sarah Cohen-Boulakia
Christine Froidevaux (Eds.)

Data Integration in the Life Sciences

5th International Workshop, DILS 2008
Evry, France, June 25-27, 2008
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

Amos Bairoch

University of Geneva, Swiss Institute of Bioinformatics
and Department of Structural Biology and Bioinformatics

CMU – 1, rue Michel Servet, 1211 Geneva 4, Switzerland

E-mail: amos.bairoch@isb-sib.ch

Sarah Cohen-Boulakia

Christine Froidevaux

Université Paris-Sud 11, Laboratoire de Recherche en Informatique

Bâtiment 490, 91405 Orsay Cedex, France

E-mail: {cohen, chris}@lri.fr

Library of Congress Control Number: Applied for

CR Subject Classification (1998): H.2, H.3, H.4, J.3

LNCS Sublibrary: SL 8 – Bioinformatics

ISSN 0302-9743

ISBN-10 3-540-69827-2 Springer Berlin Heidelberg New York

ISBN-13 978-3-540-69827-2 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 2008

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12282746 06/3180 5 4 3 2 1 0

Preface

For several years now, there has been an exponential growth of the amount of life science data (e.g., sequenced complete genomes, 3D structures, DNA chips, mass spectroscopy data), most of which are generated by high-throughput experiments. This exponential corpus of data is stored and made available through a large number of databases and resources over the Web, but unfortunately still with a high degree of semantic heterogeneity and varying levels of quality. These data must be combined together and processed by bioinformatics tools deployed on powerful and efficient platforms to permit the uncovering of patterns, similarities and in general to help in the process of discovery. Analyzing complex, voluminous, and heterogeneous data and guiding the analysis of data are thus of paramount importance and necessitate the involvement of data integration techniques.

DILS 2008 was the fifth in a workshop series that aims at fostering discussion, exchange, and innovation in research and development in the area of data integration for the life sciences. Each previous DILS workshop attracted around 100 researchers from all over the world and saw an increase of submitted papers over the preceding one. This year was not an exception and the number of submitted papers increased to 54. The Program Committee selected 18 of them. The selected papers cover a wide spectrum of theoretical and practical issues including data annotation, Semantic Web for the life sciences, and data mining on integrated biological data.

Among these 18 papers, we distinguished 8 that describe research on new models, methods, or algorithms, and 8 that deal with the description of systems or experience with systems in practice. The two remaining papers have been selected for publication in a special issue of the *BMC Bioinformatics Journal*.

In addition to the presented papers, DILS 2008 featured three keynote talks by Olivier Bodenreider, National Library of Medicine, NIH, USA; Peter Karp, SRI International, USA; and Norman Paton, University of Manchester, UK. DILS 2008 also included a tutorial on bio-ontologies and a session dedicated to updates of biomolecular resources of world-wide importance: the UniProt knowledgebase and the EBI proteomics services.

The workshop was held at the University of Evry, in what is known as the ‘Genomic Valley’ at the heart of the Ile-de-France region, in France. DILS 2008 was kindly sponsored by the University of Paris-Sud 11, Microsoft Research who also made available their conference management system, the ENFIN network of Excellence, and the following institutes: IMGT, CEA, SIB, and CNRS (LRI and GDR BIM). We are very grateful to the University of Evry for hosting DILS, the MAISEL school for providing rooms for students, and the Genopole-Evry for its help in the local organization.

As editors of this volume, we thank all the authors who submitted papers, the Program Committee members and the external reviewers for their excellent work. Special thanks go to the local organizers, webmasters, Publicity and Sponsorship Chairs: Patrick Amar, Marie-Dominique Devignes, Nicole Lefèvre-Villain, Frédéric Lemoine, Isabelle Mougenot, Bastien Rance, Malika Smail, and Fariza Tahi. Finally, we are grateful for the cooperation of Springer in putting this volume together.

June 2008

Amos Bairoch
Sarah Cohen-Boulakia
Christine Froidevaux

Organization

Executive Committee

Program Chairs

Amos Bairoch	Swiss Institute of Bioinformatics, Swiss-Prot group, University of Geneva, Switzerland
Sarah Cohen-Boulakia	Laboratoire de Recherche en Informatique, CNRS UMR 8623, University of Paris-Sud 11, France
Christine Froidevaux	Laboratoire de Recherche en Informatique, CNRS UMR 8623, University of Paris-Sud 11, France

Program Committee

Amos Bairoch	SIB, University of Geneva, Switzerland
Sarah Cohen-Boulakia	LRI, University of Paris-Sud 11, France
Susan Davidson	University of Pennsylvania, USA
Marie-Dominique Devignes	LORIA, Nancy, France
Barbara Eckman	IBM, USA
Juliana Freire	University of Utah, USA
Christine Froidevaux	LRI, University of Paris-Sud 11, France
Floris Geerts	University of Edinburgh, UK
Amarnath Gupta	University of California San Diego, USA
Henning Hermjakob	EBI, UK
Ela Hunt	ETH Zurich, Switzerland
Minoru Kanehisa	Kyoto University, Japan
Jacob Koehler	University of Tromsø, Norway
Anthony Kosky	Axiop Inc., USA
Ulf Leser	Humboldt-Universität zu Berlin, Germany
Janice Lee Mong Li	National University of Singapore, Singapore
Frédérique Lisacek	SIB, University of Geneva, Switzerland
Bertram Ludäscher	University of California Davis, USA
Victor Markowitz	Lawrence Berkeley Labs, USA
Luc Moreau	University of Southampton, UK
Peter Mork	MITRE, USA
Fouzia Moussouni	INSERM, Rennes, France
Jignesh M. Patel	University of Michigan, USA

Manuel Peitsch	SIB, University of Basel, Switzerland
Erhard Rahm	University of Leipzig, Germany
Louïqa Raschid	University of Maryland, USA
Malika Smail	LORIA, Nancy, France
Val Tannen	University of Pennsylvania, USA
Thodoros Topaloglou	University of Toronto, Canada

External Reviewers

Sören Auer	Alasdair Gray	Timothy McPhillips
Jérôme Azé	Annika Gross	Anne Morgat
Shawn Bowers	Philip Groth	Krishna Palaniappan
I-Min Chen	Michael Hartung	Marie-Anne Poursat
Adrien Coulet	Samira Jaeger	Silke Trißl
Mohamed Elati	Toralf Kirsten	Patrick Ziegler
Isabelle Phan	Jérémie Mary	

Sponsoring Institutions

University of Paris-Sud 11	http://www.u-psud.fr/en/index.html/
Microsoft Research	http://research.microsoft.com/
ENFIN Network of Excellence	http://enfin.org/
IMGT	http://imgt.cines.fr/
CEA	http://www.cea.fr/
CNRS GDR BIM	http://www.gdr-bim.u-psud.fr/
CNRS LRI	http://www.lri.fr/
Swiss Institute of Bioinformatics	http://www.isb-sib.ch/

Sponsorship

Marie-Dominique Devignes	LORIA, Nancy, France
Malika Smail	LORIA, Nancy, France

Publicity

Julie Chabaliier	Faculty of medicine, University of Rennes 1, France
Fouzia Moussouni	INSERM, Rennes, France

Webmasters

Frédéric Lemoine	LRI, University of Paris-Sud 11, France
Bastien Rance	LRI, University of Paris-Sud 11, France

Local Organization (France)

Patrick Amar	LRI, University of Paris-Sud 11
Sarah Cohen-Boulakia	LRI, University of Paris-Sud 11
Christine Froidevaux	LRI, University of Paris-Sud 11
Frédéric Lemoine	LRI, University of Paris-Sud 11
Isabelle Mougenot	LIRMM, University of Montpellier 2
Bastien Rance	LRI, University of Paris-Sud 11
Fariza Tahi	IBISC, University of Evry

DILS 2008 website	http://dils2008.lri.fr/
-------------------	---

Table of Contents

Keynote Presentations

Ontologies and Data Integration in Biomedicine: Success Stories and Challenging Issues	1
<i>Olivier Bodenreider</i>	
BioWarehouse: Relational Integration of Eleven Bioinformatics Databases and Formats	5
<i>Peter D. Karp, Thomas J. Lee, and Valerie Wagner</i>	
Data Integration in the Life Sciences: Fun, Findings and Frustrations . . .	8
<i>Norman W. Paton</i>	

Semantic Web for the Life Sciences

Analyzing the Evolution of Life Science Ontologies and Mappings	11
<i>Michael Hartung, Toralf Kirsten, and Erhard Rahm</i>	
Ontology Design Principles and Normalization Techniques in the Web	28
<i>Xiaoshu Wang, Jonas S. Almeida, and Arlindo L. Oliveira</i>	
Exploiting Ontology Structure and Patterns of Annotation to Mine Significant Associations between Pairs of Controlled Vocabulary Terms	44
<i>Woei-Jyh Lee, Louisa Raschid, Hassan Sayyadi, and Padmini Srinivasan</i>	

Designing and Evaluating Architectures to Integrate Biological Data

Automatic Methods for Integrating Biomedical Data Sources in a Mediator-Based System	61
<i>Fleur Mougín, Anita Burgun, Olivier Bodenreider, Julie Chabalier, Olivier Loréal, and Pierre Le Beux</i>	
VisGenome and Ensembl: Usability of Integrated Genome Maps	77
<i>Joanna Jakubowska, Ela Hunt, John McClure, Matthew Chalmers, Martin McBride, and Anna F. Dominiczak</i>	
An Entity Resolution Framework for Deduplicating Proteins	92
<i>Lucas Lochovsky and Thodoros Topaloglou</i>	

New Architectures and Experience on Using Systems

Semantic Representation and Querying of caBIG Data Services	108
<i>E. Patrick Shironoshita, Ray M. Bradley, Yves R. Jean-Mary, Thomas J. Taylor, Michael T. Ryan, and Mansur R. Kabuka</i>	
SisGen: A CORBA-Based Data Management Program for DNA Sequencing Projects	116
<i>Georgios J. Pappas Jr., Robson P. Miranda, Natália F. Martins, Roberto C. Togawa, and Marcos M.C. Costa</i>	
Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species	124
<i>Frederic Bastian, Gilles Parmentier, Julien Roux, Sebastien Moretti, Vincent Laudet, and Marc Robinson-Rechavi</i>	
ENFIN - An Integrative Structure for Systems Biology	132
<i>Florian Reisinger, Manuel Corpas, John Hancock, Henning Hermjakob, Ewan Birney, and Pascal Kahlem</i>	

Systems Using Technologies from the Semantic Web for the Life Sciences

A System for Ontology-Based Annotation of Biomedical Data	144
<i>Clement Jonquet, Mark A. Musen, and Nigam Shah</i>	
Bio2RDF: A Semantic Web Atlas of Post Genomic Knowledge about Human and Mouse	153
<i>François Belleau, Nicole Tourigny, Benjamin Good, and Jean Morissette</i>	
OMIE: Ontology Mapping within an Interactive and Extensible Environment	161
<i>Amel Bouzeghoub, Abdeltif Elbyed, and Fariza Tahi</i>	
Chemical Knowledge for the Semantic Web	169
<i>Mykola Konyk, Alexander De Leon, and Michel Dumontier</i>	

Mining Integrated Biological Data

Combining One-Class Classification Models Based on Diverse Biological Data for Prediction of Protein-Protein Interactions	177
<i>José A. Reyes and David Gilbert</i>	
Semi Supervised Spectral Clustering for Regulatory Module Discovery	192
<i>Alok Mishra and Duncan Gillies</i>	

New Features of Major Resources for Biomolecular Data

UniProtKB/Swiss-Prot: New and Future Developments	204
<i>Amos Bairoch</i>	

EBI Proteomics Services	207
<i>Henning Hermjakob</i>	

DILS 2008 Tutorial

Bio-ontologies Tutorial	208
<i>Olivier Dameron and Julie Chabalier</i>	

Author Index	209
-------------------------------	-----