

Lecture Notes in Bioinformatics

7348

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

Olivier Bodenreider Bastien Rance (Eds.)

Data Integration in the Life Sciences

8th International Conference, DILS 2012
College Park, MD, USA, June 28-29, 2012
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

Olivier Bodenreider

Bastien Rance

National Institutes of Health, US National Library of Medicine

8600 Rockville Pike, Bethesda, MD 20894, USA

E-mail: {olivier.bodenreider, bastien.rance}@nih.gov

ISSN 0302-9743

e-ISSN 1611-3349

ISBN 978-3-642-31039-3

e-ISBN 978-3-642-31040-9

DOI 10.1007/978-3-642-31040-9

Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012939199

CR Subject Classification (1998): H.3, J.3, I.2, H.4, H.2, H.5

LNCS Sublibrary: SL 8 – Bioinformatics

© Springer-Verlag Berlin Heidelberg 2012

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.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

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

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume of *Lecture Notes in Bioinformatics* (LNBI) contains selected papers from the 8th International Conference on Data Integration in the Life Sciences (DILS 2012), held June 28–29, 2012 at the University of Maryland in College Park, Maryland, USA (<http://sites.google.com/site/webdils2012/>).

The Data Integration in the Life Sciences (DILS) conference has been held regularly since 2004, alternating between venues in North America and Europe. Over the years, DILS has become a forum for life science researchers, a place where issues in data integration are discussed, where new avenues are explored, and where integration is extended to new domains. Through a mix of invited keynote presentations, oral presentations of peer-reviewed papers, posters and demos, a variety of ideas are discussed, ranging from reports on mature research and established systems, to exciting new prototypes and ongoing research.

This year the conference was organized around three major themes. Each session was introduced by a keynote presentation followed by paper presentations. In the “Foundations of Data Integration,” Jim Ostell from the National Center for Biotechnology Information (NCBI) presented the Entrez system. Recurring themes from the papers included ontologies, semantic similarity, mapping between ontologies and schema matching. The second theme, “New Paradigms for Data Integration,” was introduced by a presentation of the Watson system by Ken Barker from IBM Research. The papers demonstrated the benefits of Semantic Web technologies for integrating biological data and explored crowd-sourcing as a potential resource to support data integration. Finally, DILS 2012 emphasized “Integrating Clinical Data.” Jim Cimino from the National Institutes of Health (NIH) Clinical Center presented “BTRIS,” the clinical data warehouse supporting translational research at NIH. The papers explored the integration of clinical data for cancer research and the contribution of natural language processing to the integration of unstructured clinical data.

We thank the Program Committee for thoroughly reviewing and helping to select the manuscripts submitted to the conference. Our thanks also go to Louiqa Raschid, who coordinated the logistics at the University of Maryland.

June 2012

Olivier Bodenreider
Bastien Rance

Organization

DILS Steering Committee

| | |
|----------------------|---|
| Sarah Cohen-Boulakia | LRI, University of Paris-Sud 11, France |
| Graham Kemp | Chalmers University of Technology, Sweden |
| Patrick Lambrix | Linköping University, Sweden |
| Ulf Leser | Humboldt-Universität zu Berlin, Germany |
| Bertram Ludaescher | University of California, USA |
| Paolo Missier | Newcastle University, UK |
| Norman Paton | University of Manchester, UK |
| Louiqa Raschid | University of Maryland, USA |
| Erhard Rahm | University of Leipzig, Germany |

General Chairs

| | |
|---------------------|--|
| Olivier Bodenreider | National Library of Medicine, NIH, USA |
| Louiqa Raschid | University of Maryland, USA |

Webmaster

| | |
|---------------|--|
| Bastien Rance | National Library of Medicine, NIH, USA |
|---------------|--|

Local Organizers (USA)

| | |
|----------------|-----------------------------|
| Louiqa Raschid | University of Maryland, USA |
|----------------|-----------------------------|

Program Committee

| | |
|----------------------|---|
| Christopher Baker | UNB Saint John Faculty, Canada |
| Elmer Bernstam | University of Texas, USA |
| Judy Blake | The Jackson Laboratory, USA |
| Anita Burgun | Université de Rennes 1, France |
| Jim Cimino | National Library of Medicine, NIH, USA |
| Sarah Cohen-Boulakia | LRI, University of Paris-Sud 11, France |
| David De Roure | Oxford e-Research Center, UK |
| Dina Demner-Fushman | National Library of Medicine, NIH, USA |
| Michel Dumontier | Carleton University, Canada |
| Christine Froidevaux | LRI, University of Paris-Sud 11, France |
| Carole Goble | University of Manchester, UK |
| Graciela Gonzalez | Arizona State University, USA |
| Vasant Honavar | Iowa State University, USA |

| | |
|----------------------------|---|
| Tony Xiaohua Hu | Drexel University, USA |
| Hasan Jamil | Wayne State University, USA |
| Graham Kemp | Chalmers University of Technology, Sweden |
| Purvesh Khatri | Stanford University, USA |
| Patrick Lambrix | Linköping University, Sweden |
| Adam Lee | National Library of Medicine, NIH, USA |
| Mong Li Lee | National University of Singapore, Singapore |
| Ulf Leser | Humboldt-Universität zu Berlin, Germany |
| Frédérique Lisacek | Swiss Institute of Bioinformatics, Switzerland |
| Bertram Ludaescher | University of California, USA |
| Yves Lussier | University of Chicago, USA |
| Brad Malin | Vanderbilt University, USA |
| M. Scott Marshall | University of Amsterdam, The Netherlands |
| Marco Masseroli | Politecnico di Milano, Italy |
| Paolo Missier | Newcastle University, UK |
| Peter Mork | MITRE, USA |
| Fleur Mougin | University of Bordeaux 2, France |
| Shawn Murphy | Partners HealthCare, Boston, USA |
| Radhakrishnan Nagaraja | University of Arkansas for Medical Sciences, USA |
| Jyotishman Pathak | Mayo Clinic College of Medicine, USA |
| Norman Paton | University of Manchester, UK |
| Erhard Rahm | University of Leipzig, Germany |
| Bastien Rance | National Library of Medicine, NIH, USA |
| Dietrich Rebholz-Schuhmann | EBI, UK |
| Alan Ruttenberg | University at Buffalo, USA |
| Satya Sahoo | Case Western Reserve University, USA |
| Neil Sarkar | University of Vermont, USA |
| Guergana Savova | Children's Hospital, Boston, USA |
| Michael Schroeder | TU Dresden, Germany |
| Nigam Shah | Stanford University, USA |
| Amit Sheth | Wright State University, USA |
| Andrea Splendiani | Rothamsted Research, UK |
| Karin Verspoor | National ICT, Australia |
| Maria Esther Vidal | Universidad Simón Bolívar, Venezuela |
| Chris Welty | IBM, USA |
| Guo-Qiang Zhang | Case Western Reserve University, USA |

Additional Referees

| | |
|--------------------|--|
| Artjom Klein | UNB Saint John Faculty, Canada |
| Thomas Wächter | TU Dresden, Germany |
| Robert Leaman | Arizona State University, USA |
| Daniel Eisinger | TU Dresden, Germany |
| Anne Morgat | Swiss Institute of Bioinformatics, Switzerland |
| Alexandre Riazanov | UNB Saint John Faculty, Canada |

Table of Contents

Foundations of Data Integration

| | |
|---|----|
| Entrez: The NCBI Search and Discovery Engine | 1 |
| <i>James M. Ostell</i> | |
| InterOnto – Ranking Inter-Ontology Links | 5 |
| <i>Silke Trißl, Philipp Hussels, and Ulf Leser</i> | |
| Finding Cross Genome Patterns in Annotation Graphs | 21 |
| <i>Joseph Benik, Caren Chang, Louiqa Raschid, Maria-Esther Vidal, Guillermo Palma, and Andreas Thor</i> | |
| Pay-as-You-Go Ranking of Schema Mappings Using Query Logs | 37 |
| <i>Ruhaila Maskat, Norman W. Paton, and Suzanne M. Embury</i> | |

New Paradigms for Data Integration

| | |
|--|----|
| Combining Structured and Unstructured Knowledge Sources for Question Answering in Watson | 53 |
| <i>Ken Barker</i> | |
| Cancer Data Integration and Querying with GeneTegra | 56 |
| <i>E. Patrick Shironoshita, Yves R. Jean-Mary, Ray M. Bradley, Patricia Buendia, and Mansur R. Kabuka</i> | |
| Integrating Large, Disparate Biomedical Ontologies to Boost Organ Development Network Connectivity | 71 |
| <i>Chimezie Ogbuji and Rong Xu</i> | |
| Validating Candidate Gene-Mutation Relations in MEDLINE Abstracts via Crowdsourcing | 83 |
| <i>John D. Burger, Emily Doughty, Sam Bayer, David Tresner-Kirsch, Ben Wellner, John Aberdeen, Kyungjoon Lee, Maricel G. Kann, and Lynette Hirschman</i> | |

Integrating Clinical Data

| | |
|---|----|
| The Biomedical Translational Research Information System: Clinical Data Integration at the National Institutes of Health | 92 |
| <i>James J. Cimino</i> | |

| | |
|--|-----|
| ONCO-i2b2: Improve Patients Selection through Case-Based Information Retrieval Techniques | 93 |
| <i>Daniele Segagni, Matteo Gabetta, Valentina Tibollo, Alberto Zambelli, Silvia G. Priori, and Riccardo Bellazzi</i> | |
| Syntactic-Semantic Frames for Clinical Cohort Identification Queries . . . | 100 |
| <i>Dina Demner-Fushman and Swapna Abhyankar</i> | |
| Author Index | 113 |