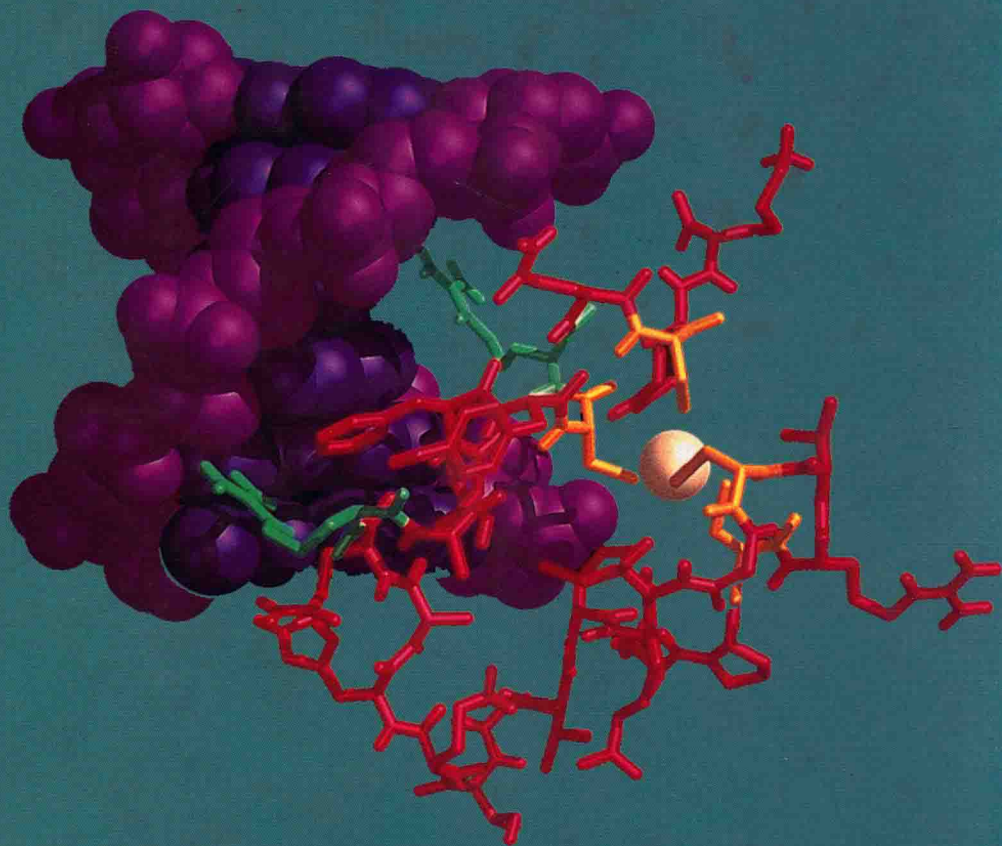


# PACIFIC SYMPOSIUM ON BIOCOMPUTING 2002



*Edited by*

**Russ B. Altman, A. Keith Dunker,  
Lawrence Hunter, Kevin Lauderdale & Teri E. Klein**

**World Scientific**

# PACIFIC SYMPOSIUM ON BIOCOMPUTING 2002

Kauai, Hawaii  
3-7 January 2002

*Edited by*

Russ B. Altman

Stanford University, USA

A. Keith Dunker

Washington State University, USA

Lawrence Hunter

University of Colorado Health Sciences Center, USA

Kevin Lauderdale

Stanford University, USA

Teri E. Klein

Stanford University, USA



**World Scientific**

New Jersey • London • Singapore • Hong Kong

*Published by*

World Scientific Publishing Co. Pte. Ltd.

P O Box 128, Farrer Road, Singapore 912805

*USA office:* Suite 1B, 1060 Main Street, River Edge, NJ 07661

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

21123047

**British Library Cataloguing-in-Publication Data**

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

**BIOCOMPUTING**

**Proceedings of the 2002 Pacific Symposium**

Copyright © 2001 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 981-02-4777-X

Printed in Singapore by World Scientific Printers

PACIFIC SYMPOSIUM ON  

---

BIOCOMPUTING 2002

## PACIFIC SYMPOSIUM ON BIOCOMPUTING 2002

The seventh Pacific Symposium on Biocomputing (PSB) marks the first PSB held following the tragic events of September 11, 2001 in New York, Pennsylvania and Washington DC. These events have affected the world at large and cannot go unnoticed by the computational biology community. The organizers would like to add their condolences to those who suffered. In spite of technical and personal difficulties that individuals incurred, we are happy able to put forth these proceedings.

PSB is sponsored by the International Society for Computational Biology (<http://www.iscb.org/>). Meeting participants benefit once again from travel grants from the U.S. Department of Energy, the National Library of Medicine/National Institutes of Health, Applied Biosystems and Boston College. We gratefully acknowledge the hardware contributions from Compaq.

We thank Professor David Botstein in advance for his plenary address on *Extracting Biologically Interesting Information from Microarrays* and Professor Rebecca Eisenberg for her plenary address on *Bioinformatics, Bioinformation and Biomolecules: the Role and Limitations of Patents*. Kevin Lauderdale has gone beyond the call of duty and once again expertly created the printed and online proceedings. Al Conde has ensured that the hardware and network systems are functional. We would especially like to acknowledge the contributions of the session organizers who solicited papers and reviews, and ensured that the quality of the meeting remains high. The session organizers (and their associated sessions) are:

*Inna Dubchak, Lior Pachter and Liping Wei (Genome-wide Analysis and Comparative Genomics)*

*Peter Karp, Pedro Romero and Eric Neumann (Genome, Pathway and Interaction Bioinformatics)*

*Willi von der Lieth (Expanding Proteomics to Glycobiology)*

*Lynette Hirschman, Jong C. Park, Junichi Tsujii, Cathy Wu and Limsoon Wong (Literature Data Mining for Biology)*

*Isaac Kohane, Clay Stephens, Julie Schneider and Francisco De La Vega (Human Genomic Variation: Disease, Drug Response, and Clinical Phenotypes)*

*Scott Stanley and Benjamin Salisbury (Phylogenetic Genomics and Genomic Phylogenetics)*

*Peter Clote, Gavin Naylor, and Ziheng Yang (Proteins: Structure, Function and Evolution)*

The PSB organizers and session leaders relied on the assistance of those who capably reviewed the submitted manuscripts. A partial list of reviewers is provided elsewhere in this volume. We thank those who have been left off this list inadvertently or who wish to remain anonymous.

Aloha!

*Pacific Symposium on Biocomputing Co-Chairs*

*October 1, 2001*

*Russ B. Altman  
Stanford University*

*A. Keith Dunker  
Washington State University*

*Lawrence Hunter  
University of Colorado Health Sciences Center*

*Teri E. Klein  
Stanford University*

## Thanks to reviewers . . .

Finally, we wish to thank the scores of paper 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, that requires a great deal of work from many people, and we are grateful to all of you listed below, and to any whose names we may have accidentally omitted.

Aram Adourian	David Paul Holden	Pedro Romero
Laura Almasy	John Holmes	Vincent Schachter
Orly Alter	Roderick V. Jensen	Steffen Schulze-Kremer
Chris Amos	Ruhong Jiang	Jody Schwartz
Mike Bada	Kenneth Karol	Thomas Seidl
Pierre Baldi	Peter Karp	Imran Shah
Serafim Batzoglou	Ju Han Kim	Ron Shamir
Jadwiga Bienkowska	Jessica Kissinger	Roded Sharan
Eckart Bindewald	Alex Lancaster	Victor Solovvey
Erich Bornberg-Bauer	Jobst Landgrebe	Terence Speed
Phil Bradley	Rick Lathrop	Paul Spellman
Richard Broughton	Hans-Peter Lenhof	Scott Stanley
Michael Brudno	Jin-Long Li	Robert Stuart
Andrea Califano	Weizhong Li	Jane Su
Matt Callow	Pat Lincoln	Xiaoping Su
Roland Carel	Jan Liphardt	Zoltan Szallasi
Vincent J. Carey	Irene Liu	Amos Tanay
Simon Cawley	Xiaole Liu	Debra Tanguay
Hue Sun Chan	Gaby Loots	Glenn Tesler
Joseph Chang	Joanne Luciano	Denis Thieffry
Andrew Clark	Andrew Martin	Glenys Thomson
Julio Collado-Vides	Kate McKusick	Jeff Thorne
Josep Comeron	William Newell	Martin Tompa
Olivier Couronne	Magnus Nordborg	Jun'ichi Tsuji
Derek Dimcheff	Gary Nunn	Jacques van Helden
Chris Ding	Matej Oresic	Mike Walker
Roland Dunbrack	Christos Ouzounis	Teresa Webster
Jeremy Edwards	Ivan Ovcharenko	Simon Whelan
Jodi Vanden Eng	Jong Park	Kelly Ewen White
Niklas Eriksen	Peter Park	Glenn Williams
George Estabrook	Hugh Pasika	Limsoon Wong
Andras Fiser	Len Pennacchio	Cathy Wu
Jennifer Gleason	Yitzhak Pilpel	Yu Xia
Richard Goldstein	Tom Plasterer	Dong Xu
Susumu Goto	Darrent Platt	Ying Xu
Douglas Greer	David Pollock	Chen-Hsiang Yeang
Igor Grigoriev	John Quackenbush	John Yin
Mark Grote	Mark Rabin	Ping Zhan
Ivo Gut	Marco Ramoni	Ge Zhang
Alexander J. Hartemink	Aviv Regev	Yingdong Zhao
Lynette Hirschman	Michael Reich	
Steve Holbrook	Markus Ringnér	

# CONTENTS

Preface	v
<b>HUMAN GENOME VARIATION: DISEASE, DRUG RESPONSE, AND CLINICAL PHENOTYPES</b>	
Session Introduction	3
<i>I. Kohane, C. Stephens, J. Schneider, and F. De La Vega</i>	
A Stability Based Method for Discovering Structure in Clustered Data	6
<i>A. Ben-Hur, A. Elisseeff, and I. Guyon</i>	
Singular Value Decomposition Regression Models for Classification of Tumors from Microarray Experiments	18
<i>D. Ghosh</i>	
An Automated Computer System to Support Ultra High Throughput SNP Genotyping	30
<i>J. Heil, S. Glanowski, J. Scott, E. Winn-Deen, I. McMullen, L. Wu, C. Gire, and A. Sprague</i>	
Inferring Genotype from Clinical Phenotype through a Knowledge Based Algorithm	41
<i>B.A. Malin and L.A. Sweeney</i>	
A Cellular Automata Approach to Detecting Interactions Among Single-nucleotide Polymorphisms in Complex Multifactorial Diseases	53
<i>J.H. Moore and L.W. Hahn</i>	
Ontology Development for a Pharmacogenetics Knowledge Base	65
<i>D.E. Oliver, D.L. Rubin, J.M. Stuart, M. Hewett, T.E. Klein, and R.B. Altman</i>	



A SOFM Approach to Predicting HIV Drug Resistance <i>R.B. Potter and S. Draghici</i>	77
Automating Data Acquisition into Ontologies from Pharmacogenetics Relational Data Sources Using Declarative Object Definitions and XML <i>D.L. Rubin, M. Hewett, D.E. Oliver, T.E. Klein, and R.B. Altman</i>	88
On a Family-Based Haplotype Pattern Mining Method for Linkage Disequilibrium Mapping <i>S. Zhang, K. Zhang, J. Li, and H. Zhao</i>	100
<b>GENOME-WIDE ANALYSIS AND COMPARATIVE GENOMICS</b>	
Session Introduction <i>I. Dubchak, L. Pachter, and L. Wei</i>	112
Scoring Pairwise Genomic Sequence Alignments <i>F. Chiaromonte, V.B. Yap, and W. Miller</i>	115
Structure-Based Comparison of Four Eukaryotic Genomes <i>M. Cline, G. Liu, A.E. Loraine, R. Shigeta, J. Cheng, G. Mei, D. Kulp, and M.A. Siani-Rose</i>	127
Constructing Comparative Genome Maps with Unresolved Marker Order <i>D. Goldberg, S. McCouch, and J. Kleinberg</i>	139
Representation and Processing of Complex DNA Spatial Architecture and its Annotated Genomic Content <i>R. Gherbi and J. Herisson</i>	151
Pairwise RNA Structure Comparison with Stochastic Context-Free Grammars <i>I. Holmes and G.M. Rubin</i>	163

Estimation of Genetic Networks and Functional Structures Between Genes by Using Bayesian Networks and Nonparametric Regression	175
<i>S. Imoto, T. Goto and S. Miyano</i>	
Automatic Annotation of Genomic Regulatory Sequences by Searching for Composite Clusters	187
<i>O.V. Kel-Margoulis, T.G. Ivanova, E. Wingender, and A.E. Kel</i>	
EULER-PCR: Finishing Experiments for Repeat Resolution	199
<i>Z. Mulyukov and P.A. Pevzner</i>	
The Accuracy of Fast Phylogenetic Methods for Large Datasets	211
<i>L. Nakhleh, B.M.E. Moret, U. Roshan, K. St. John, J. Sun, and T. Warnow</i>	
Pre-mRNA Secondary Structure Prediction Aids Splice Site Prediction	223
<i>D.J. Patterson, K. Yasuhara, and W.L. Ruzzo</i>	
Finding Weak Motifs in DNA Sequences	235
<i>S.-H. Sze, M.S. Gelfand, and P.A. Pevzner</i>	
Evidence for Sequence-Independent Evolutionary Traces in Genomics Data	247
<i>W. Volkmuth, and N. Alexandrov</i>	
Multiple Genome Rearrangement by Reversals	259
<i>S. Wu and X. Gu</i>	
High Speed Homology Search with FPGAs	271
<i>Y. Yamaguchi, T. Maruyama, and A. Konagaya</i>	
<b>EXPANDING PROTEOMICS TO GLYCOBIOLOGY</b>	
Session Introduction	283
<i>C.-W. von der Lieth</i>	

Glycosylation of Proteins: A Computer Based Method for the Rapid Exploration of Conformational Space of N-Glycans	285
<i>A. Bohne and C.-W. von der Lieth</i>	
Data Standardisation in GlycoSuiteDB	297
<i>C.A. Cooper, M.J. Harrison, J.M. Webster, M.R. Wilkins, and N.H. Packer</i>	
Prediction of Glycosylation Across the Human Proteome and the Correlation to Protein Function	310
<i>R. Gupta and S. Brunak</i>	
<b>LITERATURE DATA MINING FOR BIOLOGY</b>	
Session Introduction	323
<i>L. Hirschman, J. C. Park, J. Tsujii, C. Wu, and L. Wong</i>	
Mining MEDLINE: Abstracts, Sentences, or Phrases?	326
<i>J. Ding, D. Berleant, D. Nettleton, and E. Wurtele</i>	
Creating Knowledge Repositories from Biomedical Reports: The MEDSYNDIKATE Text Mining System	338
<i>U. Hahn, M. Romacker, and S. Schulz</i>	
Filling Preposition-Based Templates to Capture Information from Medical Abstracts	350
<i>G. Leroy and H. Chen</i>	
Robust Relational Parsing Over Biomedical Literature: Extracting Inhibit Relations	362
<i>J. Pustejovsky, J. Castaño, J. Zhang, M. Kotecki, and B. Cochran</i>	
Predicting the Sub-Cellular Location of Proteins from Text Using Support Vector Machines	374
<i>B.J. Stapley, L.A. Kelley, and M.J.E. Sternberg</i>	

A Thematic Analysis of the AIDS Literature <i>W.J. Wilbur</i>	386
--	-----

## **GENOME, PATHWAY AND INTERACTION BIOINFORMATICS**

Session Introduction <i>P. Karp, P. Romero, and E. Neumann</i>	398
---	-----

Pathway Logic: Symbolic Analysis of Biological Signaling <i>S. Eker, M. Knapp, K. Laderoute, P. Lincoln, J. Meseguer, and K. Sonmez</i>	400
--	-----

Towards the Prediction of Complete Protein-Protein Interaction Networks <i>S.M. Gomez and A. Rzhetsky</i>	413
---	-----

Identifying Muscle Regulatory Elements and Genes in the Nematode <i>Caenorhabditis Elegans</i> <i>D. Guhathakurta, L.A. Schriefer, M.C. Hresko, R.H. Waterston, and G.D. Stormo</i>	425
---	-----

Combining Location and Expression Data for Principled Discovery of Genetic Regulatory Network Models <i>A.J. Hartemink, D.K. Gifford, T.S. Jaakkola, and R.A. Young</i>	437
---	-----

The ERATO Systems Biology Workbench: Enabling Interaction and Exchange Between Software Tools for Computational Biology <i>M. Hucka, A. Finney, H.M. Sauro, H. Bolouri, J. Doyle, and H. Kitano</i>	450
---	-----

Genome-Wide Pathway Analysis and Visualization Using Gene Expression Data <i>M.P. Kurhekar, S. Adak, S. Jhunjhunwala, and K. Raghupathy</i>	462
---	-----

Exploring Gene Expression Data with Class Scores	474
<i>P. Pavlidis, D.P. Lewis, and W.S. Noble</i>	

Guiding Revision of Regulatory Models with Expression Data	486
<i>J. Shrager, P. Langley, and A. Pohorille</i>	

Discovery of Causal Relationships in a Gene-Regulation Pathway from a Mixture of Experimental and Observational DNA Microarray Data	498
<i>C. Yoo, V. Thorsson, and G.F. Cooper</i>	

## **PHYLOGENETIC GENOMICS AND GENOMIC PHYLOGENETICS**

Session Introduction	510
<i>S. Stanley and B.A. Salisbury</i>	

Shallow Genomics, Phylogenetics, and Evolution in the Family <i>Drosophilidae</i>	512
<i>M. Zilversmit P. O'Grady, and R. Desalle</i>	

Fast Phylogenetic Methods for the Analysis of Genome Rearrangement Data: An Empirical Study	524
<i>L.-S. Wang, R.K. Jansen, B.M.E. Moret, L.A. Raubeson, and T. Warnow</i>	

Vertebrate Phylogenomics: Reconciled Trees and Gene Duplications	536
<i>R.D.M. Page and J.A. Cotton</i>	

## **PROTEINS: STRUCTURE, FUNCTION AND EVOLUTION**

Session Introduction	548
<i>P. Clote, G.J.P. Naylor, and Z. Yang</i>	

Screened Charge Electrostatic Model in Protein-Protein Docking Simulations	552
<i>J. Fernandez-Recio, M. Totrov, and R. Abagyan</i>	
The Spectrum Kernel: A String Kernel for SVM Protein Classification	564
<i>C. Leslie, E. Eskin, and W.S. Noble</i>	
Detecting Positively Selected Amino Acid Sites Using Posterior Predictive P-Values	576
<i>R. Nielsen and J. P Huelsenbeck</i>	
Improving Sequence Alignments For Intrinsically Disordered Proteins	589
<i>P. Radivojac, Z. Obradovic, C.J. Brown, and A.K. Dunker</i>	
<i>ab initio</i> Folding of Multiple-Chain Proteins	601
<i>J.A. Saunders, K.D. Gibson, and H.A. Scheraga</i>	
Investigating Evolutionary Lines of Least Resistance Using the Inverse Protein-Folding Problem	613
<i>J. Schonfeld, O. Eulenstein, K. Vander Velden, and G.J.P. Naylor</i>	
Using Evolutionary Methods to Study G-Protein Coupled Receptors	625
<i>O. Soyer, M.W. Dimmic, R.R. Neubig, and R.A. Goldstein</i>	
Progress in Predicting Protein Function from Structure: Unique Features of O-Glycosidases	637
<i>E.W. Stawiski, Y. Mandel-Gutfreund, A.C. Lowenthal, and L. M. Gregoret</i>	
Support Vector Machine Prediction of Signal Peptide Cleavage Site Using a New Class of Kernels for Strings	649
<i>J.-P. Vert</i>	

Constraint-Based Hydrophobic Core Construction for Protein Structure Prediction in the Face-Centered-Cubic Lattice <i>S. Will</i>	661
Detecting Native Protein Folds Among Large Decoy Sets with Hydrophobic Moment Profiling <i>R. Zhou and B.D. Silverman</i>	673

**Session**  
**Introductions and**  
**Peer Reviewed**  
**Papers**



