DIMACS Center
Rutgers University

Special Focus on Information Processing in Biology

Annual Report

May 2006

Participants who spent 160 hours or more

PI: Fred Roberts, DIMACS

Participants who spent less than 160 hours

Ron Levy, BioMaPS, Rutgers University, Special Focus Co-Organizer

Wilma Olson, Center for Molecular Biophysics and Biophysical Chemistry, Rutgers University, Special Focus Co-Organizer

Eduardo Sontag, BioMaPS, DIMACS, Rutgers University, Special Focus Co-Organizer

BioMaPS/DIMACS/MBBC/PMMB/SYCON Short Course: Molecular Mechanisms and Models of Bacterial Signal Transduction
June 6 - 10, 2005

Organizers:
Eduardo Sontag, Rutgers University
Ann Stock, UMDNJ/HHMI

Workshop on Information Processing by Protein Structures in Molecular Recognition
June 13 - 14, 2005

Organizers:
Bhaskar DasGupta, University of Illinois at Chicago
Jie Liang, University of Illinois at Chicago

Workshop on Detecting and Processing Regularities in High Throughput Biological Data
June 20 - 22, 2005

Organizer:
Laxmi Parida, IBM T J Watson Research

Workshop on Machine Learning Approaches for Understanding Gene Regulation
August 15 - 17, 2005

Organizers:
Christina Leslie, Columbia University
Chris Wiggins, Columbia University

Working Group on DNA Barcode of Life
September 26, 2005

Organizers:
Rebecka Jornsten, Rutgers University
David Madigan, Rutgers University

Fred Roberts, DIMACS

Working Group on Evolution of Gene Regulatory Logic
January 6 - 8, 2006

Organizers:
Tanya Berger-Wolf, University of Illinois, Chicago
David Krakauer, Santa Fe Institute

Workshop on Data Mining, Systems Analysis, and Optimization in Neuroscience
February 15 - 17, 2006

Organizers:
W. Art Chaovalitwongse, Rutgers University
Leonidas D. Iasemidis, Arizona State University
Panos Pardalos, University of Florida

Short Course: A Field Guide to GenBank and NCBI Molecular Biology Resources
March 2, 2006

Organizer:
Nicholas Beckloff, UMDNJ

Co-Organizers:
Tamar Barkay, Rutgers University
Paul Ehrlich, BIOMAPS Institute
Mel Janowitz, DIMACS
Tara Matise, Rutgers University

DARPA Workshop on State-Dependent Delays in Regulatory Networks
March 2 - 3, 2006

Organizers:
Tim Buchman, Washington University
Jon Lorsch, John Hopkins University
Konstantin Mischaikow, Georgia Institute of Technology

Workshop on Computational/Experimental Approaches to Protein Defects in Human Disease
April 20 - 21, 2006

Organizers:
Jean Baum, Rutgers University
Barbara Brodsky, UMDNJ

Workshop: Sequence, Structure and Systems Approaches to Predict Protein Function
May 3 - 5, 2006

Organizers:
Anna Panchenko, NIH
Teresa Przytycka, NIH
Mona Singh, Princeton University

Workshop: Clustering Problems in Biological Networks
May 9 - 11, 2006

Organizers:
Sergiy Butenko, Texas A&M
W. Art Chaovalitwongse, Rutgers University
Panos Pardalos, University of Florida

Short Course: Exploring 3D Molecular Structures Using NCBI Tools
May 16, 2006

Organizers:
Tamar Barkay, Rutgers University
Nicholas Beckloff, UMDNJ
Paul Ehrlich, BIOMAPS Institute
Mel Janowitz, DIMACS
Tara Matise, Rutgers University

Workshop: The DNA Barcode Data Analysis Initiative (DBDAI): Developing Tools for a New Generation of Biodiversity Data
July 6 - 8, 2006

Organizers:
Javier Cabrera, Rutgers University
Fred Roberts, DIMACS
David Schindel, National Museum of Natural History
Michel Veuille, Muséum National d'Histoire Naturelle

Workshop: Machine Learning Techniques in Bioinformatics
July 11 - 12, 2006

Organizers:
Dechang Chen, Uniformed Services University of the Health Sciences
Xue-Wen Chen, University of Kansas
Sorin Draghici, Wayne State University

Working Group on Computational Tumor Modeling
August 2, 2006

Organizers:

David Axelrod, Rutgers University
Thomas S. Deisboeck, Harvard Medical School

Workshop on Computational Tumor Modeling
August 3 - 4, 2006

Organizers:

David Axelrod, Rutgers University
Thomas S. Deisboeck, Harvard Medical School

Visitors who have undertaken research under support of the project:

Boris Mirkin, Birkbeck College
12-9/05-12/21/05

Other Collaborators

Tamar Barkay, Rutgers, Co-Organizer, Short Course: A Field Guide to GenBank and NCBI Molecular Biology Resources

Nicholas Beckloff, UMDNJ, Organizer, Short Course: A Field Guide to GenBank and NCBI Molecular Biology Resources

Petra Berenbrink, Simon Fraser, Co-Organizer, Workshop on Biomolecular Networks: Topological Properties and Evolution

Tanya Berger-Wolf, University of Illinois, Chicago, Co-organizer, Working Group on Evolution of Gene Regulatory Logic

Tim Buchman, Washington University, Co-organizer, DARPA Workshop on State-Dependent Delays in Regulatory Networks

Javier Cabrera, Rutgers University, Co-Organizers Workshop: The DNA Barcode Data Analysis Initiative (DBDAI): Developing Tools for a New Generation of Biodiversity Data

W. Art Chaovalitwongse, Rutgers, Co-Organizer Workshop on Data Mining, Systems Analysis, and Optimization in Neuroscience

Dechang Chen, Uniformed Services University of the Health Sciences, Co-Organizer Workshop: Machine Learning Techniques in Bioinformatics

Xue-Wen Chen, University of Kansas, Co-Organizer Workshop: Machine Learning Techniques in Bioinformatics

Bhaskar DasGupta, University of Illinois at Chicago, Co-Organizer, Workshop on Information Processing by Protein Structures in Molecular Recognition

Sorin Draghici, Wayne State University, Co-Organizer Workshop: Machine Learning Techniques in Bioinformatics

Paul Ehrlich, BIOMAPS Institute, Co-Organizer, Short Course: A Field Guide to GenBank and NCBI Molecular Biology Resources

Leonidas D. Iasemidis, Arizona State University, Co-Organizer Workshop on Data Mining, Systems Analysis, and Optimization in Neuroscience

Mel Janowitz, DIMACS, Co-Organizer, Short Course: A Field Guide to GenBank and NCBI Molecular Biology Resources

Rebecka Jornsten, Rutgers, Co-Organizer, Working Group on DNA Barcode of Life

David Krakauer, Santa Fe Institute, Co-organizer, Working Group on Evolution of Gene Regulatory Logic

Christina Leslie, Columbia University, Co-Organizer, Workshop on Machine Learning Approaches for Understanding Gene Regulation

Jie Liang, University of Illinois at Chicago, Co-Organizer, Workshop on Information Processing by Protein Structures in Molecular Recognition

Jon Lorsch, John Hopkins University, Co-organizer, DARPA Workshop on State-Dependent Delays in Regulatory Networks

David Madigan**,** Rutgers, Co-Organizer, Working Group on DNA Barcode of Life

Tara Matise, Rutgers, Co-Organizer, Short Course: A Field Guide to GenBank and NCBI Molecular Biology Resources

Konstantin Mischaikow, Georgia Institute of Technology, Co-organizer, DARPA Workshop on State-Dependent Delays in Regulatory Networks

Panos Pardalos, University of Florida Co-Organizer Workshop on Data Mining, Systems Analysis, and Optimization in Neuroscience

Laxmi Parida, IBM T J Watson Research, Co-Organizer, Workshop on Detecting and Processing Regularities in High Throughput Biological Data

Cenk Sahinalp, Simon Frasier University, Co-Organizer, Workshop on Biomolecular Networks: Topological Properties and Evolution

David Schindel, Smithsonian Institution, Co-Organizers Workshop: The DNA Barcode Data Analysis Initiative (DBDAI): Developing Tools for a New Generation of Biodiversity Data

Eduardo Sontag, Rutgers University, Co-Organizer, BioMaPS/DIMACS/MBBC/PMMB/SYCON Short Course: Molecular Mechanisms and Models of Bacterial Signal Transduction

Ann Stock, UMDNJ/HHMI, Co-Organizer, BioMaPS/DIMACS/MBBC/PMMB/SYCON Short Course: Molecular Mechanisms and Models of Bacterial Signal Transduction

Chris Wiggins, Columbia University, Co-Organizer, Workshop on Machine Learning Approaches for Understanding Gene Regulation

Michel Veuille, Muséum National d'Histoire Naturelle, Co-Organizers Workshop: The DNA Barcode Data Analysis Initiative (DBDAI): Developing Tools for a New Generation of Biodiversity Data

Partner Organizations

Telcordia Technologies: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

AT&T Labs - Research: Collaborative Research; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research.

NEC Laboratories America: Collaborative Research; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research.

Lucent Technologies, Bell Labs: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

Princeton University: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning and workshop/working group organization.

Avaya Labs: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

HP Labs: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research and workshop/working group organization.

IBM Research: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research and workshop/working group organization.

Microsoft Research: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research and workshop/working group organization.

Stevens Institute of Technology: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning and research.

Georgia Institute of Technology: Collaborative Research

Partner organization of DIMACS. Individuals from the organization participated in the program planning and workshop/working group organization.

Sante Fe Institute: Collaborative Research.
Individuals from the organization participated in the program planning and working group/workshop organization.

The National Museum of Natural History, Paris, France: Collaborative Research.
Individuals from the organization participated in the program planning and working group/workshop organization.

Smithsonian Institution: Collaborative Research.
Individuals from the organization participated in the program planning and working group/workshop organization.

National Institutes of Health: Collaborative Research; Personnel Exchanges.
Individuals from the organization participated in the program planning and working group/workshop organization. Provided funding for a workshop.

Defense Advanced Research Projects Agency (DARPA): Collaborative Research
Provided funding for a workshop.

Activities and Findings

Overview

This special focus is jointly sponsored by the Center for Discrete Mathematics and Theoretical Computer Science (DIMACS), The Biological, Mathematical, and Physical Sciences Interfaces Institute for Quantitative Biology (BioMaPS), and the Rutgers Center for Molecular Biophysics and Biophysical Chemistry (MB Center). It is a follow up on DIMACS' highly successful special foci on "Computational Molecular Biology" and "Mathematical Support for Molecular Biology."

Increasingly, many aspects of biology can be viewed as involving the processing of information. Modern information and computer science have played an important role in such major biological accomplishments as the sequencing of the human genome. On the other hand, biological ideas can inspire new concepts and methods in information science. This special focus is motivated by these two observations. The special focus activities are organized around a series of workshops with four themes:

- Algorithmic Approaches to Biological Information Processing
- Computer Science, Engineering and Biology: Applications and Analogies
- Biological Circuits and Cellular Signaling
- Proteomics.

Two of these themes represent approaches and two represent areas of application of these approaches.

Theme 1: Algorithmic Approaches to Biological Information Processing.

A major theme of the special focus revolves around algorithms for biological information processing. We take two points of view here. One involves how biological organisms use “algorithms” to process information and another involves how we use algorithmic methods to understand how organisms process information. The two points of view are interrelated and are reflected in three workshops.

Understanding information processing in the biological organism involves dealing with huge data sets. Modern algorithmic methods for dealing with such data sets, especially algorithms involved in pattern recognition, learning, cluster analysis, and, generally speaking, data mining, are especially relevant. Biological information processing takes advantage of regularities such as repetition, structural motifs and patterns, clustering, etc. Understanding such biological processes might, by analogy, lead us to new data mining algorithms and, in turn, methods of data mining might be useful in understanding how organisms process such regularities.

The massive amounts of information gathered in recent years has made it possible to study complex cellular networks using algorithmic methods of data analysis and information science. Predictions about the structure and behavior of gene regulatory networks provide a major challenge for this kind of approach. Modern methods of machine learning are especially appropriate given the nature of the data -- copious but noisy and incomplete -- and also provide tools that have been a major area of research at DIMACS.

Theme 2: Computer Science, Engineering, and Biology: Applications and Analogies.

The study of analogies between information processing in biology and information processing in computer science and engineering offers promise for the understanding of both and we investigate these analogies. More generally, we investigate applications of ideas from the biological sciences in computer science and engineering and vice versa. Such analogies and applications are a second major theme of the special focus.

Theme 3: Biological Circuits and Cellular Signaling.

Biochemical networks in the cell are responsible for processing environmental signals, inducing appropriate cellular responses, and sequencing internal events such as gene expression. Through elaborate mechanisms, they allow cells and entire organisms to perform their basic functions. A third theme of the special focus is the elucidation of the function and role of biological circuits and cellular signaling, with an eye to how non-biological networks can be applied to biological ones and vice versa.

Theme 4: Proteomics.

The fourth theme of the special focus revolves around proteomics. We seek to build on the knowledge gained from genomics to understand the activities and interactions of proteins in the cell. Studying the complete set of proteins expressed by the genome of an organism, cell or tissue type during its lifetime is a complex problem because the number of proteins is so large compared to the number of genes, because proteins can undergo numerous modifications, and because the makeup of the proteome changes frequently in response to the environment. Understanding how information encoded in the three-dimensional structures that underlie complex protein-DNA and protein-protein network interaction is one of the fundamental challenges of biology.

Tutorials, Workshops, and Working Groups During This Reporting Period

BioMaPS/DIMACS/MBBC/PMMB/SYCON Short Course: Molecular Mechanisms and Models of Bacterial Signal Transduction

Dates: June 6 - 10, 2005
Location: Busch Campus, Rutgers University
Organizers: Eduardo Sontag, Rutgers University, and Ann Stock, UMDNJ/HHMI
Attendance: 78

The course was a five-day intensive investigation of signal transduction divided into two related parts:

1. Basic introduction to signal transduction in bacteria for participants with extensive training in the mathematical, computational, and physical sciences but with a more limited background in molecular biology
2. Advanced reviews of current contributions to the understanding of bacterial signal transduction with an emphasis on computational approaches to modeling biological systems by leading scientists and their group members

The foundation required by non-expert researchers for the understanding of signal transduction was provided by five leaders in the field in a series of 2-hour presentations during the first half of the course. Stanislav Shvartsman of Princeton started the course with a basic description of signal transduction including an outline of chemical pathways and their biological significance. Igor Zhulin of the Georgia Institute of Technology followed with a more detailed description of signal transduction in bacteria. Ann Stock of the University of Medicine and Dentistry of NJ focused on bacterial "two-component" proteins involved in phosphotransfer signaling systems, a general mechanism of signal transduction that is widespread throughout nature. Various aspects of bacterial motility and chemotaxis were covered by Robert Bourret of the University of North Carolina and Ned Wingreen of Princeton University. These introductory lectures laid the groundwork for seminars on current research in signal transduction. The lectures were designed to provide participants with a limited knowledge of molecular biology a smooth transition to the understanding and appreciation of cutting-edge research.

In the remaining presentations, participants gained an in-depth view of signal transduction both from the content of the presentations and the perspectives provided by both molecular biologists and physicists. Speakers include Bonnie Bassler (Princeton University), William Bialek (Princeton University), Mark Goulian (University of Pennsylvania), Tom Silhavy (Princeton University), and Alexander van Oudenaarden (MIT). In addition to his introductory lecture, Robert Bourret (University of North Carolina) offered advice on the important issue of achieving effective collaborations between experimentalists and modelers.

Workshop on Information Processing by Protein Structures in Molecular Recognition

Dates: June 13 - 14, 2005
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Bhaskar DasGupta, University of Illinois at Chicago, Jie Liang, University of Illinois at Chicago
Attendance: 35

Biological processes in cells are based on specific molecular recognitions, which triggers cascade of biological responses. The physical basis of complex network interaction is the three-dimensional structure of proteins and their functional regions. Understanding how information encoded in these biomolecules is recognized and processed by the interacting partners is a fundamental problem of biology.

In this workshop we discussed the development of algorithms for discovery of spatial patterns important for recognition, for uncovering deep evolutionary relationship of proteins, for predicting binding partners, and for simulating the protein-protein and protein-DNA recognition process. Specific topics of interest included protein-ligand and protein-protein binding site prediction, functional prediction of proteins with known structures but unknown functions, protein-protein interactions and docking, prediction of immune epitope, design of peptide modulators of protein-protein interactons, protein substructure matching, and evolution of structural biopattern. We hope further development in these areas formulated new research problems and motivate new algorithms in combinatorics, optimization, discrete mathematics, mathematical programming, and additional areas.

Workshop on Detecting and Processing Regularities in High Throughput Biological Data

Dates: June 20 - 22, 2005
Location: DIMACS Center, CoRE Building, Rutgers University
Organizer: Laxmi Parida, IBM T J Watson Research
Attendance: 62

The biological community is being inundated with a large amount of data and understanding this data is lagging behind the process of acquiring it. It is believed nature has left vital clues hidden in this data and there is a need for techniques and methodologies to work effectively in detecting these. Biological information processing exploits these regularities to gain understanding of the underlying model or phenomenon. For example, in its simplest form regularity could be repetition of functional or structural domains in a protein sequences or co-expression of genes in microarrays. When the data is in terms of networks, either representing protein-protein interactions or metabolic pathways, topological motifs tell a tale that will be fundamental in understanding the working of a biological system. The workshop aimed to contribute significantly to the research effort by bringing together researchers from the many different groups engaged in biological projects having the study of regularities in the data as an underlying theme.

List of Keynote Speakers:

- Alberto Apostolico, Purdue University and University of Padova
- Andrea Califano, Columbia University
- Bud Mishra, New York University
- David Mount, University of Arizona
- Andrey Rzhetsky, Columbia University
- David Sankoff, University of Ottawa

Workshop on Machine Learning Approaches for Understanding Gene Regulation

Dates: August 15 - 17, 2005
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Christina Leslie and Chris Wiggins, Columbia University
Attendance: 51

Over the last decade, biology has been transformed into a data-driven science. Through innovations in sequencing, high-throughput microscopy, mRNA expression arrays, protein-protein and protein-DNA binding assays, and numerous other high-throughput methods, it is now possible to query simultaneously the activities of thousands of genes and their products under a wide variety of experimental conditions.

The resulting data pose an exciting challenge for the field of machine learning. Many of the model organisms (most notably S. cerevisiae) are of sufficient complexity to render detailed mathematical modeling intractable. However, it is still possible to try to learn quantitative models that are rich enough to fit data, yet simple enough to generalize and to be interpretable. Work by numerous groups suggests a promising future for more complex eukaryotes (e.g., C. elegans, S. pombe, or D. melanogaster).

Qualitatively new challenges to the machine learning community include the integration of heterogeneous datasets, such as sequence, binding, and expression data; the creation of models which are interpretable even to those not trained in probabilistic reasoning or statistical learning theory; and the presentation of the resulting models in a way useful to bench biologists as well as computational biologists.

This three-day workshop encouraged interaction among innovators in computational biology and innovators in machine learning; illuminated recent successes as well as pressing challenges; and hopefully will inspire the development of novel, biologically relevant, and biologically interpretable machine learning approaches to the current problems in biology.

Working Group on DNA Barcode of Life

Dates: September 26, 2005
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Rebecka Jornsten, Rutgers University, David Madigan, Rutgers University, and Fred Roberts, DIMACS
Attendance: 25

The "Barcode of Life" project aims to create a unique "signature" for every species using a particular gene in the mitochondrial genome. See: http://www.barcodinglife.org/. The project has gathered steam over the last couple of years and has a rapidly growing database containing over 30,000 sequences. Many data analysis and modeling challenges now present themselves.

At this exploratory workshop, we did some brainstorming on these challenges and possible research directions. This was an informal event with time for hands-on work with some sequence data and discussion time.

Working Group on Evolution of Gene Regulatory Logic

Dates: January 6 - 8, 2006
Location: Santa Fe Institute, Santa Fe, New Mexico
Organizer: Tanya Berger-Wolf, University of Illinois, Chicago; David Krakauer, Santa Fe Institute
Attendance: 17

The Evolution of Gene Regulatory Logic working group was convened to discuss the idea that the discovery of the near "universal" or canonical character of the genetic code in 1953 by Watson and Crick, demonstrated that beneath the extensive diversity of forms in nature, there are conserved regularities that have made possible the development of a general science of genetics and molecular biology. The elucidation of the mechanism of action of the prokaryotic operon by Jacob and Monod extended the investigation of regular features of genetic control to include a simple form of logic based on feedback principles. The discovery of the highly conserved homeogenes in eukaryotes revived interest in general principles of morphology, extending the notion of the Bauplan into that of the Zootype, and establishing a connection between development and gene regulation.

Current work on genetic regulatory networks and cis-regulatory logic extends the operon concept to include complex networks of feedback, Boolean logic, and regular sequence motifs or codes, collectively correlated with regular patterns of gene expression. The history of studies of gene regulation and phenotypic development have repeatedly uncovered fundamental mechanisms shared by distantly related lineages, and observed essentially similar modes of control. In this workshop we stepped back from an exclusively sequence level study of the evolution of gene regulation, and asked what forms of logic are instantiated in regulatory networks and how these vary among species. We interpret logic as some set of mechanisms for encoding computations that perform adaptive functions within and between cells. Another way to think about logic is as the set of rules and memory stores the cell uses to process information about its local environment.

Is there as Jacob suggested a single "Logic of Life" or multiple logics and if so how can we best characterize these logics in order to uncover the transformation series by which they come about throughout evolutionary history?

In this meeting we reviewed our current understanding of prokaryotic and eukaryotic gene regulation and established similarities based on common modes of operation. We explored the value of regulatory architectures conceived in the engineering and computational realms when applied to biology, and explored the possibility of new hybrid forms of logic derived from biology that often possess fascinating robustness and evolveability properties. The meeting brought together biologists and computer scientists who seek to understand the logic underpinning the construction of complex adaptive functions.

Workshop on Data Mining, Systems Analysis, and Optimization in Neuroscience

Dates: February 15 - 17, 2006
Location: University of Florida, Gainseville, Florida
Organizers: W. Art Chaovalitwongse, Rutgers University; Leonidas D. Iasemidis, Arizona State University; Panos Pardalos, University of Florida
Attendance: 63

The human brain is among the most complex systems known to mankind. Neuroscientists seek to understand brain function through detailed analysis of neuronal excitability and synaptic transmission. Only in the last few years has it become feasible to capture simultaneous responses from large enough numbers of neurons to empirically test the theories of human brain function. Experimental neuroscience methods have resulted in massive amounts of data, but traditional data-processing and quantitative methods are not sophisticated enough to exploit this new flood of information. There is an increasing number of modern research efforts in data mining, systems analysis and optimization research to advance methods needed to process the large spatial and temporal data arising in quantitative neuroscience. This conference explored these new methods.

A major area of interest is the study of how neuronal circuitries of the brain support its cognitive and functioning capacities at a descriptive level of the molecular mechanisms of synaptic plasticity. Advances in the fields of signal processing, statistics, data mining and optimization have made it possible to discover and investigate complex patterns in the vast amount of information being generated by neuroimaging and neurophysiological signals. Research breakthroughs could lead to understanding more about diseases such as epilepsy, sleep disorders, movement disorders, and cognitive disorders that affect millions of people every year.

This workshop brought together a multi-disciplinary group to enable the sharing of knowledge, ideas, and techniques. This required, by necessity, collaboration among computer scientists, mathematicians, neurobiologists and clinicians. This workshop resulted in lively discussions of the cross-disciplinary research and open up a new question: How do we go from the gigabytes of experimental data that we now have to concise conclusions about the function of the brain? The answer to this question will revolutionize neuroscience research and give us a greater understanding of brain function.

Short Course: A Field Guide to GenBank and NCBI Molecular Biology Resources

Dates: March 2, 2006
Location: UMDNJ, Newark, NJ
Organizer: Nicholas Beckloff, UMDNJ
Co-Organizers: Tamar Barkay, Rutgers; Paul Ehrlich, BIOMAPS Institute; Mel Janowitz, DIMACS; Tara Matise, Rutgers
Attendance: 40

The University of Medicine and Dentistry of New Jersey (UMDNJ) hosted the National Center for Biotechnology Information workshop, "A Field Guide to GenBank and NCBI Molecular Biology Resources". This was a free workshop taught at UMDNJ, Newark, NJ on Thursday, March 2, 2006. A morning lecture that covered all of the resources of the NCBI was given followed by a hands-on workshop led by the NCBI instructors themselves. Each year new tools are added to the site so it is a great workshop for all users of the NCBI, new and old.

DARPA Workshop on State-Dependent Delays in Regulatory Networks (funded by DARPA)

Dates: March 2 - 3, 2006
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Tim Buchman, Washington University; Jon Lorsch, John Hopkins University; Konstantin Mischaikow, Georgia Institute of Technology
Attendance: 35

Multi-scale activity is a universal characteristic of biological processes. In particular, novel dynamics emerge from arrangements of components and events across different scales of distance and time that are intertwined to translate biomolecular events into recognizable phenotypes. While a quantitative understanding and predictive mathematical modeling of these complexities on the level of organisms is acknowledged to be a very long-range goal, there are opportunities for more modest but still fundamental steps. Pertinent to the modeling challenge is the notion of time. The multi-scale aspect of biology guarantees that the regulatory biological processes occur at different locations, times, and rates. Nevertheless, these processes achieve remarkable temporal and quantitative coordination. The observations suggest that accounting for - and adapting to - time delay is central to biological complexity and robustness.

Two key advances - one in biology and one in mathematics - suggest that there is an immediate opportunity for progress in this arena. In biology, new experimental tools that allow for reliable, comprehensive and serial assessments of the states and processes of gene regulatory networks have become widely available. In mathematics, new theoretical developments suggest the possibility to develop a theory of the global dynamics of nonlinear systems with multiple or variable delays.

The immediate goal of this workshop was to begin a dialogue between mathematicians with expertise in the dynamics of delay differential equations and control theory, and biologists with expertise in the mechanisms in signal transduction/gene regulatory networks. For the biologists the benefit was new models within which to understand, test, and control gene regulation. For the mathematicians the benefit was a concrete set of problems around which the development of this new theory can be focused.

Workshop on Computational/Experimental Approaches to Protein Defects in Human Disease

Dates: April 20 - 21, 2006

Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Jean Baum, Rutgers University and Barbara Brodsky, UMDNJ
Attendance: 93

This workshop brought together diverse perspectives of biophysical, computational and evolutionary scientists to understand the deleterious consequences of human disease mutations and the aggregation leading to neurodegenerative diseases. The study of amyloidogenesis and the structure of amyloid aggregates is leading to new theories and treatments, while the large amount of information on the human genome and other genomes has allowed computational and experimental approaches to comprehend why a given single nucleotide polymorphism may be deleterious and result in human disease. New computational approaches complement experimental results, making this workshop extremely timely. The workshop was designed to reach a very diverse audience consisting of roughly a 50:50 mix of experimentalists and computational scientists.

This workshop was sponsored jointly by DIMACS, the BioMAPS program, the Center for Molecular Biophysics and Biophysical Chemistry, and Robert Wood Johnson Medical School. DIMACS seeks to develop and populate the interface between the mathematical-physical sciences and molecular biology, while the mission of the BioMAPS Institute is to promote research and education at the interface between the **Bio**logical, **Ma**thematical, and **P**hysical **S**ciences. The Center for Molecular Biophysics and Biophysical Chemistry at Rutgers provides a focus for scholarly activities at the interface of the biological and physical sciences at Rutgers and Robert Wood Johnson Medical School coordinates a general academic program for graduate and undergraduate students.

Specific topics included:

- Understanding the mechanism of protein association leading to amyloidosis.
- Elucidation of the structure of the amyloid aggregates.
- Generation of native conformational states of proteins from amino acid sequence.
- Computational analyses of the effects of mutations on conformation and stability.
- Prediction of whether a SNP will be deleterious, leading to human disease.
- Consideration of compensated pathogenic mutations and their relation to disease.
- Prediction of disordered regions and the role of natively unfolded regions in diseases.

Workshop: Sequence, Structure and Systems Approaches to Predict Protein Function
Dates: May 3 - 5, 2006
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Anna Panchenko, NIH; Teresa Przytycka, NIH; Mona Singh, Princeton University
Attendance: 68

While more than two hundred complete genomes have been sequenced, a large fraction of genes in genomes do not have assigned functions, and elucidation of the biological roles of all unknown gene products remains an elusive goal. In addition to genomic sequence data, in recent years we have witnessed an explosion of structural data, along with large-scale protein network data. This has led to a number of complementary approaches to predict protein function based on heterogeneous data from diverse experimental sources. The goal of such methods is to decrease the amount of time-consuming experimental work necessary to interpret the complexity of genomes and proteomes. In general, protein function can be predicted by the analysis of specific conserved structural and sequence features, by transferring the annotation from experimentally characterized genes to their uncharacterized homologs , by genome context and cross-genomic analysis, and by analyzing proteins within the context of biological networks. Each of these methods faces unique computational and statistical challenges.

This workshop brought together biologists, computer scientists and mathematicians who work on various aspects of protein function prediction. This workshop provided both a venue for reviewing the current state-of-the-art of diverse methods as well as a platform for further cross-fertilization and integration of sequence, structure and systems approaches.

Workshop: Clustering Problems in Biological Networks

Dates: May 9 - 11, 2006
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Sergiy Butenko, Texas A&M; W. Art Chaovalitwongse, Rutgers University; and Panos Pardalos, University of Florida
Attendance: 74

Clustering techniques are essential to a wide variety of applications. Network clustering approaches are becoming common in the analysis of massive data sets arising in various branches of science, engineering, government and industry. In particular, network clustering techniques emerge as an important tool in computational biology, where they can be used for analysis of gene and protein networks and other important problems.

As an example, understanding gene expression and regulation is one of the major problems in biology. Network models have become common in this field, and clustering approaches play a central role in such models. In gene networks, the vertices correspond to genes and the edges represent functional relations between these genes that are identified using the comparative genomics methods. Solving clustering problems in gene networks allows the identification of groups of genes which have similar expression patterns. This information is crucial for understanding the nature of genetic diseases.

Similarly, in protein networks the proteins serve as nodes and nodes corresponding to two proteins are connected by an edge if they interact with each other. The importance of studying the protein networks is increasing as more information on protein interactions in various organisms is becoming available from the protein databanks. Some important properties of protein networks have been recently studied by a number of researchers. For example, it has been discovered that the degree distribution in such networks follows the power law - a property that has been observed in networks arising in a variety of diverse applications. This structure has important implication on the cell's survivability. Clustering models in protein networks are important for understanding the structure of protein interactions in a cell.

Due to a wide range of applications of network clustering techniques, a large part of the previously developed methodology can be transferred to the study of networks arising in biology. However, some of the clustering problems of interest in computational biology have their own specifics. This workshop provided a forum for leading as well as beginning researchers to discuss recent advances and identify current and future challenges arising in the research concerning clustering problems in computational biology.

Short Course: Exploring 3D Molecular Structures Using NCBI Tools

Dates: May 16, 2006
Location: DIMACS Center, CoRE Building, Rutgers University

Organizers: Tamar Barkay, Rutgers; Nicholas Beckloff, UMDNJ; Paul Ehrlich, BIOMAPS Institute; Mel Janowitz, DIMACS; Tara Matise, Rutgers
Attendance: 28

The National Center for Biotechnology Information (NCBI) will present Exploring 3D Molecular Structures Using NCBI Tools, a course including lectures and computer workshops on effectively using the NCBI databases, search services, analysis tools that focus on 3D macromolecular structure data.

After attending the course, participants should be able to do the following:

- Understand the origin and organization of 3D structural data and how these data are curated at NCBI
- Find structural neighbors using VAST and functional elements within structures using the Conserved Domain Database and RPS-BLAST
- Analyze a 3D structure, highlight features such as bound ligands and active site residues, create customized annotations, and save and export a figure
- Find and evaluate a 3D modeling template for a protein by creating multiple sequence alignments using either sequence or structure similarity searches

The course is intended for principal investigators, postdoctoral fellows, graduate students, advanced undergraduates, and other scientific staffs who either work with 3D structural data or are interested in understanding how to incorporate such data into their research. Participants do not need to have taken any other NCBI course to attend, and no prior experience with structural data is required. People who have attended The Field Guide course will find this an interesting follow-up.

Workshop: The DNA Barcode Data Analysis Initiative (DBDAI): Developing Tools for a New Generation of Biodiversity Data

Dates: July 6 - 8, 2006
Location: The National Museum of Natural History, Paris, France
Organizers: Javier Cabrera, Rutgers University; Fred Roberts, DIMACS; David Schindel, National Museum of Natural History; Michel Veuille, Muséum National d'Histoire Naturelle
Attendance: (Registration still open for this workshop)

In the past two years, a series of studies have been published in which "DNA barcoding" was proposed as a tool for differentiating species. Barcoding is based on the assumption that short gene regions evolve at a rate that produces clear interspecific sequence divergence while retaining low intraspecific sequence variability. The cytochrome c oxidase subunit 1 mitochondrial region ("COI") has emerged as a suitable barcode region for most animals. Taxonomists are in the process of identifying appropriate gene regions for barcoding other major groups of eukaryotes. Taxonomic studies of a growing number of taxa have shown that the discontinuity in the levels of barcode sequence divergence (both phenetic and diagnostic) match the species boundaries as delineated by morphological and ecological characters. These studies set the stage for a more in-depth analysis of the relationship between DNA barcode patterns and our understanding of speciation processes and mitochondrial evolution.

The Consortium for the Barcode of Life (CBOL; see www.barcoding.si.edu) is an international consortium of about 70 Member Organizations from six continents and more than 35 nations. These include natural history museums, herbaria, biodiversity and conservation organizations, university departments and other research organizations, government agencies and private sector companies. CBOL is devoted to exploring and developing the potential of DNA barcoding to become a tool for taxonomic research and for applications of species-level data to applied problems such as conservation, crop protection and sustainable development. Four Working Groups have been formed by CBOL, including the Data Analysis Working Group (DAWG) chaired by Dr. Michel Veuille, Director of the Department of Systematics and Evolution in the National Museum of Natural History, Paris.

This workshop is part of the DNA Barcode Data Analysis Initiative (DBDAI), a 24-month international interdisciplinary program of work, sponsored by CBOL and DAWG, that will bring together taxonomists, population geneticists, statisticians, applied mathematicians and computer scientists. The overarching goals of this program of work will be to better understand the relationship between DNA barcode data and population-level genetic processes, and to develop the analytical tools needed to interpret, analyze and archive DNA barcode data. A further goal will be to explore both the potential and limitations of barcoding in the study of natural populations, especially populations of pests and endangered species. This workshop will address research challenges posed by the DAWG, in collaboration with DIMACS, which is conducting the US portion of DAWG. For more information about these challenges, see http://dimacs.rutgers.edu/Workshops/BarcodeResearchChallenges/.

At this workshop, preliminary ideas for approaches to the data analysis challenges involved in DNA barcoding will be presented. Work Teams will submit abstracts of their preliminary results to the Steering Committee, from which participants in the workshop will be selected. Presenters will receive feedback from the Steering Committee and other workshop participants.

Workshop: Machine Learning Techniques in Bioinformatics

Dates: July 11 - 12, 2006
Location: DIMACS Center, CoRE Building, Rutgers University
Organizer: Dechang Chen, Uniformed Services University of the Health Sciences; Xue-Wen Chen, University of Kansas; and Sorin Draghici, Wayne State University
Attendance: (Registration still open for this workshop)

Bioinformatics aims to solve biological problems by using techniques from mathematics, statistics, computer science, and machine learning. Recent years have observed the essential use of these techniques in this rapidly growing field. Examples of such applications include those to gene expression data analysis, gene-protein interactions, protein folding and structure prediction, genetic and molecular networks, sequence and structural motifs, genomics and proteomics, text mining in bioinformatics, and so on. Bioinformatics provides opportunities for developing novel machine learning techniques; and machine learning plays a key role in advancing bioinformatics. The workshop is devoted to computational challenges of important biological problems. The goal of this workshop is to bring together researchers in both machine learning and bioinformatics to discuss state-of-the-art machine learning algorithms and their applications to various tasks in bioinformatics.

Working Group on Computational Tumor Modeling (partially funded by NIH)

Dates: August 2, 2006
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: David Axelrod, Rutgers University, and Thomas Deisboeck, Harvard Medical School
Attendance: (Registration still open for this working group)

Computational methods are coming to have intriguing roles in the analysis of diseases such as cancer, for example in understanding how genetic mutations affect multicellular behavior or how spatial-temporal patterns of angiogenesis impact the efficacy of cancer therapies. Due to the inherent non-linearity and complexity of the many networked physiological processes involved on the cellular level alone, conventional reductionism-driven approaches fail in providing answers to such questions. Given the multi-scaled patho-physiology involved, it is becoming abundantly clear that cancer research requires a cross-disciplinary, complex, systems science approach, in which innovative multi-scaled computational cancer models play a central role. Ultimately, these “systems biology” models will allow cancer researchers to properly study such critical, interconnected tumorigenesis processes as initiation, progression, invasion, angiogenesis and metastasis. The potential applications for algorithms that capture these processes therefore range from experimental clinical cancer research, such as treatment outcome predictions, to virtual trials for the pharmaceutical industry. This working group meeting will present a variety of relevant computational tumor models and algorithms, covering several scales of interest by starting from the genetic instability and the functional genomics level up to tumor cell invasion and the angiogenesis level. Work on tumor cell signaling and information processing, multicellular pattern formation and scaling laws will be discussed as well. Finally, the meeting will also focus on several key challenges related to cancer modeling, such as biomedical data acquisition, access and quality, as well as the pros and cons of combining different (e.g., discrete and continuous) modeling approaches.

The Computational Tumor Modeling Working Group is part of The National Cancer Institute's Integrative Cancer Biology Program, a new initiative in systems biology. The goal of this initiative is to promote the analysis of cancer as a complex biological system, with the ultimate goal of developing reliably predictive computational models of various cancer processes, facilitating the development of cancer interventions. This will be achieved through the integration of experimental and computational approaches towards the understanding of cancer biology. One way of achieving this goal is for the National Cancer Institute to sponsor a series of Workshops that bring together cancer biologists and scientists from fields such as mathematics, physics, information technology, imaging sciences, and computer science.

This Computational Tumor Biology Working Group at DIMACS will bring together scientists who are already working in the highly interdisciplinary field of mathematical and computational cancer modeling and simulation as well as “newcomers” who are interested in this exciting research area. The group's main goal is to introduce and discuss innovative concepts, algorithms and platforms and to establish new cross-disciplinary collaborations. The focus will be on multiscale modeling as well as on data integration and visualization techniques, i.e. on the challenging interface between experiment and theory. The meeting is also intended to introduce the Center for the Development of a Virtual Tumor (CViT), one of the National Cancer Institute's funded Integrative Cancer Biology Programs. CViT's long term aim is to develop a module-based toolkit for cancer research.

Workshop on Computational Tumor Modeling (Partially funded by NIH)

Dates: August 3 - 4, 2006
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: David Axelrod, Rutgers University, and Thomas Deisboeck, Harvard Medical School
Attendance: (Registration still open for this workshop)

This workshop will cover the topics of the working group that meets immediately before.

Findings

Protein Function Annotation Based on Ortholog Clusters Extracted from Incomplete Genomes Using Combinatorial Optimization

Reliable automatic protein function annotation requires methods for detecting orthologs with known function from closely related species. While current approaches are restricted to finding ortholog clusters from complete proteomes, most annotation problems arise in the context of partially sequenced genomes. Akshay Vashist, graduate student, Computer Science, Rutgers University worked with Casimir Kulikowski, Computer Science, Rutgers University and Ilya Muchnik, DIMACS, to develop a combinatorial optimization method for extracting candidate ortholog clusters robustly from incomplete genomes. Their proposed algorithm focuses exclusively on sequence relationships across genomes and finds a subset of sequences from multiple genomes where every sequence is highly similar to other sequences in the subset. Vashist, Kulikowski, and Muchnik then used an optimization criterion similar to the one for finding ortholog clusters to annotate the target sequences. They reported on a candidate annotation for proteins in the rice genome using ortholog clusters constructed from four partially complete cereal genomes - barley, maize, sorghum, wheat and the complete genome of *Arabidopsis* at the 10th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2006) in Venice, Italy.

Sequence Information Searching

Graham Cormode, Bell Labs, and Muthu Muthukrishnan, Computer Science, Rutgers, initiated a new class of research referred to as "Substring Compression Problems." These problems take a sequence S and search for subsequences which are the most/least compressible under a variety of standard compression techniques, as well as using this measure to compute similarity between pairs of sequences. Since it has been observed that compressibility is a strong indicator for biological function, this has immediate application to exploration of DNA coding sequences (e.g. newly discovered viruses), biological sequence comparison, etc. They presented the first known, nearly optimal algorithms for substring compression problems: Substring Compression Queries, Least/MostCompressible Sequences and their generalizations, that are exact or provably approximate. Their exact algorithms exploit the structure in strings via suffix trees and their approximate algorithms rely on new relationships between Lempel-Ziv compression and string parsings.

Extracting Fixed-length Features from Variable-length Sequences

Pai-Hsi Huang received a DIMACS winter Graduate Student award for the year 2006 supporting his research with Vladimir Pavlovic, Computer Science, Rutgers, on extracting fixed-length features from variable-length sequences. Huang followed the framework proposed by Tommi Jaakkola: combining a generative model (hidden Markov model), which is excellent in modeling variable-length sequences, with a discriminative model (logistic regression). Huang searched for the "correct," or "effective" features to extract, namely, which set of features are best for the discriminative model. His preliminary results indicate that the most probable path, also known as the Viterbi path, does not yield very good features. Huang realized that the Viterbi path has large variance, in other words, perturbing the sequence slightly might cause the path to change significantly. This motivated the study of another set of features: the sufficient statistics. Huang's results show that the sufficient statistics have great resistance to the problem of an imbalanced dataset (one class dominates the other), whereas for the original features proposed by previous researchers, the gradient of the log-likelihood of the sequence suffer significantly from the imbalanced dataset problem. Huang has proposed a linear dimensionality reduction technique (using prior knowledge) that allows a compact representation of the examples. This allows the reduction of the dimensionality of a problem from p_*S, where S is the cardinality of the alphabet set, which is 20 in Huang's specific problem, to p_*4, where p is the length of the generative model, typically on the order of 100. This is an improvement on the (nonlinear) dimensionality reduction technique proposed by Jaakkola, et. al., and reduces the dimensionality of the problem from p_*S to p_*9. Huang and Pavlovic have submitted their results to the ISMB2006 (International Conference on Intelligent Systems for Molecular Biology).

Synthetic Variables in Data Analysis

In many information processing problems in biology, datasets contain large amounts of redundant variables, and numerous “feature selection” methods have been devised to eliminate them. Irina Lozina, a Rutgers graduate student, worked with Peter Hammer, RUTCOR, Rutgers University on problems with binary (0,1) data, and examined the effect of introducing an addition to the original variables, sets of artificial variables, representing logical compositions of the given variables. Lozina associated to every subset of k variables all 2^2^k Boolean functions, creating in this way a relatively large set of artificial variables. In order to reduce the size of this set she “filtered” it, retaining only the best 20 “new” variables using various criteria. The process was iterated, considering now the dataset as including not only the original variables, but also the “new” ones. Lozina applied this procedure to several biomedical datasets from the publicly available datasets of University of California at Irvine’s repository. The major conclusion of the study is that the artificial variables obtained after 3, 4 or 5 iterations have a very high correlation with the outcome, frequently exceeding 85%. Since the artificial variables obtained in this way are in fact Boolean functions approximating closely nature’s “hidden functions”, the proposed procedure promises to become a useful tool in data mining.

Inferring (Biological) Signal Transduction Networks via Transitive Reductions of Directed Graphs

Reka Albert, Pennsylvania State University, Bhaskar DasGupta, University of Illinois at Chicago, Riccardo Dondi, Univ. Milano–Bicocca, and Eduardo Sontag, Rutgers University, considered the binary transitive reduction (BTR) problem that arises in inferring a sparsest possible (biological) signal transduction network consistent with a set of experimental observations with a goal to minimize false positive inferences even if risking false negatives. Special cases of BTR have been investigated before in different contexts; the best previous results are as follows: (1) the minimum equivalent digraph problem, that correspond to the special case of BTR with no critical edges and all edges labels being zeroes, is known to be MAX-SNP-hard, admits a polynomial time algorithm with an approximation ratio of $1.617 + \epsilon$ for any constant $\epsilon > 0$ and can be solved in linear time for directed acyclic graphs. (2) a 2-approximation algorithm exists for the special case of BTR in which all edge labels are zeroes. Albert, DasGupta, Dondi, and Sontag’s contributions include: observing that the BTR problem can be solved in linear time for directed acyclic graphs; providing a 1.78-approximation for the restricted version of BTR when all edge labels are zeroes (the same restricted version as in (2) above); and providing a 2+o(1)-approximation for BTR on general graphs.

Inapproximability Results for the Lateral Gene Transfer Problem

Bhaskar Dasgupta, University of Illinois at Chicago, Sergio Ferrarini, Politecnico de Milano, Uthra Gopalakrishnan, University of Illinois at Chicago, and Nisha Raj Paryani, University of Illinois at Chicago, established some inapproximability results for the Lateral Transfer Problem. This optimization problem, which was defined by Hallet and Lagergren, is that of finding the most parsimonious lateral gene transfer scenario for a given pair of gene and species trees. Dasgupta, Ferrarini, Gopalakrishnan, and Paryani proved that the Lateral Transfer Problem is MAX SNP-hard; thus a Polynomial Time Approximation Scheme is not possible for it unless P = NP.

Outreach Activities

This project is closely intertwined with DIMACS efforts to link mathematics and computer science with biology in the high schools. The project organizers were involved in planning a DIMACS conference on this subject in April 2006 (see http://dimacs.rutgers.edu/Workshops/Biomath/). Also, the project organizers are working closely with the Summer 2006 DIMACS Bio-math Connect Institute (BMCI), which is aimed at introducing high school math/CS and Bio teachers to topics at the interface. This project is informing the BMCI effort and specific topics from the project are being adapted for use in BMCI.

Books

Papers

R. Albert, B. DasGupta, R. Dondi and E. Sontag, "Inferring (biological) signal transduction networks via transitive reductions of directed graphs," DIMACS Technical Report 2005-41.

R. Albert, B. DasGupta, R. Dondi and E. Sontag, "Inferring (biological) signal transduction networks via transitive reductions of directed graphs," submitted to journal.

V. Choi, P.K. Agarwal, H. Edelsbrunner, and J. Rudolph, "Local search heuristic for rigid protein docking," *Proceedings 4th Workshop on Algorithms in Bioinformatics (WABI), Lecture Notes in Computer Science 3240*, (2004) 218-229.

G. Cormode and S. Muthukrishnan, "Substring compression problems," *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, 2005.

B. Dasgupta, S. Ferrarini, U. Gopalakrishnan and N-R. Paryani, "Inapproximability results for the lateral gene transfer problem," DIMACS Technical Report 2005-14.

M. Gargano, "Consensus of phylogenetic trees using genetic algorithms," *Congressus Numerantium*, submitted.

P-H. Huang and V. Pavlovic, "An interpretable method for protein homology detection," *International Conference on Intelligent Systems for Molecular Biology (ISMB2006)*, submitted.

X. Li and J. Liang, "Computational design of combinatorial peptide library for modulating protein-protein interactions," *Pacific Symposium on Biocomputing* (2005), 28-39.

I. Lozina, "Artificial variables and artificial Boolean classifiers," in preparation.

A. Vashist, C. Kulikowski, I. Muchnik, "Screening for ortholog clusters using multipartite graph clustering by quasi-concave set function optimization," *Proceedings of The Tenth International Conference on Rough Sets, Fuzzy Sets,Data Mining, and Granular Computing* , RSFDGrC, (2005), 409-419.

A. Vashist, C. Kulikowski, I. Muchnik, "Ortholog groups as clusters on a multipartite graph," *Workshop on Algorithms in Bioinformatics* (WABI 05), (2005), 328-340.

A. Vashist, C. Kulikowski, I. Muchnik, "Automating protein function annotation through candidate ortholog clusters from incomplete genomes," CSB Workshops, (2005), 73-74.

A. Vashist, C. Kulikowski, I. Muchnik, "Protein function annotation based on ortholog clusters extracted from incomplete genomes using combinatorial optimization," *ACM Conf. on Research in Computational*

Molecular Biology (RECOMB), Lecture Notes in Computer Science Series, Springer Berlin / Heidelberg **3909** (2006), 99-113.

A. Vashist, C. Kulikowski, I. Muchnik, "Protein function annotation based on ortholog clusters extracted from incomplete genomes using combinatorial optimization," *The Journal of Computational Biology*, to be submitted.

A. Vashist, C. Kulikowski, I. Muchnik, "Ortholog clustering on a multipartite graph," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCB)*, to appear.

A. Vashist, C.A. Kulikowski, I.B. Muchnik, "Ortholog clustering on a multipartite graph," in *Proceedings of 5th Workshop on Algorithms in Bioinformatics 2005*, Lecture Notes in Computer Science Series, Springer Berlin / Heidelberg, to appear.

A. Vashist, , Z. Zhao, A. Elgammal, I. Muchnik, and C. Kulikowski, "Discriminative part selection by combinatorial and statistical methods for part-based object recognition," *Proceedings of Computer Vision & Pattern Recognition (CVPR 2006) Workshop - Beyond Patches*, 2006.

Y. Wang, P.K. Agarwal, H. Edelsbrunner, and J. Rudolph, "Coarse and reliable geometric algorithm for protein docking," *Pacific Symposium on Biocomputing*, (2005), 64-75.

Talks

G. Cormode and S. Muthukrishnan, "Substring compression problems," ACM-SIAM Symposium on Discrete Algorithms, 2005.

I. Lozina, "Logical analysis of data: from combinatorial optimization to biomedical, financial, and management applications," Optimization Days 2006, The Annual Meeting of the Canadian Operational Research Society, Montreal, May 8-10, 2006.

F.S. Roberts, "The RNA detective game: Finding RNA chains from fragments," DIMACS Biomath Connect Institute Meeting, July 2006.

F.S. Roberts, "Consensus list colorings and physical mapping of DNA," Combinatorial Challenges 2006 (A Meeting in Celebration of Pavol Hell's 60th Birthday), Victoria, BC, Canada, May 2006.

A. Vashist, C. Kulikowski, and I. Muchnik, "Protein function annotation based on ortholog clusters extracted from incomplete genomes using combinatorial optimization," *ACM Conf. on Research in Computational Molecular Biology (RECOMB)* Venice, Italy, April 2-5, 2006.

A. Vashist, C.A. Kulikowski, I.B. Muchnik, "Ortholog clustering on a multipartite graph," 5th Workshop on Algorithms in Bioinformatics 2005, Mallorca, Spain, October 4, 2005.

A. Vashist, Z. Zhao, A. Elgammal, I. Muchnik, and C. Kulikowski, "Discriminative part selection by combinatorial and statistical methods for part-based object recognition," *Computer Vision & Pattern Recognition (CVPR 2006) Workshop - Beyond Patches*, New York, New York, June 17, 2006.

Main website

http://dimacs.rutgers.edu/SpecialYears/2005_IPB/

Other Specific Products

Web pages

BioMaPS/DIMACS/MBBC/PMMB/SYCON Short Course: Molecular Mechanisms and Models of Bacterial Signal Transduction
http://dimacs.rutgers.edu/Workshops/Transduction

Workshop on Information Processing by Protein Structures in Molecular Recognition
http://dimacs.rutgers.edu/Workshops/InformationProcessing/

Workshop on Detecting and Processing Regularities in High Throughput Biological Data
http://dimacs.rutgers.edu/Workshops/Detecting/

Workshop on Machine Learning Approaches for Understanding Gene Regulation
http://dimacs.rutgers.edu/Workshops/MachineLearning/

Working Group on DNA Barcode of Life
http://dimacs.rutgers.edu/Workshops/Barcode/

Working Group on Evolution of Gene Regulatory Logic
http://dimacs.rutgers.edu/Workshops/Regulatory/

Workshop on Data Mining, Systems Analysis, and Optimization in Neuroscience
http://dimacs.rutgers.edu/Workshops/Neuroscience

Short Course: A Field Guide to GenBank and NCBI Molecular Biology Resources
http://dimacs.rutgers.edu/Workshops/GenBank

DARPA Workshop on State-Dependent Delays in Regulatory Networks
http://dimacs.rutgers.edu/Workshops/Delays/

Workshop on Computational/Experimental Approaches to Protein Defects in Human Disease
http://dimacs.rutgers.edu/Workshops/Neurodegenerative/

Workshop: Sequence, Structure and Systems Approaches to Predict Protein Function
http://dimacs.rutgers.edu/Workshops/ProteinFunction/

Workshop: Clustering Problems in Biological Networks
http://dimacs.rutgers.edu/Workshops/Clustering/

Short Course: Exploring 3D Molecular Structures Using NCBI Tools
http://dimacs.rutgers.edu/Workshops/Exploring

BioMaPS/DIMACS/MBBC/PMMB Short Course: Biological Development
http://dimacs.rutgers.edu/Workshops/CellCommunication/

Workshop: The DNA Barcode Data Analysis Initiative (DBDAI): Developing Tools for a New Generation of Biodiversity Data
http://dimacs.rutgers.edu/Workshops/DNABarcode/

Workshop: Machine Learning Techniques in Bioinformatics
http://dimacs.rutgers.edu/Workshops/MLTechniques/

Working Group on Computational Tumor Modeling
http://dimacs.rutgers.edu/Workshops/TumorModelingWG/

Workshop on Computational Tumor Modeling
http://dimacs.rutgers.edu/Workshops/TumorModeling/

Reports

<u>Contributions</u>

Contributions within Discipline

This special focus is of course by nature multi-disciplinary. A major contribution is the impact on the research programs and careers of the participants. Here is a selection of comments from the participants describing this.

“I participated in a workshop about a month ago on time delays in biology, and I will be participating at a workshop in August on tumor modeling, at which I have been asked to give a talk.

Our group has been working in both of these areas, and the delays workshop was extremely informative for us, as we have only recently been working on time delay problems in biology. Most useful for us was 1) to see what others are doing in terms of time delay models, 2) to see what others are doing in terms of the mathematical analysis of time delay systems, 3) to identify what the open problems are in both areas so as to focus our future work. The time delay workshop was extremely useful to us in all three areas. Moreover, as a result, we are now preparing a manuscript on time delayed negative feedback models--at the time of my attendance, I was quite uncertain of just how useful our results were, and the workshop convinced me that they were indeed worth publishing. In the manuscript, I acknowledge the workshop's importance to the work, and I will send you a reprint when the paper has been accepted. Finally, as an IBM employee, the quality and utility of this workshop made me proud of IBM's involvement with DIMACS.

As for the tumor workshop, I expect to get more or less the same thing out of my attendance. This is again a somewhat new area of research for our group--while we have published models for cancer-related signal transduction pathways, we are now extending those signal transduction pathway models in the context of tumors (ie. spatial models for tumor growth that incoporate relevant signal transduction pathways). If the delays workshop was any indication, then this, too, will be invaluable for shaping our future research efforts in tumor modeling.”
John Wagner
Functional Genomics & Systems Biology
IBM T.J. Watson Research Center

"I participated in a Short Course in Signal Transduction last year (June 2005, I think). As a mathematician who is attempting to use mathematical and computational methods in helping to describe various biological phenomena, I think I benefitted a lot from the Short Course. I have started to collaborate on a signal transduction problem (involving embryonic cell differentiation in Xenopus) with two biologists in the Houston area -both of whom are appreciative (and knowledgeable!) of mathematical methods used in describing biology. We foresee doing a lot of research work on this in the Summer -we actually have been submitting grant proposals to several agencies to fund this project. We expect undergraduate students to help us in this effort as well. I appreciate the efforts of DIMACS to promote interdisciplinary collaborations between biologists and mathematicians."
Edwin Tecarro
Department of Computer and Mathematical Sciences
University of Houston-Downtown

"I have attended several DIMACS workshops over the past five years and I have always found them very instructive. They always lead me to a new idea, which usually develops into a publication, or a meeting with someone who becomes a collaborator. The last one I attended was on cancer and at that workshop, I met a researcher from CINJ/UMDNJ. I am now working with him on "Breast Cancer Phenotypes".
Gyan Bhanot
IBM Research

"As a result of talks & discussion during the recent "Data Mining in Neuroscience" conference in Gainesville, I am licensing a program and trying out a method on some Merck data. That would be unlikely to have happened without the conference."
Matthew Wiener
Applied Computer Science & Mathematics
Merck Research Laboratories

DIMACS and the Consortium for the Barcode of Life:

DNA barcodes are short gene sequences taken from standardized positions in the genome. DNA barcodes are being developed as a global standard for species identification. DIMACS has joined with an international Consortium for the Barcode of Life (CBOL) (www.barcoding.si.edu), supported by the Alfred P. Sloan Foundation and hosted by the Smithsonian Institution's National Museum of Natural History. CBOL has launched a number of Working Groups that are addressing scientific and technical obstacles faced by DNA barcoding projects as well as other researchers in taxonomy and systematic and evolutionary biology. One of them, the Data Analysis Working Group (DAWG), includes computer scientists, statisticians, taxonomists and population geneticists who are developing a new generation of analytical protocols and techniques for DNA barcode data. This group, chaired by Professor Michel Veuille, a population geneticist and Chairman of the Department of Systematics and Evolution, National Museum of Natural History, Paris, and DIMACS is leading the US side of the activity. We organized a "brainstorming" meeting in September, identifying important computer science and statistics problems arising from DNA barcoding. We have now laid plans for an international competition to develop new analytical tools, building on the long-standing DIMACS program of algorithm implementation challenges.

Contributions to Other Disciplines

Since the "discipline" is inherently multidisciplinary, there is no separate entry in this section.

Contributions Beyond Science and Engineering

DNA and the Barcode of Life:

As noted in the section on Contributions within the Discipline, DIMACS has entered into partnership with the Consortium for the Barcode of Life As DNA barcodes become a more commonly-used system for species identification, users will need to have access to data analytical procedures that are standardized, calibrated, and accepted by the research community. This collaboration has the potential to provide the taxonomic community and other users of DNA barcode with the tools and techniques they will need to use barcode data and other molecular markers for species-level identification. These tools and techniques will be especially valuable for non-taxonomists such as border inspectors and regulatory agencies that need to identify specimens from incomplete specimens. Thus, access to barcode data and appropriate analytical tools will be valuable assets to varied consumers. Taxonomists will be important users of these data, along with ecologists, population biologists, and ecosystem scientists. Many government agencies will also welcome the ability to identify organisms using barcodes. For example, the US Department of Agriculture's Animal and Plant Health Inspection Service is responsible for border inspection and control of agricultural pests. The same is true for the US Environmental Protection Agency's use of barcodes as environmental quality indicators and the US Food and Drug Administration's efforts to detect mislabeled fish and meat, and prohibited ingredients in cattle feed using barcode data.

Contributions to Human Resources Development

Many graduate students, undergraduates, and several postdocs are participating in the program. Local graduate students and many non-local students are involved as visitors and workshop/working group attendees. More senior people are also heavily influenced by the project, being exposed to new directions of research. The impact on the careers of the students and faculty is illustrated by the following:

“I attended the Workshop on State-Dependent Delays in Gene Regulatory Networks (March 2-3, 2006). I found the meeting interesting, because I am a mathematician who is just starting to learn about molecular biology. At Drew I am beginning to use the telegraph equation to study currents in the dendrites of neurons, and have not yet begun to consider delays in the equations. … I feel that my time at the meeting was well spent. I am sure that future meetings of this type will be most useful to me.”
James McKenna
Drew University

“This was my first interaction with this program and it is too early to report on any scientific progress that directly evolved from the meeting. However, I found the program very informative. I now have a better sense of the utility of theoretical math for the biological problems in which I am interested and will almost certainly find a way to apply some of the mathematical principles discussed to our future studies. Thank you for the opportunity to participate in this valuable program.”
Peter Espenshade
Department of Cell Biology
Johns Hopkins University School of Medicine

In addition, the following graduate students have undertaken small research projects under support of the project.

Summer 2005:

Elad Hazan, Princeton, CS

“Develop efficient convex optimization algorithms, specifically for Linear Programming (LP) and Semi-Definite Programming (SDP)”

Irina Lozina, Rutgers, RUTCOR
“Synthetic variables in data analysis”

Akshaya Kumar Vashist, Rutgers, CS
“Detecting clusters of orthologous genes in a large set of genomes”

Rong Zhang, Rutgers, CS
“Classification based on characteristic samples”

Winter 05/06:

Suhrid Balakrishnan, Rutgers, CS
“A comptuational approach to discover Gene essentiality”

Akshaya Kumar Vashist, Rutgers, CS
“Improving ortholog detection on a multipartite graph by incorporating intra-genome gene interactions”

Liming Wang, Rutgers, Math
“Investigate a paradox in DNA damage pathway”