Lecture Notes in Bioinformatics 4075

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


Subseries of Lecture Notes in Computer Science
Preface

Data management and data integration are fundamental problems in the life sciences. Advances in molecular biology and molecular medicine are almost universally underpinned by enormous efforts in data management, data integration, automatic data quality assurance, and computational data analysis. Many hot topics in the life sciences, such as systems biology, personalized medicine, and pharmacogenomics, critically depend on integrating data sets and applications produced by different experimental methods, in different research groups, and at different levels of granularity. Despite more than a decade of intensive research in these areas, there remain many unsolved problems. In some respects, these problems are becoming more severe, both due to continuous increases in data volumes and the growing diversity in types of data that need to be managed. And the next big challenge is already upon us: the need to integrate the different “omics” data sets with the vast amounts of clinical data, collected daily in thousands of hospitals and physicians’ offices all over the world.

DILS 2006 is the third in an annual 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 science. DILS 2004 in Leipzig and DILS 2005 in San Diego each attracted around 100 researchers from all over the world. This year the number of submitted papers again increased. The Program Committee selected 23 papers out of 50 strong full submissions. In an effort to include contributions that do not present a new method but that describe innovative and up-and-running practical systems, we distinguished “research papers” and “systems papers.” The seven systems papers can be found in the sections Systems I and Systems II. Among the research papers there are four short papers and 12 full papers.

In addition to the presented papers, DILS 2006 featured two invited talks by Victor M. Markowitz and James H. Kaufmann and a session with updates on projects of world-wide importance: the Taverna eScience project, the BioMoby integration framework, and the BioMart integrated genomics data warehouse. Finally, there was a lively poster session.

The workshop was held at the Wellcome Trust Conference Center on the campus of the European Bioinformatics Institute (EBI) in Hinxton, UK. It was kindly sponsored by Microsoft Research, who also made available their conference management system, IBM Research, metanomics, metanomicshealth, the EBI industry programme, and by Schering AG. We are grateful for the help of Springer in putting together and publishing these proceedings. As Program Co-chairs we thank all authors who submitted their work, and the Program Committee members for their careful (and timely) reviews.

We particularly thank Paul Kersey of the EBI, who served as Local Chair of the workshop, and thus did all the hard work.

June 2006

Ulf Leser
Felix Naumann
Barbara Eckman
# Organization

**DILS 2006 Co-chairs**

Ulf Leser  
Humboldt-Universität zu Berlin, Germany

Felix Naumann  
Humboldt-Universität zu Berlin, Germany

Barbara Eckman  
IBM Healthcare and Life Sciences, USA

**Local Chair**

Paul Kersey, European Bioinformatics Institute, Hinxton, UK

**Program Committee**

<table>
<thead>
<tr>
<th>Name</th>
<th>Organization/Institution</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emmanuel Barillot</td>
<td>Institut Curie</td>
<td>France</td>
</tr>
<tr>
<td>David Benton</td>
<td>GlaxoSmithKline</td>
<td>USA</td>
</tr>
<tr>
<td>Laure Berti-Equille</td>
<td>Universitaire de Beaulieu</td>
<td>France</td>
</tr>
<tr>
<td>Peter Bunemann</td>
<td>University of Edinburgh</td>
<td>UK</td>
</tr>
<tr>
<td>Terence Critchlow</td>
<td>Lawrence Livermore National Laboratory</td>
<td>USA</td>
</tr>
<tr>
<td>Jürgen Eils</td>
<td>Deutsches Krebsforschungszentrum DKFZ</td>
<td>Germany</td>
</tr>
<tr>
<td>Floris Geerts</td>
<td>University of Edinburgh and Limburgs Universitair Centrum</td>
<td>UK</td>
</tr>
<tr>
<td>Amarnath Gupta</td>
<td>San Diego Supercomputer Center</td>
<td>USA</td>
</tr>
<tr>
<td>Joachim Hammer</td>
<td>University of Florida</td>
<td>USA</td>
</tr>
<tr>
<td>Henning Hermjakob</td>
<td>European Bioinformatics Institute</td>
<td>UK</td>
</tr>
<tr>
<td>Mike Hogarth</td>
<td>UC Davis</td>
<td>USA</td>
</tr>
<tr>
<td>Stefan Jablonski</td>
<td>Univ. Erlangen-Nuernberg</td>
<td>Germany</td>
</tr>
<tr>
<td>H V Jagadish</td>
<td>University of Michigan</td>
<td>USA</td>
</tr>
<tr>
<td>Hasan Jamil</td>
<td>Wayne State University</td>
<td>USA</td>
</tr>
<tr>
<td>Jacob Köhler</td>
<td>Rothamsted Research</td>
<td>UK</td>
</tr>
<tr>
<td>Peter Karp</td>
<td>SRI International</td>
<td>USA</td>
</tr>
<tr>
<td>Vipul Kashyap</td>
<td>Partners HealthCare System</td>
<td>USA</td>
</tr>
<tr>
<td>Arek Kasprzyk</td>
<td>European Bioinformatics Institute</td>
<td>UK</td>
</tr>
<tr>
<td>Anthony Kosky</td>
<td>Axiope Inc</td>
<td>USA</td>
</tr>
<tr>
<td>Bertram Ludäscher</td>
<td>UC Davis</td>
<td>USA</td>
</tr>
<tr>
<td>Paula Matuszek</td>
<td>GlaxoSmithKline Beecham</td>
<td>USA</td>
</tr>
<tr>
<td>Peter Mork</td>
<td>The MITRE Corporation</td>
<td>USA</td>
</tr>
<tr>
<td>Jignesh Patel</td>
<td>University of Michigan</td>
<td>USA</td>
</tr>
<tr>
<td>Norman Paton</td>
<td>University of Manchester</td>
<td>UK</td>
</tr>
<tr>
<td>Christian Piepenbrock</td>
<td>Epigenomics AG</td>
<td>Germany</td>
</tr>
<tr>
<td>Erhard Rahm</td>
<td>Universität Leipzig</td>
<td>Germany</td>
</tr>
</tbody>
</table>
VIII Organization

Louiqa Raschid University of Maryland USA
Otto Ritter AstraZeneca USA
Monica Scannapieco University of Rome “La Sapienza” Italy
Dennis Paul Wall Harvard Medical School USA
Sharon Wang IBM Healthcare and Life Sciences USA
Bertram Weiss Schering AG Germany
Limsoon Wong Institute for Infocomm Research Singapore

Additional Reviewers

Shawn Bowers Andrew Jones Heiko Müller
Adriane Chapman Toralf Kirsten Eugene Novikov
Hon Nian Chua Judice Koh Loic Royer
Heiko Dietze Jörg Lange Donny Soh
Andreas Doms Mario Latendresse Silke Trissl
Nan Guo Christian Lawerenz Thomas Wächter
Michael Hartung Timothy M. McPhillips

Sponsoring Institutions

Microsoft Research http://research.microsoft.com/
metanomics http://www.metanomics.de/
metanomicshealth http://www.metanomics-health.de/
IBM Research http://www.research.ibm.com/
EBI Industry Program http://industry.ebi.ac.uk/
Schering http://www.schering.de/

Website

For more information please visit the DILS 2006 website at http://www.informatik.hu-berlin.de/dils2006/.
<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>An Application Driven Perspective on Biological Data Integration</td>
<td>Victor M. Markowitz</td>
<td>1</td>
</tr>
<tr>
<td>Towards a National Healthcare Information Infrastructure</td>
<td>Sarah Knoop</td>
<td>2</td>
</tr>
<tr>
<td>Data Access and Integration in the ISPIDER Proteomics Grid</td>
<td>Lucas Zamboulis, Hao Fan, Khalid Belhajjame, Jennifer Siepen, Andrew Jones, Nigel Martin, Alexandra Poulavassilis, Simon Hubbard, Suzanne M. Embury, Norman W. Paton</td>
<td>3</td>
</tr>
<tr>
<td>A Cell-Cycle Knowledge Integration Framework</td>
<td>Erick Antezana, Elena Tsiporkova, Vladimir Mironov, Martin Kuiper</td>
<td>19</td>
</tr>
<tr>
<td>Link Discovery in Graphs Derived from Biological Databases</td>
<td>Petteri Sevon, Lauri Eronen, Petteri Hintsanen, Kimmo Kulovesi, Hannu Toivonen</td>
<td>35</td>
</tr>
<tr>
<td>Towards an Automated Analysis of Biomedical Abstracts</td>
<td>Barbara Gawronska, Björn Erlendsson, Björn Olsson</td>
<td>50</td>
</tr>
<tr>
<td>Improving Text Mining with Controlled Natural Language: A Case Study for Protein Interactions</td>
<td>Tobias Kuhn, Loïc Royer, Norbert E. Fuchs, Michael Schröder</td>
<td>66</td>
</tr>
<tr>
<td>SNP-Converter: An Ontology-Based Solution to Reconcile Heterogeneous SNP Descriptions for Pharmacogenomic Studies</td>
<td>Adrien Coulet, Malika Smail-Tabbone, Pascale Benlian, Amedeo Napoli, Marie-Dominique Devignes</td>
<td>82</td>
</tr>
</tbody>
</table>
Systems I

SABIO-RK: Integration and Curation of Reaction Kinetics Data
Ulrike Wittig, Martin Golebiewski, Renate Kania, Olga Krebs, Saqib Mir, Andreas Weidemann, Stefanie Anstein, Jasmin Saric, Isabel Rojas .................................................. 94

SIBIOS Ontology: A Robust Package for the Integration and Pipelining of Bioinformatics Services
Malika Mahoui, Zina Ben Miled, Sriram Srinivasan, Mindi Dippold, Bing Yang, Li Nianhua .......................... 104

Data Structures for Genome Annotation, Alternative Splicing, and Validation
Sven Mielordt, Ivo Grosse, Jürgen Kleffe .................................. 114

BioFuice: Mapping-Based Data Integration in Bioinformatics
Toralf Kirsten, Erhard Rahm ........................................ 124

Potpourri

A Method for Similarity-Based Grouping of Biological Data
Vaida Jakonienė, David Rundqvist, Patrick Lambrix ......................... 136

On Querying OBO Ontologies Using a DAG Pattern Query Language
Amarnath Gupta, Simone Santini ........................................ 152

Using Term Lists and Inverted Files to Improve Search Speed for Metabolic Pathway Databases
Greeshma Neglur, Robert L. Grossman, Natalia Maltsev, Clement Yu ................................................... 168

Systems II

Arevir: A Secure Platform for Designing Personalized Antiretroviral Therapies Against HIV
Kirsten Roomp, Niko Beerenwinkel, Tobias Sing, Eugen Schülter, Joachim Büch, Saleta Sierra-Aragon, Martin Däumer, Daniel Hoffmann, Rolf Kaiser, Thomas Lengauer, Joachim Selbig .... 185

The Distributed Annotation System for Integration of Biological Data
Andreas Prlić, Ewan Birney, Tony Cox, Thomas A. Down, Rob Finn, Stefan Gräf, David Jackson, Andreas Kühäri, Eugene Kulesha, Roger Pettett, James Smith, Jim Stalker, Tim J.P. Hubbard .................................................. 195
An Information Management System for Collaboration Within Distributed Working Environment
Maria Samsonova, Andrei Pisarev, Konstantin Kozlov, Ekaterina Poustelnikova, Arthur Tkachenko ....................... 204

Short Papers

Ontology Analysis on Complexity and Evolution Based on Conceptual Model
Zhe Yang, Dalu Zhang, Chuan Ye ........................................ 216

Distributed Execution of Workflows in the INB
Ismael Navas-Delgado, Antonio J. Pérez, Jose F. Aldana-Montes,
Oswaldo Trelles .......................................................... 224

Knowledge Networks of Biological and Medical Data: An Exhaustive and Flexible Solution to Model Life Science Domains
Sascha Losko, Karsten Wenger, Wenzel Kalus, Andrea Ramge,
Jens Wiehler, Klaus Heumann ........................................... 232

On Characterising and Identifying Mismatches in Scientific Workflows
Khalid Belhajjame, Suzanne M. Embury, Norman W. Paton ........ 240

Workflow

Collection-Oriented Scientific Workflows for Integrating and Analyzing Biological Data
Timothy McPhillips, Shawn Bowers, Bertram Ludäscher ........... 248

Towards a Model of Provenance and User Views in Scientific Workflows
Shirley Cohen, Sarah Cohen-Boulakia, Susan Davidson ............. 264

An Extensible Light-Weight XML-Based Monitoring System for Sequence Databases
Dieter Van de Craen, Frank Neven, Kerstin Koch ..................... 280

Author Index ................................................................. 297