Lecture Notes in Bioinformatics 3615

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

Subseries of Lecture Notes in Computer Science

Bertram Ludäscher Louiqa Raschid (Eds.)

Data Integration in the Life Sciences

Second International Workshop, DILS 2005
San Diego, CA, USA, July 20-22, 2005
Proceedings

Springer

Volume Editors

Bertram Ludäscher
University of California at Davis, Department of Computer Science
One Shields Ave, Davis, CA 95616, USA
E-mail: ludaesch@ucdavis.edu

Louiqa Raschid
University of Maryland, Department of Computer Science
A.V. Williams Building, College Park, MD 20742, USA
E-mail: louiqa@umiacs.umd.edu

Library of Congress Control Number: 2005928957

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

ISSN 0302-9743
ISBN-10 3-540-27967-9 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-27967-9 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11530084 06/3142 5 4 3 2 1 0

Preface

The explosion in the number and size of life science data resources, and the rapid growth in the variety and volume of laboratory data has been fueled by world-wide research activity and the emergence of new technologies. The modeling, management and analysis of this data often requires a comprehensive integration of heterogeneous and typically semistructured data, distributed across many possibly data sources. Recent interoperability standards such as XML and WSDL solve some (easy) problems, but data and process integration often remain time-consuming and error-pone manual tasks. The difficulty of these tasks is compounded by the high degree of semantic heterogeneity across data sources, varying data quality, and other domain-specific application requirements.

DILS 2005 was the 2nd International Workshop on Data Integration in the Life Sciences, following a successful first DILS workshop, March 2004 in Leipzig, Germany. For a specialized workshop, the DILS 2005 call for papers created a large interest (over 50 abstracts and eventually 42 paper submissions; an increase of over 20% over DILS 2004), out of which the international Program Committee selected 15 full papers, as well as 5 short papers, and 8 posters/demonstrations, which are all included in this volume. They cover a wide spectrum of theoretical and practical issues including scientific/clinical workflows, ontologies, tools and systems, and integration techniques. DILS 2005 also featured keynotes by Dr. Peter Buneman, Professor at the School of Informatics, University of Edinburgh, and Dr. Shankar Subramaniam, Professor at the Department of Bioengineering and Chemistry, UC San Diego. The program also included 6 invited presentations and reports on ongoing research activities in academia and industry and a panel organized by the AMIA Geomics Working Group.

The workshop was organized by the San Diego Supercomputer Center (SDSC) and took place July 20–22, 2005 at the University of California, San Diego. Additional sponsors included Microsoft Research, the American Medical Informatics Association (AMIA), the UC Davis Genome Center, and the University of Maryland Center for Bioinformatics and Computational Biology.

As the workshop co-chairs and editors of this volume, we thank all authors who submitted papers and the Program Committee members and external reviewers for their excellent work. Special thanks also go to Amarnath Gupta who served as workshop general chair, and his team, especially Donna Turner, Jon Meyer, and LInda Ferri, all at SDSC. We thank Chani Johnson and the Microsoft CMT Team for the excellent support of their paper management system. Finally, we thank Alfred Hofmann, Erika Siebert-Cole, and the team from Springer for their cooperation and help in putting this volume together.

June 2005 Bertram Ludäscher and Louiqa Raschid

2nd International Workshop on Data Integration in the Life Sciences (DILS)

University of California, San Diego
July 20–22, 2005

DILS 2005 Co-chairs

Amarnath Gupta	(General Chair)	University of California, San Diego, USA
Bertram Ludäscher	(PC Co-chair)	University of California, Davis, USA
Louiqa Raschid	(PC Co-chair)	University of Maryland, USA

Program Committee

Vineet Bafna	University of California, San Diego, USA
Chitta Baral	Arizona State University, USA
Judith Blake	Jackson Laboratory, USA
Shawn Bowers	University of California, Davis, USA
Terence Critchlow	Lawrence Livermore National Laboratory, USA
Alin Deutsch	University of California, San Diego, USA
Barbara Eckman	IBM Life Sciences, USA
Christoph Freytag	Humboldt University, Berlin, Germany
Floris Geerts	University of Edinburgh, UK
Carole Goble	University of Manchester, UK
Amarnath Gupta	University of California, San Diego, USA
Michael Gribskov	Purdue University, USA
Ralf Hofestaedt	University of Bielefeld, Germany
Hasan Jamil	Wayne State University, USA
Matthew Jones	University of California, Santa Barbara, USA
Jessie Kennedy	Napier University, Edinburgh, UK
Zoé Lacroix	Arizona State University, USA
Ulf Leser	Humboldt University Berlin, Germany
Felix Naumann	Humboldt University Berlin, Germany
Frank Olken	Lawrence Berkeley National Laboratory, USA
Jignesh Patel	University of Michigan, USA
Erhard Rahm	University of Leipzig, Germany
Julia Rice	IBM Life Sciences, USA
Peter Tarczy-Hornoch	University of Washington, USA
Limsoon Wong	Institute for Infocomm Research, Singapore
Aidong Zhang	State University of New York at Buffalo, USA

Additional Reviewers

Alexander Bilke	Humboldt University Berlin, Germany
Jens Bleiholder	Humboldt University Berlin, Germany
Hong-Hai Do	University of Leipzig, Germany
Antoon Goderis	University of Manchester, UK
Woo-Chang Hwang	State University of New York at Buffalo, USA
Daxin Jiang	State University of New York at Buffalo, USA
Toralf Kirsten	University of Leipzig, Germany
Peter Li	University of Newcastle, UK
Phillip Lord	University of Manchester, UK
Hervé Ménager	Arizona State University, USA
Peter Mork	University of Washington, USA
HweeHwa Pang	Institute for Infocomm Research, Singapore
Pengjun Pei	State University of New York at Buffalo, USA
Benjamin Prins	University of Bielefeld, Germany
Robert Stevens	University of Manchester, UK
Thoralf Töpel	University of Bielefeld, Germany
Silke Trißl	Humboldt University Berlin, Germany
Chris Wroe	University of Manchester, UK
Xian Xu	State University of New York at Buffalo, USA

Sponsors

Microsoft Research	research.microsoft.com
San Diego Supercomputer Center	www.sdsc.edu
American Medical Informatics Association (AMIA)	www.amia.org
UC Davis Genome Center	genomics.ucdavis.edu
U Maryland Institute for Advanced Computer Studies	www.umiacs.umd.edu
University of California, San Diego	www.ucsd.edu

Organization Committee

Amarnath Gupta	San Diego Supercomputer Center
Jon C. Meyer	San Diego Supercomputer Center
Donna Turner	San Diego Supercomputer Center
Linda Ferri	San Diego Supercomputer Center

Website

For more information on the workshop please visit the workshop website at www.sdsc.edu/dils05.

Table of Contents

Keynotes

Challenges in Biological Data Integration in the Post-genome Sequence Era
Shankar Subramaniam 1

Curated Databases
Peter Buneman 2

User Applications

A User-Centric Framework for Accessing Biological Sources and Tools
Sarah Cohen-Boulakia, Susan Davidson, Christine Froidevaux 3

BioLog: A Browser Based Collaboration and Resource Navigation Assistant for BioMedical Researchers
P. Singh, R. Bhimavarapu, H. Davulcu, C. Baral, S. Kim, H. Liu, M. Bittner, IV Ramakrishnan 19

Learning Layouts of Biological Datasets Semi-automatically
Kaushik Sinha, Xuan Zhang, Ruoming Jin, Gagan Agrawal 31

Ontologies

Factors Affecting Ontology Development in Ecology
C. Maria Keet 46

Querying Ontologies in Relational Database Systems
Silke Trißl, Ulf Leser 63

Scientific Names Are Ambiguous as Identifiers for Biological Taxa: Their Context and Definition Are Required for Accurate Data Integration
Jessie B. Kennedy, Robert Kukla, Trevor Paterson 80

The Multiple Roles of Ontologies in the BioMediator Data Integration System
Peter Mork, Ron Shaker, Peter Tarczy-Hornoch 96

Data Integration I–IV

Integrating Heterogeneous Microarray Data Sources Using Correlation Signatures
Jaewoo Kang, Jiong Yang, Wanhong Xu, Pankaj Chopra 105

Knowledge-Based Integrative Framework for Hypothesis Formation in Biochemical Networks
Nam Tran, Chitta Baral, Vinay J. Nagaraj, Lokesh Joshi 121

Semantic Correspondence in Federated Life Science Data Integration Systems
Malika Mahoui, Harshad Kulkarni, Nianhua Li, Zina Ben-Miled, Katy Börner 137

Assigning Unique Keys to Chemical Compounds for Data Integration: Some Interesting Counter Examples
Greeshma Neglur, Robert L. Grossman, Bing Liu 145

Integrating and Warehousing Liver Gene Expression Data and Related Biomedical Resources in GEDAW
Emilie Guérin, Gwenaelle Marquet, Anita Burgun, Olivier Loréal, Laure Berti-Equille, Ulf Leser, Fouzia Moussouni 158

Information Integration and Knowledge Acquisition from Semantically Heterogeneous Biological Data Sources
Doina Caragea, Jyotishman Pathak, Jie Bao, Adrian Silvescu, Carson Andorf, Drena Dobbs, Vasant Honavar 175

Cluster Based Integration of Heterogeneous Biological Databases Using the AutoMed Toolkit
Michael Maibaum, Lucas Zamboulis, Galia Rimon, Christine Orengo, Nigel Martin, Alexandra Poulovassilis 191

Hybrid Integration of Molecular-Biological Annotation Data
Toralf Kirsten, Hong-Hai Do, Christine Körner, Erhard Rahm 208

Setup and Annotation of Metabolomic Experiments by Integrating Biological and Mass Spectrometric Metadata
Oliver Fiehn, Gert Wohlgemuth, Martin Scholz 224

Performance-Oriented Privacy-Preserving Data Integration
Raymond K. Pon, Terence Critchlow 240

Potpourri

Building a Generic Platform for Medical Screening Applications Based on Domain Specific Modeling and Process Orientation
Stefan Jablonski, Rainer Lay, Sascha Müller, Christian Meiler, Matthias Faerber, Victor Derhartunian, Georg Michelson 257

Automatic Generation of Data Types for Classification of Deep Web Sources
Anne H.H. Ngu, David Buttler, Terence Critchlow 266

BioNavigation: Selecting Optimum Paths Through Biological Resources to Evaluate Ontological Navigational Queries
Zoé Lacroix, Kaushal Parekh, Maria-Esther Vidal, Marelis Cardenas, Natalia Marquez 275

Posters and Demonstrations

Support for BioIndexing in BLASTgres
Ruey-Lung Hsiao, D. Stott Parker, Hung-chih Yang 284

An Environment to Define and Execute *In-Silico* Workflows Using Web Services
Rafael Targino, Maria Claudia Cavalcanti, Marta Mattoso 288

Web Service Mining for Biological Pathway Discovery
George Zheng, Athman Bouguettaya 292

SemanticBio: Building Conceptual Scientific Workflows over Web Services
Zoé Lacroix, Hervé Ménager 296

PLATCOM: Current Status and Plan for the Next Stages
Kwangmin Choi, Jeong-Hyeon Choi, Amit Saple, Zhiping Wang, Jason Lee, Sun Kim 300

SOAP API for Integrating Biological Interaction Databases
Seong Joon Yoo, Min Kyung Kim, Ho Il Lee, Hyun Seok Park 305

Collaborative Curation of Data from Bio-medical Texts and Abstracts and Its Integration
Chitta Baral, Hasan Davulcu, Mutsumi Nakamura, Prabhdeep Singh, Luis Tari, Lian Yu 309

Towards an Ontology Based Visual Query System
Serguei Krivov, Ferdinando Villa 313

Invited Briefings

Data Integration in the Biomedical Informatics Research Network
Vadim Astakhov, Amarnath Gupta, Simone Santini, Jeffrey S. Grethe 317

Data Integration and Workflow Solutions for Ecology
William Michener, James Beach, Shawn Bowers, Laura Downey, Matthew Jones, Bertram Ludäscher, Deana Pennington, Arcot Rajasekar, Samantha Romanello, Mark Schildhauer, Dave Vieglais, Jianting Zhang 321

Eco-Informatics for Decision Makers Advancing a Research Agenda
Judith Bayard Cushing, Tyrone Wilson 325

An Architecture and Application for Integrating Curation Data at the Residue Level for Proteins
Mehmet M. Dalkilic 335

The Biozon System for Complex Analysis of Heterogeneous Interrelated Biological Data and Discovery of Emergent Structures
Aaron Birkland, Golan Yona 339

Author Index 343