

Lecture Notes in Bioinformatics

4544

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

Sarah Cohen-Boulakia Val Tannen (Eds.)

Data Integration in the Life Sciences

4th International Workshop, DILS 2007
Philadelphia, PA, USA, June 27-29, 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

Sarah Cohen-Boulakia

University of Pennsylvania, Department of Computer and Information Science

303 Levine Hall, 3330 Walnut St., Philadelphia, PA 19104, USA

E-mail: sarahcb@seas.upenn.edu

Val Tannen

University of Pennsylvania, Department of Computer and Information Science

570 Levine Hall, 3330 Walnut St., Philadelphia, PA 19104, USA

E-mail: val@cis.upenn.edu

Library of Congress Control Number: 2007928915

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

LNCS Sublibrary: SL 8 – Bioinformatics

ISSN 0302-9743

ISBN-10 3-540-73254-3 Springer Berlin Heidelberg New York

ISBN-13 978-3-540-73254-9 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: 12081769 06/3180 5 4 3 2 1 0

Preface

Understanding the mechanisms involved in life (e.g., discovering the biological function of a set of proteins, inferring the evolution of a set of species) is becoming increasingly dependent on progress made in mathematics, computer science, and molecular engineering. For the past 30 years, new high-throughput technologies have been developed generating large amounts of data, distributed across many data sources on the Web, with a high degree of semantic heterogeneity and different levels of quality. However, one such dataset is not, by itself, sufficient for scientific discovery. Instead, it must be combined with other data and processed by bioinformatics tools for patterns, similarities, and unusual occurrences to be observed. Both data integration and data mining are thus of paramount importance in life science.

DILS 2007 was the fourth in a workshop series that aims at fostering discussion, exchange, and innovation in research and development in the areas of data integration and data management for the life sciences. Each previous DILS workshop attracted around 100 researchers from all over the world. This year, the number of submitted papers again increased. The Program Committee selected 19 papers out of 52 full submissions. The DILS 2007 papers cover a wide spectrum of theoretical and practical issues including scientific workflows, annotation in data integration, mapping and matching techniques, and modeling of life science data. Among the papers, we distinguished 13 papers presenting research on new models, methods, or algorithms and 6 papers presenting implementation of systems or experience with systems in practice. In addition to the presented papers, DILS 2007 featured two keynote talks by Kenneth H. Buetow, National Cancer Institute, and Junhyong Kim, University of Pennsylvania.

The workshop was held at the University of Pennsylvania, in Philadelphia, USA. It was kindly sponsored by the School of Engineering and Applied Science of the University of Pennsylvania, the Penn Genomics Institute, and Microsoft Research, who also made available their conference management system. 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 Susan Davidson, General Chair, Chris Stoeckert, PC Co-chair, as well as Olivier Biton, Tara Betterbid, and Howard Bilowsky. Finally, we are grateful for the cooperation and help of Springer in putting this volume together.

June 2007

Sarah Cohen-Boulakia
Val Tannen

Organization

Executive Committee

General Chair

Susan Davisdon, University of Pennsylvania, USA

Program Chairs

Chris Stoeckert, University of Pennsylvania, USA

Val Tannen, University of Pennsylvania, USA

Program Committee

Judith Blake	Jackson Laboratory, USA
Sarah Cohen-Boulakia	University of Pennsylvania, USA
Marie-Dominique Devignes	LORIA, Nancy, France
Barbara Eckman	IBM
Christine Froidevaux	LRI, University of Paris-Sud XI, France
Cesare Furlanello	ITC-irst, Trento, Italy
Jim French	University of Virginia, USA
Floris Geerts	University of Edinburgh, UK
Amarnath Gupta	University of California San Diego, USA
Ela Hunt	ETH Zurich, Switzerland
Jacob Koehler	Rothamsted Research, UK
Anthony Kosky	Axiopie Inc.
Hilmar Lapp	NESCENT
Ulf Leser	Humboldt-Universität zu Berlin, Germany
Bertram Ludäscher	University of California Davis, USA
Victor Markowitz	Lawrence Berkeley Labs
Peter Mork	MITRE
Tom Oinn	European Bioinformatics Institute, UK
Meral Ozsoyoglu	Case Western Reserve University, USA
John Quackenbush	Harvard, USA
Louïqa Raschid	University of Maryland, USA
Fritz Roth	Harvard Medical School, USA
Susanna-Assunta Sansone	European Bioinformatics Institute, UK
Kai-Uwe Sattler	Technical University of Ilmenau, Germany
Chris Stoeckert	University of Pennsylvania, USA
Val Tannen	University of Pennsylvania, USA
Olga Troyanskaya	Princeton University, USA

External Reviewers

Jérôme Azé
Jana Bauckmann
Julie Bernauer
Shawn Bowers
William Bug
Amy Chen

Adnan Derti
Francis Gibbons
Philip Groth
Timothy McPhillips
Joe Mellor
Krishna Palaniappan

Norbert Podhorszki
Murat Tasan
Weidong Tian
Silke Trißl
Daniel Zinn

Sponsorship Chair

Howard Bilofsky, University of Pennsylvania

Sponsoring Institutions

School of Engineering and Applied Science at the University of Pennsylvania
<http://www.seas.upenn.edu/>

Penn Genomics Institute
<http://www.genomics.upenn.edu/>

Microsoft Research
<http://research.microsoft.com/>

Web Site and Publicity Chairs

Olivier Biton University of Pennsylvania
Sarah Cohen-Boulakia University of Pennsylvania

DILS 2007 Web site <http://dils07.cis.upenn.edu/>

Table of Contents

Keynote Presentations

- Enabling the Molecular Medicine Revolution Through Network-Centric Biomedicine 1
Kenneth H. Buetow
- Phyl-O'Data (POD) from Tree of Life: Integration Challenges from Yellow Slimy Things to Black Crunchy Stuff 3
Junhyong Kim

New Architectures and Experience on Using Systems

- Automatically Constructing a Directory of Molecular Biology Databases 6
Luciano Barbosa, Sumit Tandon, and Juliana Freire
- The Allen Brain Atlas: Delivering Neuroscience to the Web on a Genome Wide Scale 17
Chinh Dang, Andrew Sodt, Chris Lau, Brian Youngstrom, Lydia Ng, Leonard Kuan, Sayan Pathak, Allan Jones, and Mike Hawrylycz
- Toward an Integrated RNA Motif Database 27
Jason T.L. Wang, Dongrong Wen, Bruce A. Shapiro, Katherine G. Herbert, Jing Li, and Kaushik Ghosh
- B-Fabric: A Data and Application Integration Framework for Life Sciences Research 37
Can Türker, Etzard Stolte, Dieter Joho, and Ralph Schlapbach
- SWAMI: Integrating Biological Databases and Analysis Tools Within User Friendly Environment 48
Rami Rifaieh, Roger Unwin, Jeremy Carver, and Mark A. Miller
- ^{my}Grid and UTOPIA: An Integrated Approach to Enacting and Visualising in Silico Experiments in the Life Sciences 59
Steve Pettifer, Katy Wolstencroft, Pinar Alper, Teresa Attwood, Alain Coletta, Carole Goble, Peter Li, Philip McDermott, James Marsh, Tom Oinn, James Sinnott, and David Thorne

Managing and Designing Scientific Workflows

- A High-Throughput Bioinformatics Platform for Mass Spectrometry-Based Proteomics 71
Thodoros Topaloglou, Moyez Dharsee, Rob M. Ewing, and Yury Bukhman

Bioinformatics Service Reconciliation by Heterogeneous Schema Transformation	89
<i>Lucas Zamboulis, Nigel Martin, and Alexandra Poulouvassilis</i>	
A Formal Model of Dataflow Repositories	105
<i>Jan Hidders, Natalia Kwasnikowska, Jacek Sroka, Jerzy Tyszkiewicz, and Jan Van den Bussche</i>	
Project Histories: Managing Data Provenance Across Collection-Oriented Scientific Workflow Runs	122
<i>Shawn Bowers, Timothy McPhillips, Martin Wu, and Bertram Ludäscher</i>	
Mapping and Matching Techniques	
Fast Approximate Duplicate Detection for 2D-NMR Spectra	139
<i>Björn Egert, Steffen Neumann, and Alexander Hinneburg</i>	
Ontology-Supported Machine Learning and Decision Support in Biomedicine	156
<i>Alexey Tsybmal, Sonja Zillner, and Martin Huber</i>	
Instance-Based Matching of Large Life Science Ontologies	172
<i>Toralf Kirsten, Andreas Thor, and Erhard Rahm</i>	
Modeling of Life Science Data	
Data Integration and Pattern-Finding in Biological Sequence with TESS’s Annotation Grammar and Extraction Language (AnGEL)	188
<i>Jonathan Schug, Max Mintz, and Christian J. Stoeckert Jr.</i>	
Inferring Gene Regulatory Networks from Multiple Data Sources Via a Dynamic Bayesian Network with Structural EM	204
<i>Yu Zhang, Zhidong Deng, Hongshan Jiang, and Peifa Jia</i>	
Accelerating Disease Gene Identification Through Integrated SNP Data Analysis	215
<i>Paolo Missier, Suzanne Embury, Conny Hedeler, Mark Greenwood, Joanne Pennock, and Andy Brass</i>	
Annotation in Data Integration	
What’s New? What’s Certain? – Scoring Search Results in the Presence of Overlapping Data Sources	231
<i>Philipp Hussels, Silke Triffl, and Ulf Leser</i>	
Using Annotations from Controlled Vocabularies to Find Meaningful Associations	247
<i>Woei-Jyh Lee, Louiqa Raschid, Padmini Srinivasan, Nigam Shah, Daniel Rubin, and Natasha Noy</i>	

CONANN: An Online Biomedical Concept Annotator 264
Lawrence H. Reeve and Hyoil Han

Author Index 281