DOE GRANT NO. DE-FG02-05ER63568

FINAL PROGRESS REPORT – YEAR 3 OF 3-YEAR

ISMB Conference Funding for Students and Young Scientists

Historical Description

The Intelligent Systems for Molecular Biology (ISMB) conference has provided a general forum for disseminating the latest developments in bioinformatics on an annual basis for the past 13 years. ISMB is a multidisciplinary conference that brings together scientists from computer science, molecular biology, mathematics and statistics. The goal of the ISMB meeting is to bring together biologists and computational scientists in a focus on actual biological problems, i.e., not simply theoretical calculations. The combined focus on "intelligent systems" and actual biological data makes ISMB a unique and highly important meeting, and 13 years of experience in holding the conference has resulted in a consistently well organized, well attended, and highly respected annual conference.

"Intelligent systems" include any software which goes beyond straightforward, closed-form algorithms or standard database technologies, and encompasses those that view data in a symbolic fashion, learn from examples, consolidate multiple levels of abstraction, or synthesize results to be cognitively tractable to a human, including the development and application of advanced computational methods for biological problems. Relevant computational techniques include, but are not limited to: machine learning, pattern recognition, knowledge representation, databases, combinatorics, stochastic modeling, string and graph algorithms, linguistic methods, robotics, constraint satisfaction, and parallel computation. Biological areas of interest include molecular structure, genomics, molecular sequence analysis, evolution and phylogenetics, molecular interactions, metabolic pathways, regulatory networks, developmental control, and molecular biology generally. Emphasis is placed on the validation of methods using real data sets, on practical applications in the biological sciences, and on development of novel computational techniques.

The ISMB conferences are distinguished from many other conferences in computational biology or artificial intelligence by an insistence that the researchers work with real molecular biology data, not theoretical or toy examples; and from many other biological conferences by providing a forum for technical advances as they occur, which otherwise may be shunned until a firm experimental result is published. The resulting intellectual richness and cross-disciplinary diversity provides an important opportunity for both students and senior researchers. ISMB has become the premier conference series in this field with refereed, published proceedings, establishing an infrastructure to promote the growing body of research.

ISMB 2005 Summary Report

The ISMB 2005 meeting was held June 25-29, 2005 at the Renaissance Center in Detroit, Michigan. The meeting attracted over 1,730 attendees. The science presented was exceptional, and in the course of the five-day meeting, 56 scientific papers, 710 posters, 47 Oral Abstracts, 76 Software demonstrations, and 14 tutorials were presented. The attendees represented a broad spectrum of backgrounds with 7% from commercial companies, over 28% qualifying for student registration, and 41 countries were represented at the conference, emphasizing its important international aspect.

Over the years another event has become a key component of the ISMB meetings: in 2005 there were seven special interest group (SIG) meetings organized by researchers, who used the presence of a considerable fraction of the scientific community to spawn off more specialized meetings. In 2005 there were four one-day SIG meetings and three two-day SIG meetings. The following subjects constituted one-day meetings held before the main meeting in 2005: (1) Automated Function Prediction, (2), Bioinformatics and

Disease, (3) BioLINK, (4) Bio-Ontologies . The two-day meetings consisted of the following subjects: (1) Alternative Splicing, (2) Bioinformatics Open Source Conference-BOSC'2005, (3), Joint Meeting: BioPathways and SIGSM.

At the main meeting each session was introduced by a keynote address (see below). The keynote speakers are leaders in their fields and are especially selected to emphasize new areas related to the conference, biological problems ripe for investigation by computational means, and exciting research areas that are less well known to the ISMB audience. The keynote addresses, therefore, always play a key role in the continuing education of all attendees, and most especially the younger scientists who are often participating and gaining exposure to the international bioinformatics community and its leaders for the first time.

The ISMB conference is especially important because the cultures of computer science and biology are so disparate. ISMB, as a full-scale technical conference with refereed proceedings that have been indexed by both MEDLINE and Current Contents since 1996, bridges this cultural gap. Since 2001 Oxford University Press, as a special issue of the journal *Bioinformatics*, has published the *ISMB Proceedings*.

Student support was granted based on applications by the respective individuals. Preference and amount of reimbursement was determined based on the following factors:

- Status as co-author of a paper submitted or accepted. (A poster or paper presentation is required of funding student recipients);
- Degree of financial hardship;
- Status as woman, minority, or disabled scientist;
- Status as junior scientist (student or post-graduate researcher within 5 years of degree);
- Amount of reimbursement required;
- Geographic distribution of awards (emphasis on greater distance/cost of travel).

No DOE funds were used for travel of conference organizers. No DOE funds were used for the presentation of the conference itself. The grant did not fund the PI or the PI's research in any way.

Nearly 24% of the participants at ISMB 2005 were female and 24% of the ISMB2005's travel fellowship awardees were female. (Funded by the DOE and other sources.)

Previous and Future Meetings

The ISMB conference series was established in 1993. The conference has been held regularly and drawn increasing numbers of participants starting at about 130 in 1993 and growing to over 1 ,700 in 2005. The conference purposefully alternates its venue between North American and non-North American sites to foster international exchange and cooperation. The locations of previous and future ISMB conferences were/are as follows:

1993 National Library of Medicine, Bethesda, USA
1994 Stanford University, USA
1995 Cambridge University, UK
1996 Washington University, USA
1997 Halkidiki, Greece
1998 Montreal, Canada
1999 Heidelberg, Germany
2000 San Diego, USA
2001 Copenhagen, Denmark
2002 Edmonton, Canada
2003 Brisbane, Australia
2004 Glasgow, Scotland
2005 Detroit, Michigan
2006 Fortaleza, Brazil

The ISMB conference is the primary meeting sponsored by the International Society for Computational Biology (ISCB). The ISCB is dedicated to advancing the scientific understanding of living systems through computation, emphasizing the role of computing and informatics in advancing molecular biology. The Society was formally incorporated in 1998 and has been granted nonprofit tax-exempt status in the United States. The Society now has an annual budget in excess of one million dollars, with the majority of that allocated to producing the ISMB conference. In 2005 there were almost 1,900 active ISCB members from over 50 countries including over 700 students and postdoctoral trainee members.

DOE Grant for Travel - ISMB 2005, Detroit, Michigan

Average award per person $746.24

NAME	TITLE OF PAPER OR POSTER
Goldberg, Debra	Breaking the Power Law: Improved Model Selection Suggests that Many Biological Networks Have Multiple Hidden Edge Types
Hibbs, Matthew	Visual Methods for Statistical Analysis of Microarray Clusters
Kall, Lukas	An HMM Posterior Decoder for Sequence Feature Prediction that Includes Homology Information
Komandurglayavilli, Ravikumar	Sequence graphs for Phosphorylation motifs
Mahony, Shaun	Improved Detection of DNA Motifs Using a Self-organized Clustering of Familial Binding Profiles
Marinescu, Voichita	The MAPPER platform for the computational identification of transcription factor binding sites
Mayrose, Itay	A Gamma mixture model better accounts for among site rate heterogeneity
Mesirov, Jill	Gene Expression Analysis: A Knowledge-based Approach
Minin, Vladimir	Mapping Recombination Hot-Spots onto Gaussian Markov Random Fields
Neeman, Yossef	Are antisense transcripts prone to A-to-I RNA editing ?
Nimrod, Guy	In silico Identification of Functional Regions in Proteins
Novotny, Marian	On the importance of being left-handed
Opiyo, Stephen	Using Partial Least Square Regression to Classify Protein Family with Weak Sequence Similarities
Portugaly, Elon	Automatic Identification and Classification of Protein Domains
Sadka, Tali	Families of Membranous Proteins can be Characterized by the Amino Acid Composition of their Transmembrane Domains
Sakabe, Noboru	Building and surveying transcript-based alternative splicing databases
Sarkar, Indra	TEPC: Total Evidence Phylogenetic Correlation of Microbial Phenotypes and Genotypes
Schaub, Marc	Alternate Splicing in Cancer: Challenges in the Statistical Analysis of EST Data
Shafer, Paul	Hubs of Knowledge: using the functional link structure in Biozon to mine for biologically significant entities
Shmygelska, Alena	Search for Folding Nuclei in Native Protein Structures
Singer, Greg	The detection of novel alternative promoters in the human genome and their importance in various cancers
Song, Yun	Efficient Computation of Close Lower and Upper Bounds on the Minimum Number of Recombinations in Biological Sequence Evolution
Soumyaroop, Bhattacharya	Transformation of Expression Intensities Across Generations of Affymetrix Microarrays using Sequence Matching and Regression Modeling
Stadler, Michael	Inferring Splicing Regulatory Activity of Short Oligonuecleotides from Sequence Neighborhood
Strope, Pooja	Comparative Analysis of Alignment-based and Alignment-free Protein Classifiers
Su, Rong	A CASE STUDY: Web-Based Informatics System for Mouse Knockout Production
Tabach, Yuval	Genome-wide transcription regulatory circuits controlling cellular malignant
Tae, Hongseok	A polyketide database and analysis system for microbial genomes.
Taylor, Todd	Two applications of Delaunay contact matrices to the analysis of protein structures

Terribilini, Michael	A computational method to identify amino acid residues involved in protein-DNA interactions
Thomas, Asha	Fusion of Multiple Decision Models in Proteomic Biomarker Discovery
Thornton, Janet	From Proteins to Life - Old and New Challenges
Tran, Nam	Knowledge-Based Framework for Hypothesis Formation in Biochemical Networks
Tseng, Tsai-Tien	Detection of Horizontal Gene Transfer in Whole Metabolic Pathways
Vashist, Akshay	Candidate ortholog clusters in human, mouse and chicken genomes
Vijaya, Parthiban	Computational analysis of RNA binding proteins based on composition, sequence and structural information
Wang, Kai	Identifying Functional Signatures from Structural Alignments
Wang, Ting	Combining Phylogenetic Data and Network Topology to Identify Regulatory Motifs
Wei, Lai	Literature Based Functional Analysis of Microarray Data
Woo, Yong	Experimental Design for Three-Color and Four-Color Gene Expression Microarrays
Wu, Chunlei	Sequence dependence of cross hybridization on short oligo microarrays
Yamanishi, Yoshihiro	Supervised Enzyme Network Inference from the Integration of Genomic Data and Chemical Information
Yan, Changhui	Computational Prediction of RNA-Binding Sites in Proteins Based on Amino Acid Sequence
Yang, Jack	IUPUI: Intrinsic Unstructured Protein Unsupervised-supervised Identifier - A Software Tool
Yang, Mary	A Recursive Maximum Contrast Tree Approach to Learning Protein Functional Classes
Ye, Ping	Commensurate Distances and Motifs from Genetic Congruence and Protein Interaction Networks in Yeast
Zhang, Miao	Improved EST and genomic sequence alignment

TOTAL spent: $35,073.37
DOE SUM committed: $35,000.00
Overspent funds were covered by other sources from ISMB budget

Keynote Addresses 2005

- Ewan Birney-EMBL-EB
- Janet Thornton-EBI-EMBL
- Howard Cash-President of Gene Codes and Gene Codes Forensics, Inc., inc.
- Peter Hunter-University of Auckland
- Jill P. Mesirov-Broad Institute of MIT and Harvard
- Satoru Miyano, University of Tokyo
- Pavel Pevzner-University of California at San Diego
- Gunnar von Heijne-Stockholm Bioinformatics Center

Scientific Papers 2005

Of the 426 manuscripts that were submitted; the following 56 were selected for presentation and publication in the special issue of *Bioinformatics*:

CORRESPONDING AUTHOR	TITLE
Gopalacharyulu, Lindfors, Bounsaythip, Kivioja, Yetukuri, Hollmen, Oresic	Data Integration and Visualization System for Enabling Conceptual Biology
Sharan, Myers	A Motif-based Framework for Recognizing Sequence Families
Dolan, Ni, Camon, Blake	A Procedure for Assessing GO Annotation Consistency
Käll, Krogh, Sonnhammer	An HMM Posterior Decoder for Sequence Feature Prediction that Includes Homology Information
Cheung, Yip, Smith, deKnikker, Masiar, Gerstein	YeastHub: A Semantic Web Use Case for Integrating Data in the Life Sciences Domain
Mahony, Golden, Smith, Benos	Improved Detection of DNA Motifs Using a Self-organized Clustering of Familial

	Binding Profiles
Kou, Cohen, Murphy	High-recall Protein Entity Recognition Using a Dictionary
S. Mika and B. Rost	Protein Names Precisely Peeled Off Free Text
Prakash, Tompa	Statistics of Local Multiple Alignments
Narayanaswamy, Ravikumar, Vijay-Shanker	Beyond The Clause: Extraction of Phosphorylation Information from Medline Abstracts
Nagarajan, Jones, Keich	Computing the P-value of the Information Content from an Alignment of Multiple Sequences
Batt, Ropers, de Jong, Geiselmann, Mateescu, Page, Schneider	Validation of Qualitative Models of Genetic Regulatory Networks by Model Checking: Analysis of the Nutritional Stress Response in Escherichia coli
Hahn, Lee	Identification of Nine Human-specific Frameshift Mutations by Comparative Analysis of the Human and the Chimpanzee Genome Sequences
Ben-Hur, Noble	Kernel Methods for Predicting Protein-protein Interactions
Tuller, Chor	Maximum Likelihood of Evolutionary Trees: Hardness and Approximation
Ye, Osterman, Overbeek, Godzik	Automatic Detection of Subsystem/Pathway Variants in Genome Analysis
Dimmic, Hubisz, Bustamante, Nielsen	Detecting Coevolving Amino Acid sites using Bayesian Mutational Mapping
Jönsson, Heisler, Reddy, Agrawal, Gor, Shapiro, Mjolsness, Meyerowitz	Modeling the Organization of the WUSCHEL Expression Domain in the Shoot Apical Meristem
Song, Wu, Gusfield	Efficient Computation of Close Lower and Upper Bounds on the Minimum Number of Recombinations in Biological Sequence Evolution
Yu, Zaveljevski, Stevens, Yackovich, Reifman	Classifying Noisy Protein Sequence Data: A Case study of Immunoglobulin Light Chains
Cline, Blume, Cawley, Clark, Hu, Lu, Salomonis, Wang, Williams	A Statistical Method for Detecting Splice Variation from Expression Data
Borgwardt, Ong, Schoenauer, Vishwanathan, Smola, Kriegel	Protein Function Prediction via Graph Kernels
Soukup, Lee, Cho	Robust Classification Modeling on Microarray Data Using Misclassification Penalized Posterior
Garay-Malpartida, Occhiucci, Alves, Belizário	CaSPredictor: A new Computer-based Tool for Caspase Substrate Prediction
Woo, Krueger, Kaur, Churchill	Experimental Design for Three-Color and Four-Color Gene Expression Microarrays
Sadka, Linial	Families of Membranous Proteins can be Characterized by the Amino Acid Composition of their Transmembrane Domains
Li, Meyer, Liu	A Hidden Markov Model for Analyzing ChIP-chip Experiments on Genome Tiling Arrays and its Application to p53 Binding Sequences
Mettu, Lilien, Donald	High-Throughput Inference of Protein-Protein Interfaces from Unassigned NMR Data
Dueck, Morris, Frey	Multi-way Clustering of Microarray Data using Probabilistic Sparse Matrix Factorization
Halperin, Kimmel, Shamir	Tag SNP Selection in Genotype Data for Maximizing SNP Prediction Accuracy
Cheng, Baldi	Three-Stage Prediction of Protein Beta-Sheets by Neural Networks, Alignments, and Graph Algorithms
Smith, Sumazin, Das, Zhang	Mining ChIP-chip Data for Transcription Factor and Cofactor Binding Sites
Brunette, Brock	Improving Protein Structure Prediction With Model-Based Search
Huang, Morris, Frey Paper	GenXHC: A Probabilistic Generative Model for Cross-hybridization Compensation in High-density Genome-wide Microarray Data
Ko, Murga, Ondrechen	Prediction of Active Sites for Protein Structures from Computed Chemical Properties
Raetsch, Sonnenburg, Schoelkopf	RASE: Recognition of Alternatively Spliced Exons in C. elegans
Nimrod, Glaser, Steinberg, Ben-Tal, Pupko	*In silico* Identification of Functional Regions in Proteins
Hannenhalli, Wang	Enhanced Position Weight Matrices using Mixture Models

Winstanley, Abeln, Deane	How Old is Your Fold?
Price**,** Jones, Pevzner	De Novo Identification of Repeat Families in Large Genomes
Shmygelska	Search for Folding Nuclei in Native Protein Structures
Tang**,** Mechref, Novotny	Automatic Interpretation of MS/MS Spectra of Oligosaccharides
Cortes, Simeon**,** Ruiz de Angulo, Guieysse, Remaud-Simeon, Tran	A Path Planning Approach for Computing Large-Amplitude Motions of Flexible Molecules
Ernst**,** Nau, Bar-Joseph	Clustering Short Time Series Gene Expression Data
Edgar**,** Myers	PILER: Identification and Classification of Genomic Repeats
Noble**,** Kuehn, Thurman, Yu, Stamatoyannopoulos	Predicting the *in vivo* Signature of Human Gene Regulatory Sequences
Tharakaraman, Marino-Ramirez, Sheetlin, Landsman, Spouge	The Precise Positioning of Genomic Landmarks Can Aid in the Identification of Regulatory Elements
Jothi**,** Kann, Przytycka	Predicting Protein-Protein Interaction by Searching Evolutionary Tree Automorphism Space
Zheng, Lenert, Sankoff	Reversal Distance for Partially Ordered Genomes
Yamanishi**,** Vert, Kanehisa	Supervised Enzyme Network Inference from the Integration of Genomic Data and Chemical Information
Brejova**,** Brown, Li, Vinar	ExonHunter: A Comprehensive Approach to Gene Finding
Nabieva**,** *Jim,* Agarwal, Chazelle, Singh	Whole-proteome Prediction of Protein Function via Graph-theoretic Analysis of Interaction Maps
Apostolico, Comin**,** Parida	Conservative Extraction of Over-represented Extensible Motifs
Hu, Yan, Huang, Han, Zhou	Mining Coherent Dense Subgraphs Across Massive Biological Networks for Functional Discovery
Beissbarth, Tye-Din, Smyth, Speed, Anderson	A Systematic Approach for Comprehensive T-cell Epitope Discovery Using Peptide Libraries
Yu, Chen	Bayesian Neural Network Approaches to Ovarian Cancer Identification from High-resolution Mass Spectrometry Data
Ralaivola, Chen, Phung, Bruand, Swamidass, Baldi	Kernels for Small Molecules and the Prediction of Mutagenicity, Toxicity, and Anti-Cancer Activity

Tutorials 2005

Of the 42 tutorials proposed, the following 14 were presented:

FACILITATOR	TITLE OF TUTORIAL
Olga Troyanskaya	Integration and Analysis of Diverse Genomic Data
Ernst Wit and John McClure	Optimal Design and Analysis of Microarray Experiments
Peter Clote	RNA: Algorithms for Structure Prediction and Gene-finders
Martin Gollery	Developing and Using Special Purpose Hidden Markov Model Databases
Corinna Cortes and Mehryar Mohri	Weighted Finite-State Transducers in Computational Biology
Joanne Luciano and Jeremy Zucker	Semantic Aggregation, Integration and Inference of Pathway Data
Ruud van der Pas and Karl V. Steiner	Attacking Performance Bottlenecks
Suzanna Lewis	Principles of Ontology Construction
Daniel Huson	Introduction to Phylogenetic Networks
Hagit Shatkay	Mining the Biomedical Literature : State of the Art, Challenges and Evaluation Issues
Steffen Möller and Robert Hoffman	Gene Expression Levels as Traits in Genetic Linkage Analysis
Andrew Boyd and Abhijit Bose	A Bioinformatics Introduction to Cluster Computing
Iosif Vaisman	Computational Geometry of Protein Structure and Function
Gyan Bhanot and Bob Germain	A Massively Parallel High Performance Computing Environment for Computational Biology