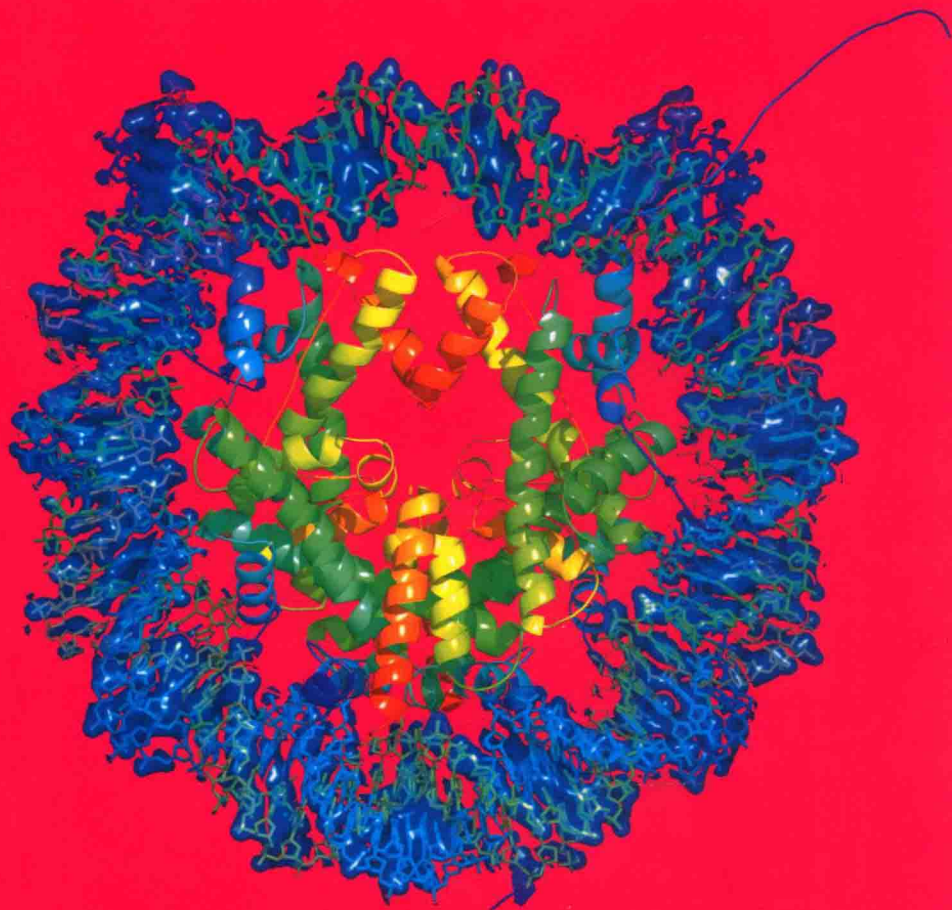


PACIFIC SYMPOSIUM ON BIOCOMPUTING 2007



Edited by

**Russ B. Altman, A. Keith Dunker,
Lawrence Hunter, Tiffany Murray & Teri E. Klein**

PACIFIC SYMPOSIUM ON BIOCOMPUTING 2007

Maui, Hawaii
3-7 January 2007

Edited by

Russ B. Altman

Stanford University, USA

A. Keith Dunker

Indiana University, USA

Lawrence Hunter

University of Colorado Health Sciences Center, USA

Tiffany Murray

Stanford University, USA

Teri E. Klein

Stanford University, USA

 **World Scientific**

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

BIOCOMPUTING 2007

Proceedings of the Pacific Symposium

Copyright © 2007 by World Scientific Publishing Co. Pte. Ltd.

All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the Publisher.

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. In this case permission to photocopy is not required from the publisher.

ISBN 981-270-417-5

Printed in Singapore by Mainland Press

PACIFIC SYMPOSIUM ON

BIOCOMPUTING 2007

PACIFIC SYMPOSIUM ON BIOCOMPUTING 2007

Biomedical computing has become a key component in the biomedical research infrastructure. In 2004 and 2005, the U.S. National Institutes of Health established seven National Centers for Biomedical Computation, focusing on a wide range of application areas and enabling technologies, including simulation, systems biology, clinical genomics, imaging, ontologies and others (see <http://www.bisti.nih.gov/ncbc/>). The goal of these centers is to help seed an information infrastructure to support biomedical research. The Pacific Symposium on Biocomputing (PSB) presented critical early sessions in most of the areas covered by these National Centers, and we are proud to continue the tradition of helping to define new areas of focus within biomedical computation.

Once again, we are fortunate to host two outstanding keynote speakers. Dr. Elizabeth Blackburn, Professor of Biology and Physiology in the Department of Biochemistry and Biophysics at the University of California, San Francisco will speak on “Interactions among telomeres, telomerase, and signaling pathways.” Her work has led our understanding of overall organization and control of chromosomal dynamics. Our keynote speaker in the area of Ethical, Legal and Social implications of technology will be Marc Rotenberg, Executive Director of the Electronic Privacy Information Center (EPIC) in Washington, D.C. He will speak on “Data mining and privacy: the role of public policy.” Many biomedical computation professionals have had and continue to grapple with privacy issues as interest in mining human genotype-phenotype data collections has increased.

PSB has a history of providing early sessions focusing on hot new areas in biomedical computation. These sessions are often conceived during the previous PSB meeting, as trends and new results are pondered and discussed. Very often, new sessions are lead by new faculty members trying to define a scientific niche and bring together leaders in the emerging areas. We are proud that many areas in biocomputing received their first significant focused attention at PSB. If you have an idea for a new session, we the organizers, are available to talk with you, either at the meeting or later by e-mail.

Again, the diligence and efforts of a dedicated group of researchers has led to an outstanding set of sessions, with associated introductory tutorials. These organizers provide the scientific core of PSB, and their sessions are as follows:

Indra Neil Sarkar

Biodiversity Informatics: Managing Knowledge Beyond Humans and Model Organisms

Bobbie-Jo Webb-Robertson & Bill Cannon

Computational Proteomics: High-throughput Analysis for Systems Biology

Martha Bulyk, Ernest Fraenkel, Alexander Hartemink, & Gary Stormo

DNA-Protein Interactions and Gene Regulation: Integrating Structure, Sequence and Function

Russ Greiner & David Wishart

Computational Approaches to Metabolomics

Pierre Zweigenbaum, Dina Demner-Fushman, Kevin Bretonnel Cohen, & Hong Yu

New Frontiers in Biomedical Text Mining

Maricel Kann, Yanay Ofran, Marco Punta, & Predrag Radivojac

Protein Interactions in Disease

In addition to the sessions and survey tutorials, this year's program includes two in depth tutorials. The presenters and titles of these tutorials are:

Giselle M. Knudsen, Reza A. Ghiladi, & D. Rey Banatao

Integration Between Experimental and Computational Biology for Studying Protein Function

Michael A Province & Ingrid B Borecki

Searching for the Mountains of the Moon: Genome Wide Association Studies of Complex Traits

We thank the Department of Energy and the National Institutes of Health for their continuing support of this meeting. Their support provides travel grants to many of the participants. Applied Biosystems and the International Society for Computational Biology continue to sponsor PSB, and as a result, we are able to provide additional travel grants to many meeting participants.

We would like to acknowledge the many busy researchers who reviewed the submitted manuscripts on a very tight schedule. The partial list following this preface does not include many who wished to remain anonymous, and of course we apologize to any who may have been left out by mistake.

Aloha!

Russ B. Altman

Departments of Genetics & Bioengineering, Stanford University

A. Keith Dunker

Department of Biochemistry and Molecular Biology, Indiana University School of Medicine

Lawrence Hunter

Department of Pharmacology, University of Colorado Health Sciences Center

Teri E. Klein

Department of Genetics, Stanford University

Pacific Symposium on Biocomputing Co-Chairs
September 28, 2006

Thanks to the reviewers...

Finally, we wish to thank the scores of reviewers. PSB requires that every paper in this volume be reviewed by at least three independent referees. Since there is a large volume of submitted papers, paper reviews require a great deal of work from many people. We are grateful to all of you listed below and to anyone whose name we may have accidentally omitted or who wished to remain anonymous.

Joshua Adkins	Vlado Dancik	Lynette Hirschman
Eugene Agichtein	Rina Das	Terence Hwa
Gelio Alves	Tjil De Bie	Sven Hyberts
Sophia Ananiadou	Dina Demner-	Lilia Iakoucheva
Alan Aronson	Fushman	Navdeep Jaitly
Ken Baclawski	Rob DeSalle	Helen Jenkins
Joel Bader	Luis DeSilva	Kent Johnson
Breck Baldwin	Diego Di Bernardo	Andrew Joyce
Ziv Bar-Joseph	Chuong Do	James Kadin
Serafim Batzoglou	Michel Dumontier	Martin R. Kalfatovic
Asa Ben-Hur	Mary G. Egan	Manpreet S. Katari
Sabine Bergler	Roman Eisner	Sun Kim
Olivier Bodenreider	Emilio Espisito	Oliver King
Alvis Brazma	Mark Fasnacht	Tanja Kortemme
Kevin Bretonnel	Oliver Fiehn	Harri Lahdesmaki
Yana Bromberg	Alessandro Flammini	Ney Lemke
Harmen Bussemaker	Fabian Fontaine	Gondy Leroy
Andrea Califano	Lynne Fox	Christina Leslie
Bob Carpenter	Ari Frank	Li Liao
Michele Cascella	Kristofer Franzen	John C. Lindon
Saikat Chakrabarti	Tema Fridman	Chunmei Liu
Shih-Fu Chang	Carol Friedman	Yves Lussier
Pierre Chaurand	Robert Futrelle	Hongwu Ma
Ting Chen	Feng Gao	Kenzie MacIsaac
Hsinchun Chen	Adam Godzik	Tom Madej
Nawei Chen	Roy Goodacre	Ana Maguitman
Praveen Cherukuri	Michael Grusak	Askenazi Manor
Wei Chu	Melissa A. Haendel	Costas Maranas
James Cimino	Henk Harkema	Leonardo Marino
Aaron Cohen	Marti Hearst	John Markley
Nigel Collier	P. Bryan Heidorn	Pedro Mendes
Matteo Dal Peraro	Bill Hersh	Ivana Mihalek

Leonid Mirny	Santiago Schnell	John Wilbur
Joyce Mitchell	Rob Schumaker	Kazimierz O.
Matthew Monroe	Robert D. Sedgewick	Wrzeszczynski
Sean Mooney	Eran Segal	Dong Xu
Rafael Najmanovich	Kia Sepassi	Yoshihiro Yamanishi
Preslav Nakov	Anuj Shah	Yuzhen Ye
Leelavati Narlikar	Paul Shapshak	Hong Yu
Adeline Nazarenko	Hagit Shatkay	Peng Yue
Jack Newton	Mark Siddall	Pierre Zweigenbaum
William Noble	Mona Singh	
Christopher Oehmen	Mudita Singhal	
Christopher Oldfield	Saurabh Sinha	
Zoltan Oltvai	Thereza Amelia	
Matej Oresic	Soares	
Bernhard Palsson	Bruno Sobral	
Chrysanthi	Ray Sommorjai	
Paranavitana	Orkun Soyer	
Matteo Pellegrini	Irina Spasic	
Aloysius Phillips	Padmini Srinivasan	
Paul J. Planet	Paul Stothard	
Christian Posse	Eric Strittmatter	
Natasa Przulj	Shamil Sunyaev	
Teresa Przytycka	Silpa Suthram	
Bin Qian	Lorrie Tanabe	
Weijun Qian	Haixu Tang	
Arun Ramani	Igor Tetko	
Kathryn Rankin	Jun'ichi Tsujii	
Andreas Rechtsteiner	Peter Uetz	
Haluk Resat	Vladimir Uversky	
Tom Rindflesch	Vladimir Vacic	
Martin Ringwald	Alfonso Valencia	
Elizabeth Rogers	Karin Verspoor	
Pedro Romero	Mark Viant	
Graciela Rosemblat	K. Vijay-Shanker	
Andrea Rossi	Hans Vogel	
Erik Rytting	Slobodan Vucetic	
Jasmin Saric	Alessandro Vullo	
Indra Neil Sarkar	Wyeth Wasserman	
Yutaka Sasaki	Bonnie Webber	
Tetsuya Sato	Aalim Weljie	

CONTENTS

Preface	v
PROTEIN INTERACTIONS AND DISEASE	
Session Introduction	1
<i>Maricel Kann, Yanay Ofra, Marco Punta, and Predrag Radivojac</i>	
Graph Kernels for Disease Outcome Prediction from Protein-Protein Interaction Networks	4
<i>Karsten M. Borgwardt, Hans-Peter Kriegel, S.V.N. Vishwanathan, and Nicol N. Schraudolph</i>	
Chalkboard: Ontology-Based Pathway Modeling and Qualitative Inference of Disease Mechanisms	16
<i>Daniel L. Cook, Jesse C. Wiley, and John H. Gennari</i>	
Mining Gene-Disease Relationships from Biomedical Literature Weighting Protein-Protein Interactions and Connectivity Measures	28
<i>Graciela Gonzalez, Juan C. Uribe, Luis Tari, Colleen Brophy, and Chitta Baral</i>	
Predicting Structure and Dynamics of Loosely-Ordered Protein Complexes: Influenza Hemagglutinin Fusion Peptide	40
<i>Peter M. Kasson and Vijay S. Pande</i>	
Protein Interactions and Disease Phenotypes in the ABC Transporter Superfamily	51
<i>Libusha Kelly, Rachel Karchin, and Andrej Sali</i>	
LTHREADER: Prediction of Ligand-Receptor Interactions Using Localized Threading	64
<i>Vinay Pulim, Jadwiga Bienkowska, and Bonnie Berger</i>	
Discovery of Protein Interaction Networks Shared by Diseases	76
<i>Lee Sam, Yang Liu, Jianrong Li, Carol Friedman, and Yves A. Lussier</i>	

An Iterative Algorithm for Metabolic Network-Based Drug Target Identification	88
---	----

Padmavati Sridhar, Tamer Kahveci, and Sanjay Ranka

Transcriptional Interactions During Smallpox Infection and Identification of Early Infection Biomarkers	100
---	-----

Willy A. Valdivia-Granda, Maricel G. Kann, and Jose Malaga

COMPUTATIONAL APPROACHES TO METABOLOMICS

Session Introduction	112
----------------------	-----

David S. Wishart and Russell Greiner

Leveraging Latent Information in NMR Spectra for Robust Predictive Models	115
---	-----

David Chang, Aalim Weljie, and Jack Newton

Bioinformatics Data Profiling Tools: A Prelude to Metabolic Profiling	127
---	-----

Natarajan Ganesan, Bala Kalyanasundaram, and Mahe Velauthapllai

Comparative QSAR Analysis of Bacterial, Fungal, Plant and Human Metabolites	133
---	-----

Emre Karakoc, S. Cenk Sahinalp, and Artem Cherkasov

BioSpider: A Web Server for Automating Metabolome Annotations	145
---	-----

Craig Knox, Savita Shrivastava, Paul Stothard, Roman Eisner, and David S. Wishart

New Bioinformatics Resources for Metabolomics	157
---	-----

John L. Markley, Mark E. Anderson, Qiu Cui, Hamid R. Eghbalnia, Ian A. Lewis, Adrian D. Hegeman, Jing Li, Christopher F. Schulte, Michael R. Sussman, William M. Westler, Eldon L. Ulrich, and Zsolt Zolnai

Setup X — A Public Study Design Database for Metabolomic Projects	169
---	-----

Martin Scholz and Oliver Fiehn

Comparative Metabolomics of Breast Cancer	181
<i>Chen Yang, Adam D. Richardson, Jeffrey W. Smith, and Andrei Osterman</i>	

Metabolic Flux Profiling of Reaction Modules in Liver Drug Transformation	193
<i>Jeongah Yoon and Kyongbum Lee</i>	

NEW FRONTIERS IN BIOMEDICAL TEXT MINING

Session Introduction	205
<i>Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and K. Bretonnel Cohen</i>	

Extracting Semantic Predications from Medline Citations for Pharmacogenomics	209
<i>Caroline B. Ahlers, Marcelo Fiszman, Dina Demner-Fushman, François-Michel Lang, and Thomas C. Rindflesch</i>	

Annotating Genes Using Textual Patterns	221
<i>Ali Cakmak and Gultekin Ozsoyoglu</i>	

A Fault Model for Ontology Mapping, Alignment, and Linking Systems	233
<i>Helen L. Johnson, K. Bretonnel Cohen, and Lawrence Hunter</i>	

Integrating Natural Language Processing with Flybase Curation	245
<i>Nikiforos Karamanis Y, Ian Lewin, Ruth Seal, Rachel Drysdale, and Edward Briscoe</i>	

A Stacked Graphical Model for Associating Sub-Images with Sub-Captions	257
<i>Zhenzhen Kou, William W. Cohen, and Robert F. Murphy</i>	

GeneRIF Quality Assurance as Summary Revision	269
<i>Zhiyong Lu, K. Bretonnel Cohen, and Lawrence Hunter</i>	

Evaluating the Automatic Mapping of Human Gene and Protein Mentions to Unique Identifiers	281
<i>Alexander A. Morgan, Benjamin Wellner, Jeffrey B. Colombe, Robert Arens, Marc E. Colosimo, and Lynette Hirschman</i>	
Multiple Approaches to Fine-Grained Indexing of the Biomedical Literature	292
<i>Aurelie Neveol, Sonya E. Shooshan, Susanne M. Humphrey, Thomas C. Rindflesh, and Alan R. Aronson</i>	
Mining Patents Using Molecular Similarity Search	304
<i>James Rhodes, Stephen Boyer, Jeffrey Kreulen, Ying Chen, and Patricia Ordonez</i>	
Discovering Implicit Associations Between Genes and Hereditary Diseases	316
<i>Kazuhiro Seki and Javed Mostafa</i>	
A Cognitive Evaluation of Four Online Search Engines for Answering Definitional Questions Posed by Physicians	328
<i>Hong Yu and David Kaufman</i>	
BIODIVERSITY INFORMATICS: MANAGING KNOWLEDGE BEYOND HUMANS AND MODEL ORGANISMS	
Session Introduction	340
<i>Indra Neil Sarkar</i>	
Biomediator Data Integration and Inference for Functional Annotation of Anonymous Sequences	343
<i>Eithon Cadag, Brent Louie, Peter J. Myler, and Peter Tarczy-Hornoch</i>	
Absent Sequences: Nullomers and Primes	355
<i>Greg Hampikian and Tim Andersen</i>	

An Anatomical Ontology for Amphibians	367
<i>Anne M. Maglia, Jennifer L. Leopold, L. Analía Pugener, and Susan Gauch</i>	
Recommending Pathway Genes Using a Compendium of Clustering Solutions	379
<i>David M. Ng, Marcos H. Woehrmann, and Joshua M. Stuart</i>	
Semi-Automated XML Markup of Biosystematic Legacy Literature with the Goldengate Editor	391
<i>Guido Sautter, Klemens Böhm, and Donat Agosti</i>	
 COMPUTATIONAL PROTEOMICS: HIGH-THROUGHPUT ANALYSIS FOR SYSTEMS BIOLOGY	
Session Introduction	403
<i>William Cannon and Bobbie-Jo Webb-Robertson</i>	
Advancement in Protein Inference from Shotgun Proteomics Using Peptide Detectability	409
<i>Pedro Alves, Randy J. Arnold, Milos V. Novotny, Predrag Radivojac, James P. Reilly, and Haixu Tang</i>	
Mining Tandem Mass Spectral Data to Develop a More Accurate Mass Error Model for Peptide Identification	421
<i>Yan Fu, Wen Gao, Simin He, Ruixiang Sun, Hu Zhou, and Rong Zeng</i>	
Assessing and Combining Reliability of Protein Interaction Sources	433
<i>Sonia Leach, Aaron Gabow, Lawrence Hunter, and Debra S. Goldberg</i>	
Probabilistic Modeling of Systematic Errors in Two-Hybrid Experiments	445
<i>David Sontag, Rohit Singh, and Bonnie Berger</i>	

Prospective Exploration of Biochemical Tissue Composition via Imaging Mass Spectrometry Guided by Principal Component Analysis <i>Raf Van de Plas, Fabian Ojeda, Maarten Dewil, Ludo Van Den Bosch, Bart De Moor, and Etienne Waelkens</i>	458
--	-----

DNA-PROTEIN INTERACTIONS: INTEGRATING STRUCTURE, SEQUENCE, AND FUNCTION

Session Introduction <i>Martha L. Bulyk, Alexander J. Hartemink, Ernest Fraenkel, and Gary Stormo</i>	470
Discovering Motifs With Transcription Factor Domain Knowledge <i>Henry C.M. Leung, Francis Y.L. Chin, and Bethany M.Y. Chan</i>	472
<i>Ab initio</i> Prediction of Transcription Factor Binding Sites <i>L. Angela Liu and Joel S. Bader</i>	484
Comparative Pathway Annotation with Protein-DNA Interaction and Operon Information via Graph Tree Decomposition <i>Jizhen Zhao, Dongsheng Che, and Liming Cai</i>	496

PROTEIN INTERACTIONS AND DISEASE

MARICEL KANN

*National Center for Biotechnology Information, NIH
Bethesda, MD 20894, U.S.A.*

YANAY OFRAN

*Department of Biochemistry & Molecular Biophysics, Columbia University
New York, NY 10032, U.S.A.*

MARCO PUNTA

*Department of Biochemistry & Molecular Biophysics, Columbia University
New York, NY 10032, U.S.A.*

PREDRAG RADIVOJAC

*School of Informatics, Indiana University
Bloomington, IN 47408, U.S.A.*

In 2003, the US National Human Genome Research Institute (NHGRI) articulated grand challenges for the genomics community in which the translation of genome-based knowledge into disease understanding, diagnostics, prognostics, drug response and clinical therapy is one of the three fundamental directions (“genomics to biology,” “genomics to health” and “genomics to society”).¹ At the same time the National Institutes of Health (NIH) laid out a similar roadmap for biomedical sciences.² Both the NHGRI grand challenges and the NIH roadmap recognized bioinformatics as an integral part in the future of life sciences. While this recognition is gratifying for the bioinformatics community, its task now is to answer the challenge of making a direct impact to the medical science and benefiting human health. Innovative use of informatics in the “translation from bench to bedside” becomes a key for bioinformaticians.

In 2005, the Pacific Symposium on Biocomputing (PSB) first solicited papers related to one aspect of this challenge, protein interactions and disease, which directly addresses computational approaches in search for the molecular basis of disease. The goal of the session was to bring together scientists interested in both bioinformatics and medical sciences to present their research progress. The session generated great interest resulting in a number of high quality papers and testable hypothesis regarding the involvement of proteins in various disease pathways. This year, the papers accepted for the session on Protein Interactions and Disease at PSB 2007 follow the same trend.

The first group of papers explored structural aspects of protein-protein interactions. Kelly et al. study ABC transporter proteins which are involved in substrate transport through the membrane. By investigating intra-transporter domain interfaces they conclude that nucleotide-binding interfaces are more conserved than those of transmembrane domains. Disease-related mutations were mapped into these interfaces. Pulim et al. developed a novel threading algorithm that predicts interactions between receptors (membrane proteins) and ligands. The method was tested on cytokines, proteins implicated in intra-cellular communication and immune system response. Novel candidate interactions, which may be implicated in disease, were predicted. Kasson and Pande use molecular dynamics to address high-order molecular organization in cell membranes. A large number of molecular dynamics trajectories provided clues into structural aspects of the insertion of about 20-residue long fusion peptide into a cell membrane by a trimer hemagglutinin of the influenza virus. The authors explain effects of mutations that preserve peptide's monomeric structure but incur loss of viral infectivity.

The second group of studies focused on analysis of protein interaction networks. Sam et al. investigate molecular factors responsible for the diseases with different causes but similar phenotypes and postulate that some are related to breakdowns in the shared protein-protein interaction networks. A statistical method is proposed to identify protein networks shared by diseases. Sridhar et al. developed an efficient algorithm for perturbing metabolic networks in order to stop the production of target compounds, while minimizing unwanted effects. The algorithm is aimed at drug development where toxicity of the drug should be reduced. Borgwardt et al. were interested in predicting clinical outcome by combining microarray and protein-protein interaction data. They use graph kernels as a measure of similarity between graphs and develop methods to improve their scalability to large graphs. Support vector machines were used to predict disease outcome. Gonzalez et al. extracted a large number of gene-disease relationships by parsing literature and mapping them to the known protein-protein interaction networks. They propose a method for ranking proteins for their involvement in disease. The method was tested on atherosclerosis. Valdivia-Granda et al. devised a method to integrate protein-protein interaction data along with other genomic annotation features with microarray data. They applied it to microarray data from a study of non-human primates infected with variola and identified early infection biomarkers. The study was complemented with a comparative protein domain analysis between host and pathogen. This work contributes to the understanding of the mechanisms of infectivity, disease and suggests potential therapeutic targets. Finally, Cook et al. worked on the novel ontology of biochemical pathways. They present Chalkboard, a tool for