

# Lecture Notes in Bioinformatics

6004

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

Christian Blaschke Hagit Shatkay (Eds.)

# Linking Literature, Information, and Knowledge for Biology

Workshop of the BioLink Special Interest Group,  
ISMB/ECCB 2009  
Stockholm, June 28-29, 2009  
Revised Selected Papers

## 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

Christian Blaschke

Bioalma

C/Ronda de Poniente, 4, 2-C

28760 Tres Cantos, Madrid, Spain

E-mail: blaschke@bioalma.com

Hagit Shatkay

Computational Biology and Machine Learning Lab

School of Computing, Queen's University

Kingston, ON, K7L 3N6 Canada

E-mail: shatkay@cs.queensu.ca

Library of Congress Control Number: 2010927026

CR Subject Classification (1998): J.3, I.2, F.1, I.4, I.5, I.2.7

LNCS Sublibrary: SL 8 – Bioinformatics

ISSN 0302-9743

ISBN-10 3-642-13130-1 Springer Berlin Heidelberg New York

ISBN-13 978-3-642-13130-1 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.com

© Springer-Verlag Berlin Heidelberg 2010

Printed in Germany

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

Printed on acid-free paper 06/3180

# Preface

This volume of *Lecture Notes in Bioinformatics (LNBI)* contains selected papers from the workshop of the BioLINK Special Interest Group (SIG) on Linking Literature, Information and Knowledge for Biology. The workshop was held June 28–29, 2009 at the ISMB/ECCB 2009 conference in Stockholm (<http://www.iscb.org/ismbecb2009>).

The BioLINK SIG meeting has been regularly held in association with the ISMB conferences since 2001, focusing on the development and application of resources and tools for biomedical text mining. The SIG is interdisciplinary in nature, and brings together researchers applying natural language processing, text mining, information extraction and retrieval in the biomedical domain, with scientists from bioinformatics and biology. This year the SIG included two special sessions, one on the analysis of images and figures, which play a critical role in scientific papers, and the other on the future of scientific publication. Overview papers concerned with both these topics are included in this volume.

The meeting featured invited talks, presentations of peer-reviewed contributed papers, reports from recent evaluations and workshops, as well as a poster session. The SIG solicited papers that discuss current biological image- and text-related needs, the challenges of meeting them, and the tools that may help address them. Specific topics include: biological image analysis and its relationship to biomedical text mining, application of text-mining tools to biomedical research, automated indexing of the biomedical literature, generation of structured digital abstracts, new evaluation measures to measure utility and usability of text mining tools integrated into the biologist's workflow, as well as new tools and applications, and future directions in biomedical text mining. The publications in this volume were selected from the workshop papers and represent extended versions that have undergone an additional review process by the members of the Program Committee.

We thank the Program Committee for thoroughly reviewing the submitted manuscripts and for their valuable comments that significantly improved the final publications.

December 2009

Christian Blaschke  
Hagit Shatkay

# Organization

## Workshop Program Chairs

|                    |                                                    |
|--------------------|----------------------------------------------------|
| Christian Blaschke | Bioalma, Spain                                     |
| Lynette Hirschman  | MITRE Corporation, USA                             |
| Scott Markel       | Accelrys, USA                                      |
| Robert Murphy      | Carnegie Mellon University, USA                    |
| Hagit Shatkay      | School of Computing, Queen's University,<br>Canada |
| Alfonso Valencia   | Spanish National Cancer Research Centre,<br>Spain  |

## Program Committee

|                         |                                                        |
|-------------------------|--------------------------------------------------------|
| Alan Aronson            | NLM, USA                                               |
| Olivier Bodenreider     | NLM, USA                                               |
| Monica Chagoyen         | Spanish National Centre for Biotechnology,<br>Spain    |
| Aaron Cohen             | Oregon Health Sciences University, USA                 |
| Kevin B. Cohen          | University of Colorado, Health Sciences Center,<br>USA |
| Mark Craven             | University of Wisconsin, Madison, USA                  |
| Bob Futrelle            | Northeastern University, USA                           |
| William Hayes           | Biogen IDEC, USA                                       |
| Martin Krallinger       | Spanish National Cancer Research Centre,<br>Spain      |
| Alberto Pascual Montano | Complutense University of Madrid, Spain                |
| Andreas Rechsteiner     | University of California, Santa Cruz, USA              |
| Luis Rocha              | Indiana University, Bloomington, USA                   |
| Karin Verspoor          | Los Alamos National Laboratories, USA                  |
| John Wilbur             | NCBI, NLM, USA                                         |
| Hong Yu                 | University of Wisconsin, Milwaukee, USA                |
| Ralf Zimmer             | Ludwig-Maximilians University, Munich,<br>Germany      |

## Sponsoring Institutions

We gratefully acknowledge:

The International Society for Computational Biology (ISCB)

The Spanish National Cancer Research Centre

BioCreative II.5 challenge

# Table of Contents

## BioLINK 2009

|                                                                                                                                                      |    |
|------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Overview of the Ninth Annual Meeting of the BioLINK SIG at ISMB:<br>Linking Literature, Information and Knowledge for Biology .....                  | 1  |
| <i>Christian Blaschke, Lynette Hirschman, Hagit Shatkay, and<br/>Alfonso Valencia</i>                                                                |    |
| Principles of Bioimage Informatics: Focus on Machine Learning of Cell<br>Patterns .....                                                              | 8  |
| <i>Luis Pedro Coelho, Estelle Glory-Afshar, Joshua Kangas,<br/>Shannon Quinn, Aabid Shariff, and Robert F. Murphy</i>                                |    |
| Summary of the BioLINK Special Interest Group Session on the Future<br>of Scientific Publishing .....                                                | 19 |
| <i>Scott Markel</i>                                                                                                                                  |    |
| Structured Literature Image Finder: Extracting Information from Text<br>and Images in Biomedical Literature .....                                    | 23 |
| <i>Luis Pedro Coelho, Amr Ahmed, Andrew Arnold, Joshua Kangas,<br/>Abdul-Saboor Sheikh, Eric P. Xing, William W. Cohen, and<br/>Robert F. Murphy</i> |    |
| Toward Computer-Assisted Text Curation: Classification Is Easy<br>(Choosing Training Data Can Be Hard...) .....                                      | 33 |
| <i>Robert Denroche, Ramana Madupu, Shibu Yooseph,<br/>Granger Sutton, and Hagit Shatkay</i>                                                          |    |
| Mining Protein-Protein Interactions from GeneRIFs with<br>OpenDMPAP .....                                                                            | 43 |
| <i>Andrew D. Fox, William A. Baumgartner Jr., Helen L. Johnson,<br/>Lawrence E. Hunter, and Donna K. Slonim</i>                                      |    |
| Combining Semantic Relations and DNA Microarray Data for Novel<br>Hypotheses Generation .....                                                        | 53 |
| <i>Dimitar Hristovski, Andrej Kastrin, Borut Peterlin, and<br/>Thomas C. Rindflesch</i>                                                              |    |
| Learning from Positive and Unlabeled Documents for Retrieval of<br>Bacterial Protein-Protein Interaction Literature .....                            | 62 |
| <i>Hongfang Liu, Manabu Torii, Guixian Xu, Zhangzhi Hu, and<br/>Johannes Goll</i>                                                                    |    |

|                                                                                                                  |           |
|------------------------------------------------------------------------------------------------------------------|-----------|
| Extracting and Normalizing Gene/Protein Mentions with the Flexible<br>and Trainable Moara Java Library . . . . . | 71        |
| <i>Mariana L. Neves, José Maria Carazo, and Alberto Pascual-Montano</i>                                          |           |
| <b>Author Index . . . . .</b>                                                                                    | <b>81</b> |