



Life Sciences Society

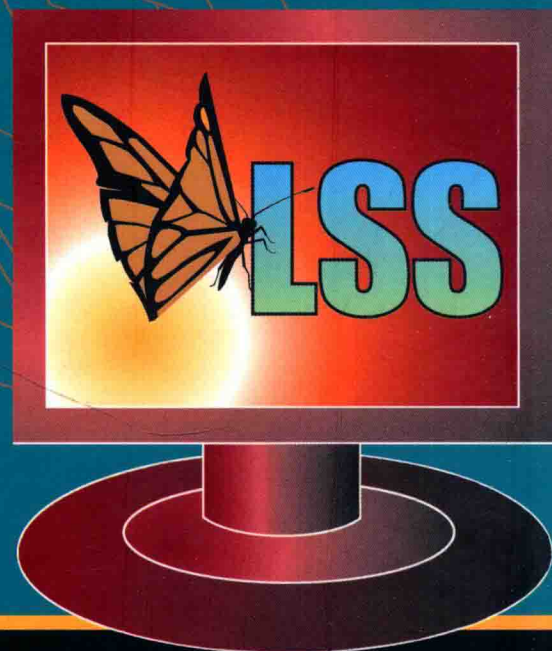
COMPUTATIONAL SYSTEMS BIOINFORMATICS

CSB2007 CONFERENCE PROCEEDINGS
Volume 6

**University of California
San Diego, USA
13–17 August 2007**

Editors

**Peter Markstein
Ying Xu**



Imperial College Press



Life Sciences Society

COMPUTATIONAL SYSTEMS BIOINFORMATICS

CSB2007 CONFERENCE PROCEEDINGS

Volume 6

**University of California
San Diego, USA
13–17 August 2007**

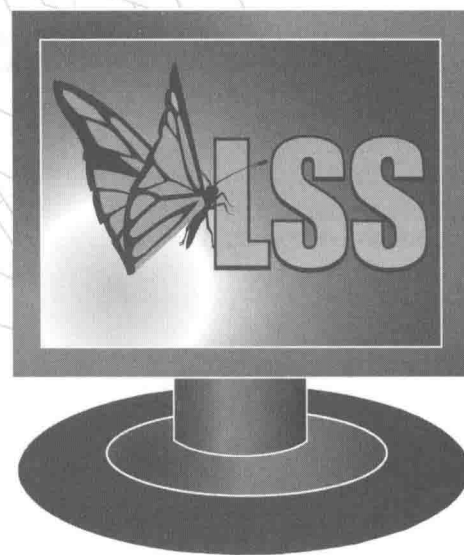
Editors

Peter Markstein

in silico Labs, LLC, USA

Ying Xu

University of Georgia, USA



Published by

Imperial College Press
57 Shelton Street
Covent Garden
London WC2H 9HE

Distributed 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.

COMPUTATIONAL SYSTEMS BIOINFORMATICS

Proceedings of the Conference CSB 2007 — Vol. 6

Copyright © 2007 by Imperial College Press

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-13 978-1-86094-872-5

ISBN-10 1-86094-872-3

Printed by Fulisland Offset Printing (S) Pte Ltd, Singapore



Life Sciences Society

**COMPUTATIONAL
SYSTEMS
BIOINFORMATICS**



Thank You CSB2007 Gold Sponsor

The Life Sciences Society, LSS Directors,
together with the CSB2007 Program Committee and
Conference Organizing Committee are extremely grateful to the

Hewlett-Packard Company

for their Gold Sponsorship of the
Sixth Annual Computational Systems Bioinformatics Conference, CSB2007
at University of California San Diego,
La Jolla, California, August 13-17, 2007



PREFACE

The 21st Century has seen the emergence of tremendous vigor and excitement at the interface between computing and biology. Following years of investment by industry and then private foundations, the US Federal Government has greatly increased its support. Increasingly, the experimental findings from all of the biological sciences are becoming data rich and their practitioners are turning to the use of computational methods to manage and analyze the data. In light of the growing opportunities and excitement at the frontier interface between computing and biology, a few scientists turned conference planners organized the first Computational Systems Bioinformatics (CSB) conference in 2001 at Stanford, CA; CSB continued each August at Stanford over the next five years. During this time, many computer scientists and other engineers, as well as biologists, have attended CSB meetings, which have particularly served to introduce cutting edge biological inquiry and challenge problems to investigators from quantitative science backgrounds.

CSB, more recently, became the public venue for the not-for-profit Life Sciences Society, or LSS, which was founded, in part, to enhance the opportunities at the interface between the quantitative / engineering sciences and the biological sciences. In 2006, LSS was honored to be invited to hold CSB2007 on the campus of the University of California at San Diego, UCSD. Some future meetings at UCSD, and ultimately, at other universities, as well as satellite sessions at bioinformatics meetings around the world, are anticipated over the next several years. The Stanford / Bay Area / Silicon Valley venue, along with presenting highlights in bioinformatics, has been especially valuable for connecting individuals from the computer and electronics industry with investigators in the pure and applied life sciences.

The current venue should provide some connections to telecommunications, while sustaining some of the earlier opportunities, but we anticipate an enhanced interaction with basic and applied biotechnology. The University of California at San Diego grew from the Scripps Institute of Oceanography and began as a graduate school with a strong focus on

the natural sciences. The rich research culture around UCSD and many neighboring institutions would generally be termed the venue of the Torrey Pines Mesa, an area very rich in biotechnology research activities. Today, San Diego has the largest cluster of Life Sciences centers with 26 research institutes (including UCSD and a suite of Institutes: Salk, Neurosciences, Scripps Research, Burnham Medical, as well as smaller not for profits) located in an area less than 10 square miles. For more on San Diego's R&D Life Sciences Centers and the vibrant biotechnology and pharmacology efforts, do visit the BioCom website: http://www.biocom.org/Portals/0/SanDiegoLifeScienceNumbers_Fall06.pdf.

CSB2007 will continue to be a 5 day single track conference, with three core plenary presentations days sandwiched between a day of practical tutorials, long a very popular feature, and a day workshops exploring the future. Thus, CSB2007 includes several half day tutorials, 30 refereed papers plus keynote and invited speakers, and posters, during its five full days. Special events for the evenings are planned.

CSB2007, as in each of its previous years, owes a lot to its many hard working volunteers, who are listed under the Committees. The indefatigable energy of Vicky and Peter Markstein continues to sustain the magnitude and amplitude of the extraordinary science and technology vector that is CSB, and their partnership with Ying Xu also remains essential. The efforts to manage and enhance local arrangements, by Kayo Arima, Patrick Shih, Lydia Grech and Ed Buckingham, should also be acknowledged.

A few words, naturally, about SoCal: bring family and guests to enjoy San Diego's world-famous attractions as SeaWorld, the San Diego Zoo, the Wild Animal Park and LEGOLAND California, as well as historic cultural gems Balboa Park and Old Town, and of course, the "endless" beach.

John Wooley, General Conference Chair

COMMITTEES

Steering Committee

Phil Bourne - University of California, San Diego
 Eric Davidson - California Institute of Technology
 Steven Salzberg - The Institute for Genomic Research
 John Wooley - University of California San Diego, San Diego Supercomputer Center

Organizing Committee

Kayo Arima – University of California San Diego, Local Arrangements
 Pat Blauvelt - Communications
 Ed Buckingham – LSS VP Conferences
 Kass Goldfein - Finance Consultant
 Lydia Grech – University of California San Diego, Local Arrangements
 Fenglou Mao – University of Georgia, On-Line Registration and Refereeing Website
 Vicky Markstein - Life Sciences Society, **Co-Chair**, LSS President
 Patrick Shih – University of California San Diego, Local Arrangements
 Jean Tsukamoto - Graphics Design
 Bill Wang - Sun Microsystems Inc, LSS Information Technology
 John Wooley - University of California San Diego, San Diego Supercomputer Center, **Co-Chair**

Program Committee

Tatsuya Akutsu - Kyoto University
 Phil Bourne – University of California San Diego
 Jake Chen - Indiana University
 Amar Das - Stanford University
 Chris Ding – Lawrence Berkeley Laboratory
 Roderic Guigo, IMIM, Barcelona
 Tao Jiang - University of California Riverside
 Lydia Kavraki – Rice University
 Hoong-Chien Lee - National Central University, Taiwan
 Ann Loraine - University of Alabama
 Michele Markstein – Harvard University
 Peter Markstein - Hewlett-Packard Co., **Co-chair**
 Satoru Miyano - University of Tokyo
 Sean Mooney - Indiana University
 Jan Mrazek - University of Georgia
 Isidore Rigoutsos - IBM TJ Watson Research Center
 Andrey Rzhetsky - Columbia University

Hershel M. Safer, Weizmann Institute of Science
David States - University of Michigan
Anna Tramontano - University of Rome
Olga Troyanskaya - Princeton University
Alfonso Valencia - Centro Nacional de Biotecnología, Spain
Eberhard Voit - Georgia Tech
Limsoon Wong - Institute for Infocomm Research
Ying Xu - University of Georgia, **Co-chair**
Aidong Zhang - SUNY Buffalo
Michael Zhang - Cold Spring Harbor Laboratory
Xianghong Jasmine Zhou - University of Southern California
Yaoqi Zhou - Indiana University

Assistants to the Program Co-Chairs

Ann Terka – University of Georgia
Joan Yantko – University of Georgia

Poster Committee

Nigam Shah - Stanford University, **Chair**
Patrick Shih – University of California San Diego

Tutorial Committee

Weizhong Li – University of California San Diego,
Al Shpuntoff – Syngenta Biotechnology Institute, **Chair**
John Wooley – University of California San Diego

Workshop Committee

Iddo Friedberg – University of California San Diego, **Co-Chair**
Weizhong Li – University of California San Diego, **Co-Chair**
Patrick Shih – University of California San Diego

REFEREES

Tatsuya Akutsu	Yuki Kato	Andrey Rzhetsky
Mar Alba	Lydia Kavradi	
	Melissa Kemp	Hershel M. Safer
Takis Benos		Sudipto Saha
	HC Lee	David States
Phil Bourne	Hoong-Chien Lee	Wing-Kin Sung
	Haiquan Li	
Liming Cai	Jing Li	Takeyuki Tamura
Ildefonso Cases	Xiaoman Shawn Li	Anna Tramontano
Robert Castelo	Guohui Lin	Olga Troyanskaya
Dongsheng Che	Chun-Chi Liu	Aristotelis Tsirigos
Jake Chen	Guimei Liu	
Liang Chen	Huiqing Liu	Alfonso Valencia
David Chew	Yunlong Liu	Siren Veflingstad
Young-Rae Cho	Ann Loraine	Eberhard Voit
I-Chun Chou		
Xiangqin Cui	Michia Ma	John Wagner
	Fenglou Mao	Mingyi Wang
PhuongAn Dam	Peter Markstein	Limsoon Wong
Amar Das	David Martin	Hongwei Wu
David de Juan	Satoru Miyano	Jialiang Wu
Chris Ding	Mark Moll	
	Jan Mrazek	Min Xu
Iakes Ezkurdia		Ying Xu
	Masao Nagasaki	
Matteo Floris	Luay Nakleh	Weiwei Yin
Sylvain Foissac	Christoforous Nikolau	
	Juan Nunez-Iglesias	Kangyu Zhang
David Gilley		Michael Zhang
Gautam Goel	Victor Olman	Shiju Zhang
Roderic Guigo		Fengfeng Zhou
	Miguel Padilla	Ruhong Zhou
Scott Harrison	Grier Page	Wen Zhou
Nurit Haspel	Florencio Pazos	Xianghong Jasmine Zhou
Jianjun Hu	Daniel Platt	Yaoqi Zhou
Woochang Hwang		
	Zhen Qi	
Seiya Imoto		
	Predrag Radivojac	
Tao Jiang	Isidore Rigoutsos	

CONTENTS

Preface	vii
Committees	ix
Referees	xi
Keynote Address	
Quantitative Aspects of Gene Regulation in Bacteria: Amplification, Threshold, and Combinatorial Control <i>Terry Hwa</i>	3
Whole-Genome Analysis of Dorsal Gradient Thresholds in the <i>Drosophila</i> Embryo <i>Julia Zeitlinger, Rob Zinzen, Dmitri Papatsenko et al.</i>	5
Invited Talks	
Learning Predictive Models of Gene Regulation <i>Christina Leslie</i>	9
The Phylofacts Phylogenomic Encyclopedias: Structural Phylogenomic Analysis Across the Tree of Life <i>Kimmen Sjölander</i>	11
Mapping and Analysis of the Human Interactome Network <i>Kavitha Venkatesan</i>	13
Gene-Centered Protein-DNA Interactome Mapping <i>A.J. Marian Walhout</i>	15
Proteomics	
Algorithm for Peptide Sequencing by Tandem Mass Spectrometry Based on Better Preprocessing and Anti-Symmetric Computational Model <i>Kang Ning and Hon Wai Leong</i>	19
Algorithms for Selecting Breakpoint Locations to Optimize Diversity in Protein Engineering by Site-Directed Protein Recombination <i>Wei Zheng, Xiaoduan Ye, Alan M. Friedman and Chris Bailey-Kellogg</i>	31
An Algorithmic Approach to Automated High-Throughput Identification of Disulfide Connectivity in Proteins Using Tandem Mass Spectrometry <i>Timothy Lee, Rahul Singh, Ten-Yang Yen and Bruce Macher</i>	41

Biomedical Application

- Cancer Molecular Pattern Discovery by Subspace Consensus Kernel Classification
Xiaoxu Han 55
- Efficient Algorithms for Genome-Wide tagSNP Selection Across Populations via the Linkage Disequilibrium Criterion
Lan Liu, Yonghui Wu, Stefano Lonardi and Tao Jiang 67
- Transcriptional Profiling of Definitive Endoderm Derived from Human Embryonic Stem Cells
Huiqing Liu, Stephen Dalton and Ying Xu 79

Pathways, Networks and Systems Biology

- Bayesian Integration of Biological Prior Knowledge into the Reconstruction of Gene Regulatory Networks with Bayesian Networks
Dirk Husmeier and Adriano V. Werhli 85
- Using Indirect Protein-Protein Interactions for Protein Complex Predication
Hon Nian Chua, Kang Ning, Wing-Kin Sung et al. 97
- Finding Linear Motif Pairs from Protein Interaction Networks: A Probabilistic Approach
Henry C.M. Leung, M.H. Siu, S.M. Yiu et al. 111
- A Markov Model Based Analysis of Stochastic Biochemical Systems
Preetam Ghosh, Samik Ghosh, Kalyan Basu and Sajal K. Das 121
- An Information Theoretic Method for Reconstructing Local Regulatory Network Modules from Polymorphic Samples
Manjunatha Jagalur and David Kulp 133
- Using Directed Information to Build Biologically Relevant Influence
Arvind Rao, Alfred O. Hero III, David J. States and James Douglas Engel 145
- Discovering Protein Complexes in Dense Reliable Neighborhoods of Protein Interaction Networks
Xiao-Li Li, Chuan-Sheng Foo and See-Kiong Ng 157
- Mining Molecular Contexts of Cancer via In-Silico Conditioning
Seungchan Kim, Ina Sen and Micheal Bittner 169

Genomics

- Prediction of Transcription Start Sites Based on Feature Selection Using AMOSA
Xi Wang, Sanghamitra Bandyopadhyay, Zhenyu Xuan et al. 183
- Clustering of Main Orthologs for Multiple Genomes
Zheng Fu and Tao Jiang 195
- Deconvoluting the BAC-Gene Relationships Using a Physical Map
Yonghui Wu, Lan Liu, Timothy J. Close and Stefano Lonardi 203

A Grammar Based Methodology for Structural Motif Finding in ncRNA Database Search <i>Daniel Quest, William Tappich and Hesham Ali</i>	215
IEM: An Algorithm for Iterative Enhancement of Motifs Using Comparative Genomics Data <i>Erliang Zeng, Kalai Mathee and Giri Narasimhan</i>	227
MANGO: A New Approach to Multiple Sequence Alignment <i>Zefeng Zhang, Hao Lin and Ming Li</i>	237
Learning Position Weight Matrices from Sequence and Expression Data <i>Xin Chen, Lingqiong Guo, Zhaocheng Fan and Tao Jiang</i>	249
Structural Bioinformatics	
Effective Labeling of Molecular Surface Points for Cavity Detection and Location of Putative Binding Sites <i>Mary Ellen Bock, Claudio Garutti and Conettina Guerra</i>	263
Extraction, Quantification and Visualization of Protein Pockets <i>Xiaoyu Zhang and Chandrajit Bajaj</i>	275
Uncovering the Structural Basis of Protein Interactions with Efficient Clustering of 3-D Interaction Interfaces <i>Zeyar Aung, Soon-Heng Tan, See-Kiong Ng and Kian-Lee Tan</i>	287
Enhanced Partial Order Curve Comparison Over Multiple Protein Folding Trajectories <i>Hong Sun, Hakan Ferhatosmanoglu, Motonori Ota and Yusu Wang</i>	299
fRMSDPred: Predicting Local RMSD Between Structural Fragments Using Sequence Information <i>Huzefa Rangwala and George Karypis</i>	311
Consensus Contact Prediction by Linear Programming <i>Xin Gao, Dongbo Bu, Shuai Cheng Li et al.</i>	323
Improvement in Protein Sequence-Structure Alignment Using Insertion/Deletion Frequency Arrays <i>Kyle Ellrott, Jun-Tao Guo, Victor Olman and Ying Xu</i>	335
Composite Motifs Integrating Multiple Protein Structures Increase Sensitivity for Function Prediction <i>Brian Y. Chen, Drew H. Bryant, Amanda E. Cruess et al.</i>	343
Ontology, Database and Text Mining	
An Active Visual Search Interface for Medline <i>Weijian Xuan, Manhong Dai, Barbara Mirel et al.</i>	359
Rule-Based Human Gene Normalization in Biomedical Text with Confidence Estimation <i>William W. Lau, Calvin A. Johnson and Kevin G. Becker</i>	371

CBioC: Beyond a Prototype for Collaborative Annotation of Molecular Interactions from the Literature <i>Chitta Baral, Graciela Gonzalez, Anthony Gitter et al.</i>	381
Biocomputing	
Supercomputing with Toys: Harnessing the Power of NVIDIA 8800GTX and Playstation 3 for Bioinformatics Problem <i>Justin Wilson, Manhong Dai, Elvis Jakupovic et al.</i>	387
Exact and Heuristic Algorithms for Weighted Cluster Editing <i>Sven Rahmann, Tobias Wittkop, Jan Baumbach et al.</i>	391
Method for Effective Virtual Screening and Scaffold-Hopping in Chemical Compounds <i>Nikil Wale, George Karypis and Ian A. Watson</i>	403
Transcriptomics and Phylogeny	
Improving the Design of Genechip Arrays by Combining Placement and Embedding <i>Sérgio Anibal de Carvalho Jr. and Sven Rahmann</i>	417
Modeling Species-Genes Data for Efficient Phylogenetic Inference <i>Wenyuan Li and Ying Liu</i>	429
Reconciliation with Non-Binary Species Trees <i>Benjamin Vernot, Maureen Stolzer, Aiton Goldman and Dannie Durand</i>	441
Author Index	453

Computational Systems Bioinformatics 2007



Keynote Address

QUANTITATIVE ASPECTS OF GENE REGULATION IN BACTERIA: AMPLIFICATION, THRESHOLD, AND COMBINATORIAL CONTROL

Terry Hwa

*Center for Theoretical Biological Physics and Department of Physics
University of California San Diego
9500 Gilman Drive
La Jolla, CA 92093-0374*

Biological organisms possess an enormous repertoire of genetic responses to ever-changing combinations of cellular and environmental signals. Unlike digital electronic circuits however, signal processing in cells is carried out by a limited number of asynchronous devices in fluctuating aqueous environments.

In this talk, I will discuss the control of genetic responses in bacteria. Theoretical analysis of the known mechanisms of transcriptional control suggests "programmable" mechanisms for implementing a broad class of combinatorial control. Further analysis of post-

transcriptional control suggests mechanisms for signal amplification, threshold response, and noise attenuation. I will present experimental characterization of some of these bio-computational "devices", as well as experiments illustrating how promoter sequences may be "trained" by directed evolution. Quantitative characterization and controlled manipulation of these devices may bring about predictive understanding of biological control systems, and reveal interesting, novel strategies of distributed computation.

