DIMACS Center
Rutgers University

Special Focus on Computational Molecular Biology

Final Report

September 2004

Ia. Participants from the program

Participants:

Special Focus Organizers:

Martin Farach-Colton, Rutgers U. Computer Science
Craig Nevill-Manning, Rutgers U. Computer Science
Wilma Olson, Rutgers U. Molecular Biophysics, Biochemistry, and Chemistry
Fred Roberts, Rutgers U. Mathematics, DIMACS

Additional Organizer:

Ronald Levy, Rutgers U. Molecular Biophysics, Biochemistry, and Chemistry

Steering Committee:

Bonnie Berger, MIT
Helen Berman, Rutgers University
Douglas Brutlag, Stanford University
Andrea Califano, First Genetic Trust
Nick Cozzarelli, UC Berkeley
Dannie Durand, Carnegie Mellon University
Jim Fickett, SmithKlineBeecham
Sorin Istrail, Celera Genomics
Thomas Lengauer, GMD
Ron Levy, Rutgers University
Michael Liebman, Roche Bioscience
Mona Singh, Princeton University
Martin Vingron, National Cancer Center of Germany
Michael Waterman, University of Southern California

Workshop Organizers:

Adam Arkin, Lawrence Berkeley Labs and UC Berkeley
Stephen Bryant, NIH
Andrea Califano, First Genetic Trust
Danny Chen, University of Notre Dame
Andy Clark, Cornell University
Bernard Coleman, Rutgers University
William H.E. Day
Thomas Deisboeck, Harvard Medical School
Paul Ehrlich, Rutgers University
Mark Gerstein, Yale University
Conrad Gilliam, Columbia University
Dimitrios Gunopulos, University of California – Riverside
Laurie Heyer, Davidson College
Richard L.X. Ho, R.W. Johnson Pharmaceutical Research Institute
Frank Hwang, National Chiao Tung University
Sorin Istrail, Celera Genomics
Mel Janowitz, Rutgers University
Nikolaous Koudas, AT&T Labs - Research
Casimir Kulikowski, Rutgers University
Francois-Joseph LaPointe, University of Montreal

Jean-Claude Latombe, Stanford University
Ron Levy, Rutgers University
Michael Liebman, Abramson Family Cancer Center, University of Pennsylvania, School of Medicine
Randy Linder, University of Texas
Tara Matise, Department of Genetics
Fred McMorris, Illinois Institute of Technology
Boris Mirkin, University of London
Gaetano Montelione, Rutgers University
Bernard Moret, University of New Mexico
Ilya Muchnik, Rutgers University
Joseph H. Nadeau, Case Western Reserve University
Wilma Olson, Rutgers University
Teresa Przytycka, Johns Hopkins University
Fred Roberts, Rutgers University
Steven L. Salzberg, The Institute for Genomic Research
David Sankoff, University of Montreal
Lee Segel, Weizmann Institute
Anirvan Sengupta, Rutgers University
Mona Singh, Princeton University
William Sofer, Rutgers University
Eduardo Sontag, Rutgers University
Gustavo Stolovitzky, IBM
David Torney, Los Alamos National Labs
Michael W Trosset, College of William & Mary
Vijay Vazirani, Georgia Tech
Michael Waterman, University of Southern California and Celera
Raimond Winslow, Johns Hopkins
Denise Wolf, Lawrence Berkeley Labs
Shibu Yooseph, Celera Genomics
Victor Zhurkin, NIH

Visitors:

Lloyd Demetrius, Harvard University, 12/10/2001 - 12/24/2001
Alair Pereira do Lago, University of São Paulo, 9/15/2000 - 1/31/2002
Raffaele Esposito, Universita di L'Aquila, 2/25/2001 - 3/25/2001
Gregory Kucherov, INRIA, 8/4/2003 - 8/29/2003
Rossana Marra, Universita di Roma 2, 2/25/2001 - 3/25/2001
Boris Mirkin, Birkbeck College, 10/23/2000 - 11/5/2000; 12/14/2000 - 12/20/2000; 9/27/2001 - 10/7/2001; 9/17/2004 - 9/23/2004
Vadim Mottl', Tula State University, 12/5/2000 - 12/19/2000
Leonid Shvartser, Ness A.T. Ltd., 12/13/2000 - 12/27/2000
Catherine Womack, Bridgewater State College, 8/1/2003 - 8/31/2003

Graduate Students:

Gabrielle Alexe, Rutgers University, RUTCOR, Summer 2002
Sorin Alexe, Rutgers University, RUTCOR, Summer 2002
Ziv Bar-Joseph, MIT, Celera Genomics, 2003
Jie Chen, Princeton University, Computer Science, Winter 2000-2001, and Winter 2003-2004
Barry Cohen, SUNY - Stony Brook, Celera Genomics, 2001
Khaled Elbassioni, Rutgers University, Computer Science, Winter 2001-2002

German Encisco, Rutgers University, Mathematics, Summer 2003
Rohan Fernandes, Rutgers University, Computer Science, Summer 2003
Jessica Fong, Princeton University, Celera Genomics 2004
Jaewook Joo, Rutgers University, Physics, Fall 2002 and Spring 2003
Carl Kinsford, Princeton University, Computer Science, Summer 2004
Steven Kleinstein, Princeton University, Computer Science, Summer 2001 and Summer 2002
Chan-Su Lee, Rutgers University, CAIP Center, Summer 2004
Zhong Jun Luo, Rutgers University, Computer Science, Winter 2001-2002
Jin Ma, Rutgers University, Computer Science, Winter 2002-2003
Matthew Menke, MIT, Celera Genomics 2004
Itsik Pe'er, Tel Aviv University, Celera Genomics 2001
Mihaela Pertea, Johns Hopkins University, Celera Genomics 2001
Scott Rifkin, Yale University, Celera Genomics 2002
Erich Schmidt, Princeton University, Computer Science, Summer 2001
Sandor Szesmak, Rutgers University, RUTCOR, Summer 2000
Akshay Vashisht, Rutgers University, Computer Science, Summer 2002

Ib. Participating Organizations

Telcordia Technologies: Facilities; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and workshops.

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

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

Lucent Technologies, Bell Labs: Facilities; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and workshops.

Princeton University: Facilities; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and workshops.

Avaya Labs: Facilities; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and workshops.

HP Labs: Facilities; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and workshops.

IBM Research: Facilities; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and workshops.

Microsoft Research: Facilities; Personnel Exchanges
Partner organization of DIMACS. Individuals from the organization participated in the program planning and workshops.

The Alfred P. Sloan Foundation

The New Jersey Commission on Science and Technology

SmithKline Beecham

Celera Genomics

1c. Other Collaborators

The project involved scientists from numerous institutions in numerous countries. There were hundreds of attendees at our workshops, coming from a variety of types of institutions and disciplines. The resulting collaborations also involved individuals from many institutions in many countries.

II. Project Activities

The Special Focus started in September 2000 and was motivated by major developments at the interface between biology and information science.

First, because of new technologies, we are receiving massive amounts of data and, moreover, the data are arriving in diverse forms and are often unstructured and non-homogeneous. This requires the intervention of experts at handling databases and processing massive amounts of information. Moreover, data is available at much higher levels of organization than heretofore (for example at the whole genome level rather than at the gene level), which makes possible dramatic new biological understanding if we could find ways to handle the complex algorithmic questions that arise.

Second, new biotechnologies such as gene expression arrays are powerful new experimental tools, but these tools and the development of new ones need to be integrated with fundamental research concerning algorithms. Thus, close collaborations between biological scientists and computer scientists needed to be developed and/or strengthened.

Third, new mathematical and computational tools have made it possible to understand biological processes at a much more complex level than before, for example at the level of cellular communities, the immune system, protein structure and design, and processes of learning and memory. This comes at a time when biologists were finding tools to explore how parts of a biological system (e.g., genes or molecules) interact, something that has long been understood to be just as important as understanding the parts themselves but which has been largely inaccessible to experimental or analytical techniques. These new tools make it possible to formulate powerful new mathematical models of biological systems, grounded in the communication and information transmission fundamental to those systems. This calls out for the expertise of mathematical modelers and information scientists who are best able to take advantage of the powerful new mathematical methods while at the same time keeping their models realistic through interactions with biological scientists.

The DIMACS “Special Year” on Mathematical Support for Molecular Biology (1994-2000) commenced in 1994-95 with a very intense level of activity and had continued at a significantly lower level since then. Through that special year, we focused the interests of a large number of computer scientists and mathematicians on the field of molecular biology, created real partnerships between mathematical and biological scientists, and trained outstanding young people to work in computational biology. We felt that the time was right for another highly visible special year thrust, like the one in 1994-95. At the time of our first special year, there was a need to address, review, and coordinate a variety of topics and methods, as well as to introduce many people to the field. Computational biology prior to the first special year was largely focused on a few problems such as alignments, physical mapping, and reconstructing phylogenies. Many of the topics emphasized in the first special year became "standard" topics of research in computational biology and the field had become much broader in its focus. Other topics began to be investigated in a much more far-reaching way using methods of computer science, for example problems involving gene finding and motif recognition, protein and RNA folding, protein structure prediction, and linkage analysis. New biotechnologies such as SBH (sequencing by hybridization), optical mapping, and EST (expressed sequence tags) had been brought to the attention of a much larger number of computer scientists so that they could formulate and research the right questions concerning new technologies. But a whole host of new problems that lay at the interface between the computational/mathematical sciences and the biological sciences arose, stimulated, as we have noted, by the availability of massive amounts of new data, the integration of new experimental methods with algorithmic methods, and the development of powerful new mathematical tools for modeling ever-more-complex biological systems.

The title of the special focus, "Computational Molecular Biology," is certainly too broad to define its focus. We concentrated on those areas where discrete mathematics (DM) and theoretical computer science (TCS) seemed likely to have a major impact. A fundamental premise of the special focus was that some of the most central problems in modern molecular biology were essentially problems involving the combinatorial and algorithmic questions of DM and TCS and it was a basic objective of the focus to create partnerships between biological and mathematical/computer scientists so that some of these central biological questions could be precisely formulated and analyzed. By bringing together some of the world's leading mathematicians/computer scientists with leading biological scientists, major progress was made. Moreover, we established and nurtured lines of communication and collaboration so that the new methods of information science that arose from the special focus could be quickly adapted to biological applications, and could be readily communicated to the biological community.

The major organized events were a series of workshops and "working group" meetings.

Workshop and Working Group Meeting on Bioconsensus

Date: October 25 - 26, 2000
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Mel Janowitz, Rutgers University; Francois-Joseph LaPointe, University of Montreal; Fred McMorris, Illinois Institute of Technology; Boris Mirkin, University of London; Fred Roberts, Rutgers University
Attendance: 35

Consensus methods developed in the context of voting, decision making, and other areas of the social and behavioral sciences have begun to have a variety of applications in the biological sciences, originally in taxonomy and evolutionary biology, and more recently in molecular biology. Typically, several alternatives (such as possible taxonomies, alternative phylogenetic trees, alternative molecular sequences, or alternative alignments) are produced using different methods or under different models and then one needs to find a consensus solution. There are, already, several hundred papers in this developing field of "BioConsensus." In this workshop, we explored ways to make use of the consensus methods of social choice theory in solving problems of biology, with emphasis on molecular biology.

Here are several of the major themes of the workshop. How have consensus methods of social choice theory already found use in biology? In turn, how have some of the specific problems of the biological sciences given rise to new concepts of consensus? Algorithms for some of the well-known consensus methods of social choice theory have the potential for application to biology, with appropriate modification, but many of these consensus problems are NP-complete in their most general setting and call for approximate algorithms or heuristic methods. What would be involved in applying traditional consensus methods to molecular biology problems? Consensus methods in molecular biology tend to be chosen because they seem mathematically interesting or useful rather than on the basis of any reasonable biological model. What is a reasonable basis for choosing a consensus method? Are there reasonable axioms having biological meaning that characterize different consensus methods?

This workshop brought together mathematicians, computer scientists, and biological scientists to discuss these and other issues. The workshop was viewed as forming a basis for a continuing dialog that led to future collaborations. In particular, there was a follow-up Tutorial and Workshop on Bioconsensus II, October 2-5, 2001.

Directly following the workshop, there was a “DIMACS distinguished lecture” by Gene Myers, Vice President of Celera Genomics, on the topic "Whole Genome Assemblies of the Drosophila and Human Genomes".

Distinguished Lecture: Gene Myers, Vice-President, Informatics Research, Celera Genomics
Whole Genome Assemblies of the Drosophila and Human Genomes
Date: October 26, 2000
Location: CoRE Building, First Floor Auditorium, Rutgers University
Attendance: 35

Dr. Myers reported on the design of a whole genome shotgun assembler and its application to the sequencing of the Drosophila and Human genomes. Celera's whole genome strategy consists of randomly sampling pairs of sequence reads of length 500-600 that are at approximately known distances from each other - short pairs at a distance of 2Kbp, long pairs at 10Kbp, and ultra-long pairs at 50-150Kbp. Reads are collected in a 1-to-1 ratio of short to long pairs, and enough ultra-long pairs to give 20-30X clone coverage is desirable. The experimental accuracy of the read sequences is roughly 99.5% with all but 1 in 10,000 being better than 98% accurate. Given such a data set, the computational problem is to infer the sequence of the euchromatic portion of the genome.

For Drosophila, Celera collected 1.6 million pairs whereby the sum of the lengths of the reads is roughly 13 times the length of the genome (~120Mbp), a so called 13X shotgun data set. For the human genome 12.5 million pairs for a 4.5X data set was generated at Celera, and then an additional 2X of faux reads was added by shredding the rough draft data obtainable at Genbank, for an aggregate 6.5X data set consisting of 37 million reads totaling 20Gbp of sequence.

By layering the ideas of uncontested interval graph collapsing, confirmed read pairs, and mutually confirming paths, one obtains an assembly algorithm that makes remarkably few errors. The assembler correctly identifies all unique stretches of a genome, correctly building contigs for each and ordering them into scaffolds, spanning each of the chromosomes. Thus all useful proteomic information is firmly assembled. For Drosophila, with a 13X data set, the results of assembly, *without* any of the finishing effort that ensues for all projects, meets the community standards set by Chromosome 22 and C. Elegans, for completion and accuracy of *finished* sequence. This assembler also completed a preliminary whole genome assembly of the 6.5X human data set in 3 weeks using 160 Alpha porcessors and a 64Gbp memory.

In order to cross-validate the results of thei human whole genome assembly, Celera also built a regional assembler that combines Celera's data with the BAC-localized contigs being produced by the Human Genome Project. This assembler orders the contigs, fills roughly 2/3rds of the resulting gaps, and then, with the help of user curation, tiles the assembled BACs into meta-level assemblies that cover megabase-sized regions of the genome.

Working Group Meeting: The Informatics of Protein Classification

Date: December 15, 2000
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Casimir Kulikowski, Gaetano Montelione and Ilya Muchnik, Rutgers University
Co-sponsored by The New Jersey Commission on Science and Technology Initiative on Structural Genomics and Bioinformatics and by Rutgers Bioinformatics Initiative.
Attendance: 35

This workshop investigated models of protein classification and protein interaction. It seems both impossible and unnecessary to describe in detail every unique protein. To understand how a protein's machinery works in a cell, one wants to develop an efficient classification of proteins. It is well accepted that such a classification should be based on particular pieces of proteins that are called domains. Unfortunately, there don't exist precise formal definitions of domains that are accepted by most experts, either for amino acid sequence or 3D-structure presentations of proteins. There are a few databases that present classifications and one can consider them to infer experimentally presented definitions. For instance, the SCOP database or FSSP are such databases for protein domains formed by structural data and PRODOM and PFAM are protein domain databases organized around sequence information. Acceptable formal definitions are important for genetic engineering, drug design, and other fields of modern biotechnology, because only through them can one build software to manipulate proteins in silico, and thus we explored them in this workshop.

Comparison analysis is a fundamental part of many protein studies. We paid particular attention to similarity measures for proteins in different representations. There are a lot of tools for such analysis, but we don't yet know the limits of their usefulness. Moreover, we know very little about how existing tools can work together in some integrated way. We discussed these questions in the workshop.

Mini-Workshop: System Based Modeling in Bioinformatics

Dates: February 19 - 20, 2001
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Michael Liebman, Abramson Family Cancer Center, University of Pennsylvania, School of Medicine; Richard L.X. Ho, R.W. Johnson Pharmaceutical Research Institute
Attendance: 55

As biologists model complex systems whose properties are not fully explained by the properties of their component parts, they have long understood that it is important to investigate the interactions of those component parts, interactions such as those in which different cells work together in such tasks as determining when a cell divides and how gene expression is regulated. Bioinformatics, with an expansion to a systems-based perspective taking advantage of the expertise of mathematicians, computer scientists, engineers, and physicists, is well positioned to play a major role in achieving this.

Bioinformatics has evolved to focus on the molecular basis of genomic data, attempting to identify, qualify and quantify genes and gene products. The ultimate goal for the application of bioinformatics in practice, for example in the pharmaceutical and medical areas, is in the development of knowledge to impact the practice of medicine (i.e., diagnosis and treatment of predisposition and disease). Biomedical Informatics is relatively early in its evolution in that it examines the bioinformatic data from this systems-based perspective and attempts to integrate observations and knowledge about clinical disease to analyze the underlying biological processes. Success in these separate developments will come from their convergent evolution. To enable the interface between computation and experiment, stochastic and deterministic modeling, including graph theoretical methods, are being applied to the representation and evaluation of biological pathways and processes in normal and disease states. These computational approaches attempt to deal with incomplete information, unresolved molecular interactions and multiple modeling hierarchies. It is hoped that progress on them will result in their application in the analysis and interpretation of clinical disease, e.g., cancer, coagulation disorders, diabetes, in terms of gene identification for use in diagnostic and therapeutic target design. This workshop investigated these computational approaches and explored the system-based approach to bioinformatics.

Workshop: Whole Genome Comparison

Dates: February 28 - March 2, 2001
Location: DIMACS Center, CoRE Building, Rutgers University
Organizer: Dannie Durand, Carnegie Mellon University
Program Committee: Joseph H. Nadeau, Case Western Reserve University; Steven L. Salzberg, The Institute for Genomic Research; David Sankoff, University of Montreal
Attendance: 92

This workshop brought together biological and computational scientists to discuss whole genome comparison at three levels:

- Whole genome sequence comparison, including the implications for identification of protein coding and regulatory regions and the origin and role of non-coding DNA.

- Whole genome map comparison, including the importance of spatial genomic organization and the role of duplications and rearrangements in genome evolution and structure.

- Functional genomic comparison: comparison of comprehensive functional data sets such as the set of orthologs, protein folds, expression patterns, pathways or protein-protein interactions found in a given organism.

Work that addressed whole genome comparison on two or more of these levels was of particular interest. The workshop focused on research comparing the genomes of different organisms, as well as the comparison of a single genome with itself.

Workshop: Protein Structure and Structural Genomics: Prediction, Determination, Technology and Algorithms

Dates: March 8 - 9, 2001
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Mona Singh, Princeton University; Sorin Istrail, Celera Genomics; Ron Levy, Rutgers University
Attendance: 123

The growth of both protein sequence and protein structure databases has had considerable influence on the development of computational and experimental methods for understanding protein structure. Within this context, this workshop focused on:

- How the integration of sequence and structure databases has changed computational methods for predicting protein structure.

- Structural genomics and high-throughput structure determination.

This workshop brought together experimentalists and theoreticians to assess the strengths and weaknesses of current methods for understanding protein structure, and brainstorm about how the growing biological databases would influence development of future methods.

Workshop: DNA Sequence and Topology

Dates: April 19 - 20, 2001
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Wilma Olson and Bernard Coleman, Rutgers University; Victor Zhurkin, NIH
Attendance: 95

The packaging of DNA within the close confines of the cell imposes a higher order structure on the long threadlike molecule. The chain must fold and adopt arrangements that allow for correct recognition and processing of the genetic message. This organizational structure is only beginning to be understood. We know, for example, that the linear sequence of genetic information also includes a base sequence-dependent spatial and energetic code that governs the global folding of the double helix and its susceptibility to interactions with other molecules and that basic biological processes, such as the transcription of the genetic code, the replication of DNA, and the repair of damaged DNA, are based on mechanisms sensitive to these physical properties.

Access to the genetic code requires severe changes in local DNA structure. Proteins usually gain access to the atoms holding the code by a partial unwinding and bending of the double helix, processes that can give rise in the DNA to “supercoiling.” Deciphering the sequence-dependent structural and deformational codes in DNA as well as the interplay between the local and long-range structure associated with its biological activity require a variety of mathematical and computational approaches: tools to extract knowledge-based “energies” from structural and thermodynamic nucleic acid databases, such as those organized at Rutgers; techniques to simulate the dynamical structures and equilibrium properties of ensembles of DNA molecules; explicit and exact solutions of the non-linear equations governing the supercoiled shapes of a deformable DNA elastic rod; analysis of the topological forms of DNA produced by enzymatic action; explicit expressions to incorporate base sequence-dependent structural information in genomic analyses.

This workshop brought together diverse perspectives on DNA topology to build bridges not only between mathematics and molecular biology but also between the mathematical and physical scientists currently focused on either the local or global view of DNA structure. This workshop followed two previous DIMACS workshops on DNA topology. Those dealt exclusively with the global folding of idealized, sequence-independent DNA rods and made no connection to the structural and energetic information embedded in the genetic code. The databases of DNA 3D structures have grown to the point where only now is it possible to extract the sequence-dependent information required to connect macromolecular structure and properties to the sequence of DNA base pairs, making this workshop extremely timely.

Workshop: Integration of Diverse Biological Data

Dates: June 21 - 22, 2001
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Andrea Califano, First Genetic Trust; Conrad Gilliam, Columbia University; Fred S. Roberts, Rutgers University
Attendance: 112

This workshop focused on combinatorial algorithms and probabilistic models for the analysis and cross-annotation of biological data in diverse databases. The rapid accumulation of biological data has led to one of history's largest and growing collections of unstructured, loosely related databases. Most of the useful relationships in this haphazard collection of data are the direct result of a laborious task of manual annotation and experimentation.

At the current pace at which new biological information is becoming available, the ability to automatically cluster, classify, and annotate it across the traditional boundaries of individual databases is becoming an increasingly critical need. For instance, DNA sequences for promoter and enhancer regions, structural motifs in transcription factors, metabolic pathway databases, and gene expression analysis are all tightly bound and interconnected. However, they tend to be studied in isolation. When this happens, clustering techniques are often less effective because in the absence of additional constraints they have to deal directly with the high dimensionality of the solution space. Functional clustering of protein sequences, for instance, can help reduce the complexity of structural clustering and vice versa. Analogously, functional clustering in the gene expression domain has been shown to significantly reduce the complexity of promoter region analysis.

We considered high-dimensional combinatorial algorithms and probabilistic models central to discrete mathematics for the analysis, clustering, and classification of complex data patterns that can be used to integrate diverse information from many biological sources. Motivating approaches included the work of Roth, Hughes, Estep and Church tying together mRNA monitoring and transcription factors; of Bystroff and Baker using 1D database information to improve 3D fold recognition; of Eckman, et al. on large-scale diverse data; and of Karp and Paley integrating genomic data and metabolic pathways data.

Summer School Tutorial on New Frontiers in Data Mining

Dates: August 13 - 17, 2001
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Dimitrios Gunopulos, University of California - Riverside and Nikolaous Koudas, AT&T Labs - Research
Attendance: 83

This "summer school" tutorial program was aimed at providing background, vocabulary, and theoretical methodology to non-specialists in data mining and to others who wished to explore this field and at bringing together students, postdocs, and researchers working on algorithms for data mining with those working in various application areas. More specifically, we aimed to introduce the attendees to the fundamental theoretical/algorithmic issues that arise in data mining and its applications.

Data mining is an exciting new field of computer science research, encompassing several diverse techniques for analyzing large datasets. The goal of data mining is to obtain new, interesting and actionable pieces of information. Vast amounts of data are accumulated in diverse application domains, including bioinformatics, epidemiology, business, physical sciences, web applications, and networking. Data mining research is stimulated by hard real life problems in analyzing data in all those areas. Data mining is fundamentally an interdisciplinary field, borrowing and combining techniques from theory, statistics, databases and machine learning, and ultimately producing new approaches.

A goal of this tutorial was to bring together students, postdocs, and researchers from the fields of data mining, bioinformatics, networking, and the web, and to facilitate the collaboration between fields, as well as to introduce the field of data mining to those who are not yet working in it or are not yet working in it from an algorithmic point of view.

In the tutorial we concentrated on new research directions that are currently emerging in the field: data mining applications in bioinformatics, networking, and the web. We explored new problems that have come up in these areas, identified common threads among the various applications, and considered new paradigms, methods and techniques that are being developed to address these problems. In the tutorial we emphasized the algorithmic aspects of analyzing large datasets. There are different general ways to approach this problem, such as approximate algorithms and data summarization techniques. We looked at new techniques on stream processing and online algorithms, and their applications to specific problems.

Biological research is undergoing a major revolution as new technologies, such as high-throughput DNA sequencing and DNA microarrays, are creating large amounts of data. New techniques in analyzing such data are important in the understanding of biological processes. Many bioinformatics problems can be formulated as generalized search problems in a large space. We looked at general lattice search techniques with different constraints, as well as new string algorithms. We also looked at applications of classification techniques in the area.

Networking and telecommunications applications produce large amounts of data that can be mined for various properties of interest. Time series data prevail in such domains and algorithms for time series matching, sequential pattern identification are of great interest. We concentrated on incremental and one pass algorithms for networking problems and explored the connection between these problems and similar incremental and one pass problems arising in the biological sciences.

The web has emerged as a vast datastore, containing diverse pieces of information. We examined recent approaches to mine information on the World Wide Web, including efficient web searching and web site personalization efforts. We also looked at data and resource management issues in the web environment, with emphasis on bioinformatics and telecommunications applications.

Tutorial on Bioconsensus II

Dates: October 2, 2001
Location: DIMACS Center, CoRE Building, Rutgers University
Tutorial Organizers: Fred McMorris, Illinois Institute of Technology and William H.E. Day
Attendance: 32

Workshop on Bioconsensus II

Dates: October 3 - 5, 2001
Location: DIMACS Center, CoRE Building, Rutgers University
Workshop Organizers: Mel Janowitz, Rutgers University; Francois Lapointe, Universite de Montreal; Fred McMorris, Illinois Institute of Technology; Boris Mirkin, University of London; Fred Roberts, Rutgers University
Attendance: 32

This tutorial and the following workshop was a follow-up to a working group meeting on Bioconsensus that was held October 25-26, 2000 at DIMACS. There was a one-day tutorial on consensus theory followed by a series of public lectures in workshop format.

Consensus methods developed in the context of voting, decision making, and other areas of the social and behavioral sciences have begun to have a variety of applications in the biological sciences, originally in taxonomy and evolutionary biology, and more recently in molecular biology. Typically, several alternatives (such as possible taxonomies, alternative phylogenetic trees, alternative molecular sequences, or alternative alignments) are produced using different methods or under different models and then one needs to find a consensus solution. There are, already, several hundred papers in this developing field of "Bioconsensus." In this workshop, we explored ways to make use of the consensus methods of social choice theory in solving problems of biology, with emphasis on molecular biology.

Here are several of the major themes of the workshop (similar to those of its predecessor workshop). How have consensus methods of social choice theory already been used in biology? In turn, how have some of the specific problems of the biological sciences given rise to new concepts of consensus? Algorithms for some of the well-known consensus methods of social choice theory have the potential for application to biology, with appropriate modification, but many of these consensus problems are NP-complete in their most general setting and call for approximate algorithms or heuristic methods. What would be involved in applying traditional consensus methods to molecular biology problems? To what extent are consensus methods in molecular biology chosen because they seem mathematically interesting rather than on the basis of some reasonable biological model. What then is a reasonable basis for choosing a consensus method? Are there reasonable axioms having biological meaning that characterize different consensus methods?

The workshop brought together mathematicians, computer scientists, and biological scientists to discuss these and other issues. The meeting formed a basis for a continuing dialog that will lead to future collaborations. Though the basic format was that of a workshop, ample time was left for questions, discussions, and informal interactions among the participants.

The tutorial included a discussion of how an axiomatic approach can help either characterize or establish the nonexistence of families of consensus methods. Included were median procedures, majority and plurality rules, and rules based on concepts of center and mean. The objects on which the consensus rules operate included weak orderings of preferences, hierarchies, undirected phylogenetic trees, and molecular sequences.

Workshop: Analysis of Gene Expression Data

Dates: October 24 - 26, 2001
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Laurie Heyer, Davidson College; Gustavo Stolovitzky, IBM; Shibu Yooseph, Celera Genomics
Attendance: 167

The gene expression array is a significant new technology aimed at providing a top down picture of the intimate genetic processes of an organism.

Whether an array is implemented using photolithography and silicon-based fabrication, capillary printing heads on a glass slide, or ink-jet technology, it allows quantification of transcription levels of large numbers of genes simultaneously.

Ever since the first microarray-based biological results were published, it has seemed unthinkable to tackle the complexities of the workings of the cell without these devices. However there remain unsolved image processing as well as computational and mathematical difficulties associated with the extraction and validation of data from gene expression microarray assays.

In the image processing domain, active research areas include noise estimation, background subtraction, and quality assessment in the feature recognition process. Downstream of the image analysis are the problems of reconstruction of biochemical pathways and genetic networks from transcriptional data. Subproblems include clustering transcription profiles of multiple genes over time and/or over populations, graph problems such as constructing metabolic pathways consistent with gene transcription clusters, string problems such as finding promoter sequences in genomic sequences that explain co-regulation, and probabilistic problems such as assessing the statistical significance of transcriptional patterns.

This workshop focused on these issues.

Workshop: Complexity in Biosystems: Innovative Approaches at the Interface of Experimental and Computational Modeling

Dates: April 8 - 10, 2002
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Thomas Deisboeck, Harvard Medical School; Lee Segel, Weizmann Institute; Eduardo Sontag, Rutgers University; Raimond Winslow, Johns Hopkins
Attendance: 125

This workshop brought together scientists who were already working in the novel field of complex biosystems modeling and simulation from different points of view and "newcomers" who were interested in getting involved in this exciting research area. The focus was on the challenging interface between experimental modeling (e.g., assay design and engineering) and computational simulation. The goal was to introduce and discuss innovative concepts, experimental and computational biosystems models and mathematical algorithms as well as to establish new collaborations beyond institutional or departmental boundaries.

Recently, complex biosystems science has attracted substantial interest. This is especially so because conventional biological science has produced a vast amount of data over the last few decades so that questions arise as to how to find patterns and how to relate multi-quality data sets in the quest for underlying mechanisms. The most visible example is the Human Genome Project and its spin-off sciences: genomics and proteomics. However, the dynamics of such complex biological systems cannot be simply explained by combining the separately measured behavioral features of its monomers. As such, a living cell is more than the sum of its organelles; other examples for biosystems include the neuronal networks in the brain, the immune system, disease processes as well as population dynamics and entire ecosystems. It is clear that conventional, reductionism-determined research approaches must fail in understanding the mechanisms behind the complex pattern generation, the self-organization and nonlinear interaction of these multi-scaled systems. Thus to develop novel research approaches with complex biosystems science will be one of the grand challenges of the next decade. Tremendously increased computational power will help in the efforts ahead to analyze the immense amount of data in the biological and biomedical sciences in order to guide promising new experimental work. Such "experimental biosystems modeling" means e.g. the design, the development and use of novel assays, i.e. in vitro models, based on and driven by the mathematical and computational modeling. We need to come up with such new experimental methods to test and refine the predictions made with novel theoretical models - especially if based on an underlying ‘complex systems concept’. Most conventional experimental models, however, have been developed in the reductionism-era, i.e., they focus on one endpoint and emphasize one feature - with little dynamical information and lacking the possibility of studying more than one to two features of the system (reproducibly) at the same time. The concept was for a long time that one simply has to dissect the biology, investigate it separately, and finally put it all back together. Most scientists would now admit that although this approach has led to very significant discoveries in the past, it will not be able to explain the complex behavior of most biological systems.

Therefore, experimental approaches also have to change - and we hope that the ongoing theoretical, computational and mathematical modeling and simulation efforts will support and push this development. We further hope that theoretical modeling will give us ‘hints’ as to where to investigate in even greater detail in experiments (thus guiding future conventional approaches) and how to design and engineer these experiments settings properly to take the complexity into account - not to take it out of the calculation. This workshop bridged the gap between experimental and computational modeling experts. Innovative complex biosystems research requires even more than only biology-inspired computational science. In fact, it can only be successful if multiple seemingly disparate disciplines combine their techniques and expertise, including biology, physics, engineering, mathematics and medicine, even economics and sociology. In summary, the need for truly interdisciplinary teams is apparent. This workshop linked more closely the groups already working in this area and got scientists involved who are just starting in this emerging scientific field. Topics included experimental modeling concepts on the intracellular, the supracellular and the tissue level and the required combination with computational approaches such as genetic net modeling, cell signaling modeling as well as continuum and discrete modeling for multi-element systems.

From a biological point of view, the meeting was organized by (modeling) scales - ranging from the molecular to the cell, multicellular as well as the organ and disease process levels, with cognizance of the fact that perhaps the most interesting problems span more than one scale. Contributors structured their presentations in terms of the categories:

(1) biomedical background and experimental platform(s)
(2) computational/mathematical modeling concept and modeling platform
(3) discussion of the modeling results (with emphasis on future work on both sides, experiment & simulations and its potential impact on biomedicine).

Given that the exploration of the input-parameter sets has been one of the main bottlenecks for modeling efforts, one focus of the discussion was to define the requirements for novel experimental biomedical model systems. As such, the workshop was structured in "case-studies", which represent the scale-related modeling efforts. Nonetheless, after each session we discussed the multiscale "environment", i.e., interface and challenges involved in passing to lower/higher scales so that a linkage with the other sessions was achieved. We also had a panel discussion with audience participation on non-reductive experimentation considering new modeling paths and discussing collaborative projects.

DIMACS-CTS (National Chiao Tung University) Conference on the Interconnections Among Codes, Designs, Graphs and Molecular Biology

Dates: May 24 - 26, 2002
Location: National Center of Theoretical Science, National Chiao Tung University, Hsinchu, Taiwan
Organizers: Frank Hwang, National Chiao Tung University; Fred Roberts, DIMACS, Rutgers University; David Torney, Los Alamos National Labs
Attendance: 96

Discrete structures have been used to formulate and solve some of the most fundamental concepts of molecular biology. In this workshop, we explored:

- The connections between molecular biology and such critical and interrelated ideas of discrete mathematics as combinatorial designs, codes, graphs, and categorical sequences;

- Some of the fascinating new insights into these discrete concepts that have arisen from work in molecular biology and;

- Recent applications of concepts from discrete mathematics (broadly defined) in molecular biology.

We explored combinatorial design issues arising in the identification of clones containing a specific DNA sequence, coding theory problems with codewords implemented as DNA sequences and graph theoretical concepts that arose originally from DNA reconstruction when the ends of some of the clones are radioactively tagged.

The following are some of the interconnected topics explored in this workshop:

We explored the use of group testing methods in identifying positive clones, which is a crucial step in physical-map-based sequencing. The method of group testing, using combinatorial designs, tests a group of clones at a time, hence reducing the number of samples to be collected and assayed. The focus was on nonadaptive algorithms for fast implementation. More generally, we explored new types of design desiderata motivated by screening assays undertaken in various applications such as pooling to find "positive" subsets of molecules, in the study of inhibitory molecules, and in data mining problems in functional genomics.

There is a long history of interrelations among combinatorial design and coding theory and we explored the use of DNA sequences as codewords in the design of new types of codes. Similar considerations of finite sequences from a finite alphabet (sometimes called categorical sequences) arose in the analysis of genome-scan genotype data (such as marker data from the GAW 12 dataset) and constitute a generalized type of group testing. Various metrics underlie the development of codes, and we explored related work on categorical sequences involving the development of new metrics for sequence similarity.

A very useful tool in the physical mapping of DNA is the optical mapping technique that starts with many single, partially digested copies of a DNA molecule. Realistic algorithms, dealing with false positive and false negative and otherwise inaccurate data, use algorithms that combine continuous optimization with discrete methods and have recently found connections to the types of considerations arising in DNA coding.

In early approaches to realistic physical mapping of DNA, when typically only partial overlap data among clones exists, tagging the ends of some clones radioactively formed the basis for useful mapping procedures. The task of DNA reconstruction could then be looked at as the problem of characterizing the graphs called tagged probe interval graphs and constructing tagged representations for them when they existed. Since this earlier work, special classes of tagged probe interval graphs and other types of intersection graphs have arisen in fascinating ways from overlap of categorical sequences of letters from a finite alphabet and other bioinformatics applications. We explored a variety of graph-theoretical and other algorithms for such problems as sequence-fragment assembly.

Workshop: Computational Methods for SNPs and Haplotype Inference
Dates: November 21 - 22, 2002
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Sorin Istrail, Celera Genomics; Andy Clark, Cornell University; Michael Waterman, University of Southern California and Celera
Attendance: 114

The ability to score large numbers of DNA variants (SNPs) in large samples of humans is rapidly accelerating, as is the demand to apply these data to tests of association with diseased states. The problem suffers from excessive dimensionality, so any means of reducing the number of dimensions to the space of genotype classes in a biologically meaningful way would likely be of benefit. Linked SNPs are often statistically associated with one another (in "linkage disequilibrium"), and the number of distinct configurations of multiple tightly linked SNPs in a sample is often far lower than one would expect from independent sampling. These joint configurations, or haplotypes, might be a more biologically meaningful unit, since they represent sets of SNPs that co-occur in a population. There has been much excitement over the idea that such haplotypes occur as blocks across the genome, as these blocks suggest that fewer distinct SNPs need to be scored to capture the information about genotype identity. This workshop focused on the formal analysis of this dimension reduction problem, the formal treatment of the hierarchical structure of haplotypes, and the consideration of the utility of these approaches toward meeting the end goal of finding genetic variants associated with complex disease.

Workshop: Protein Domains: Identification, Classification and Evolution
Dates: February 27 - 28, 2003
Location: DIMACS Center, Rutgers University, Piscataway, NJ
Organizers: Stephen Bryant, NIH; Teresa Przytycka, Johns Hopkins University
Attendance: 118

Protein domains may be defined as elementary units of protein structure and evolution, capable, to some extent, of folding and functioning autonomously. Understanding of domain structure, function and evolution is a fundamental step towards understanding of a living organism. Identification of protein domains provides an insight into protein function.

Current databases provide a vast amount of protein sequence and structure data, and it is likely that members of the majority of protein families have been already observed. Thus we are in the position to ask a question that could not be addressed previously: Can one use this information to identify domain sequences and structures computationally? How can we extract their properties?

By analyzing the database of known sequences and structures and the properties of conserved domains we hope to get some insight into domain identification, properties and evolution. The amount of the data, it's complexity and biological context open challenges of an interdisciplinary nature. For example, identification of a structural domain requires on one hand sophisticated algorithms that can search for patterns in 3-dimensional data and on the other hand a clear, biologically meaningful and computationally tractable, definition of protein domain.

The workshop was devoted to computational challenges in this new phase of understanding protein domain organization. The goal of the workshop was to bring together biological and computational/mathematical scientists to discuss the state of the art and the open questions focusing on the following aspects of protein domains:

- Methods for identification of protein domains.
- Protein domain comparison and classification.
- Mechanisms of domain evolution.
- Topological and geometrical properties of protein domains.
- Relation between sequence and structure conservation.

Working Group: Mathematical and Computational Aspects Related to the Study of The Tree of Life
Dates: March 11 - 14, 2003
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Melvin F. Janowitz, DIMACS; Francois-Joseph Lapointe, Universite de Montreal; F.R. McMorris, Illinois Institute of Technology; Fred Roberts, DIMACS
Attendance: 26

DIMACS held two working group meetings on Bioconsensus. The general goal of these meetings was to investigate the use of consensus techniques in evolutionary biology, and in particular their applications to phylogenetics. The meetings were highly successful and resulted in the publication of a volume in the DIMACS book series. There was a strong feeling among the participants in these meetings that there should be a third event, but that it should not focus on the specific topic of Bioconsensus. There was some sentiment for a further meeting on supertrees, and more generally for a working group designed to explore algorithmic and mathematical aspects related to the study of the Tree of Life. There is considerable interest in this type of project in both the Biology and the Mathematics community.

Vast quantities of molecular data are becoming available, and there is a need to provide efficient computer algorithms that will appropriately scale to accommodate the size and quality of the underlying data sets. We recognize that for many reasons viral evolution may have features not necessarily present in the evolution of organisms on a multicellular scale. Here are some of the questions addressed by the meeting. First of all should this Tree of Life really be a tree, or is some other data structure a more plausible model? Certainly certain local portions of evolutionary structure should be tree-like, but when these local structures are assembled, should they form a supertree or some more general structure? Here we wished to compare mathematical theory with current research trends in the biological community. We hoped that the biologists would suggest areas they find useful, as opposed to mathematicians just suggesting models of interest to them.

Some other questions that were addressed: How does one handle possible errors in the data? What about missing data? Certainly statistics must play a role here, but we need also to study combinatorial methods, consensus methods, some notion of approximate compatibility analysis, generalized pyramids, and some form of cluster analysis. How, for example, should one deal with reticulate evolution? Should errors in data be dealt with by allowing alternate models of evolution, or by allowing reversals and overlap or is there some other technique dependent upon the data? There is a need to develop better techniques for handling microarray data as well. Here new consensus or classification techniques need to be developed. Thought should also be given to the development of appropriate measures of dissimilarity. Are there any techniques other than string matching of interest to the biological community?

Existing methods of supertree construction should be compared with simulations. The NP-completeness of current techniques should be tested, and where appropriate fast approximate algorithms (heuristics) should be developed. Is there a role for compatibility analysis? What about parsimony? There is a need to compare MRP methods with other methods of constructing a supertree. There is a need even to arrive at an acceptable definition of a supertree. How does one deal with massive data sets to derive an accurate Tree of Life?

Workshop: Medical Applications in Computational Geometry

Dates: April 2 - 4, 2003
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Danny Chen, University of Notre Dame; Jean-Claude Latombe, Stanford University
Attendance: 66

Computer technology plays an increasingly important role in modern medicine and life sciences. Many medical problems are of a strong geometric nature and may benefit from computational geometry techniques. The DIMACS workshop on Computational Geometry and Medical Applications aimed to provide a forum for researchers working in computational geometry, medicine, and other related areas to get together and exchange ideas, and to promote cross-fertilization and collaborations among these areas. The theme of the workshop was the exploration of the applicability of computational geometry to medical problems and the new challenges posed by current medical research and practice to geometric computing. Examples of topics included surgical simulation and planning, geometric representation and modeling of medical objects and human-body tissue structures, geometric problems in medical imaging, computational anatomy, registration and matching of medical objects, etc.

BioMaPS/DIMACS Tutorial: Introduction to Modern Concepts in Biology for Mathematical and Physical Scientists

Dates: June 23 - July 3, 2003
Location: DIMACS Center, CoRE Buildingm Rutgers University
Organizers: William Sofer and Paul Ehrlich, Rutgers University
Attendance: 40

The BioMaPS/DIMACS tutorial "Introduction to Modern Concepts in Biology for Mathematical and Physical Scientists" was a 2-week intensive program divided into three related sections:

1) molecular biology, genetics, and biotechnology
2) cell biology,
3) presentations of current research

The tutorial introduced participants to topics in molecular and cell biology that are relevant to those who wish to explore or initiate work at the interface among biology, mathematics, computer science, chemistry, and physics. The tutorial was appropriate for graduate students, postdocs, and mid-career scientists who wished to change direction. The first week of the tutorial, entitled "The DNA Revolution," was designed by William Sofer, Professor of Genetics and member of the faculty of the Waksman Institute at Rutgers University, to introduce the fundamentals of modern molecular biology, genetics, and biotechnology to participants who have little background in biology or biochemistry. In addition, the tutorial examined scientific questions that interest present-day biomedical researchers. The emphasis was on problems that could be addressed through quantitative approaches that are currently considered the domain of physics, chemistry, mathematics, or computer science. Lively discussions at the level of postgraduate training of physical scientists and mathematicians occurred during these sessions.

Although the first week was focused on providing participants with a clear understanding of the major issues and accomplishments of modern biology, participants also acquired an essential factual background through the use of an interactive "Internet Textbook." This primer of biological information covered important chemical structures and biochemical processes as well as classical genetics. Building on this foundation were descriptions of several methods and research projects in genetic engineering and bioinformatics. This material equipped participants with the framework for understanding most other relevant technologies. In addition, a virtual laboratory was designed to offer participants realistic exercises that illustrate the day-to-day activities of modern-day bench scientists. Finally, participants were given the opportunity to develop bioinformatics research projects as part of the "hands-on" approach to training that Professor Sofer believes is most effective.

The second week of this tutorial started with a review of cell biology that included lectures given by Paul Ehrlich, Adjunct Professor of Cell Biology and member of the BioMaPS Institute at Rutgers University. He examined energy generation, cell division including cell cycle control, cell communication including adhesion, and apoptosis as well as the structure and function of the cytoskeleton, membranes, and intracellular compartments. Basic information was supplemented with a summary of open questions published by leaders of each discipline. In addition, major topics were illustrated by an in-depth analysis of recently published biophysical studies that have made a significant contribution to the understanding of biological phenomena, e.g., mathematical models of cell and viral kinetics that explain the emergence of drug-resistant viruses despite the use of highly effective drugs. Finally, participants became familiar with important Websites useful to researchers at the interface of biology with the mathematical and physical sciences.

The second week ended with presentations on current research in fields at the interface among the biological, mathematical, and physical sciences. The first two parts of the tutorial provided the participants with adequate preparation to gain insight and understanding from these presentations. They were given by faculty of DIMACS and the BioMaPS Institute at Rutgers University.

Workshop: Data Mining Techniques in Bioinformatics
Dates: October 30 - 31, 2003
Location: DIMACS Center, CoRE Building, Rutgers University

Organizer: Mona Singh, Princeton University; Mark Gerstein, Yale University
Attendance: 127

In recent years, high-throughput experimental methods in molecular biology have resulted in enormous amounts of data, both in terms of volume and in terms of new types of data. In addition to complete genome sequences, we have gene expression data, protein structural data, protein-protein interaction data, and protein-DNA interaction data. Data mining techniques will play a large role in analyzing and integrating these large biological datasets, as well as in discovering the biological processes underlying these data.

This workshop brought together researchers in both data mining and bioinformatics, with the aim of exploring new techniques and insights for deciphering biological databases.

Workshop on Information Processing in the Biological Organism (A Systems Biology Approach)
Dates: November 4-5, 2003
Location: Four Points Sheraton, Bethesda, Maryland
Organizers: Fred S. Roberts, DIMACS; Eduardo Sontag, Rutgers University

Note that this workshop was part of the special focus but was supported under a separate NSF grant.

This workshop investigated information processing in biological organisms from the general point of view of systems biology. Traditional biological research has aimed to understand isolated parts of a cell or organism. It has achieved dramatic technological breakthroughs in understanding genes and proteins. The potential for dramatic new biological knowledge arises from investigating the complex interactions of many different levels of biological information. This is the heart of the systems approach to biology. It is aimed at studying genomic DNA, mRNA, proteins, and informational pathways and informational networks in conjunction, looking for "system-level" understanding. Systems biology requires an understanding of the basic system structures (the networks of gene interactions and biochemical pathways) and it requires an understanding of the dynamics of systems - how they change over time through metabolic changes, modifications in biochemical makeup, etc. Understanding biological systems from this point of view can be greatly aided by the use of powerful mathematical and computer models. In turn, the systems understanding of biological systems can provide insights that might be useful for computer and information science.

Information processing is a key aspect of biological systems and the workshop concentrated on this aspect of systems biology. In general terms, the workshop sought to answer the following questions:
(1). What model systems concerning information processing in the biological organism are of broad utility to systems biology?
(2). What are the mathematical foundations relevant to the systems biology approach to the study of information processing in biological organisms, and specifically what algorithms are of broad and central utility for this topic?
(3). What is the next generation of reference resources needed for research in this field?
(4). How can the study of selected information processing processes within biological systems inform other disciplines, including computer science?

The workshop was organized around four main topics:
(a). Genetics to gene-product information flows, including temporal and spatial aspects.
(b). Signal fusion within the cell.
(c). Cell-to-cell communication.
(d). Information flow at the system level, including environmental interaction.

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

Dates: December 10 - 11, 2003
Location: DIMACS Center, CoRE Building, Rutgers University
Organizer: Paul Ehrlich, BIOMAPS Institute; Mel Janowitz, DIMACS; Tara Matise, Department of Genetics
Attendance: 111

The National Center for Biotechnology Information (NCBI) presented 'A Field Guide to GenBank and NCBI Molecular Biology Resources', a lecture and hands-on computer workshop on GenBank and related databases covering effective use of the Entrez databases and search service, the BLAST similarity search engine, genome data and related resources. Further information about NCBI may be found at http://www.ncbi.nlm.nih.gov.

Topics covered included:

- GenBank Database: description and scope
- The NCBI Derivative Databases: RefSeqs
- Database Searching using Entrez
- Neighboring and Links
- Entrez searching
- The NCBI Structures Database
- The Molecular Modeling Database (MMDB)
- Structural Alignments
- Viewing Structures and Structural Alignments with Cn3D
- Similarity Searching using NCBI BLAST
- Local Alignment Statistics
- Scoring Systems
- Using BLAST 2.2.6 web services
- PSI-BLAST
- RPS-BLAST (CDD Search)
- Specialized BLAST pages
- Genomic Resources at NCBI
- Complete Microbial Genomes in Entrez
- Higher Genome Resources
- RefSeq and LocusLink
- UniGene
- Variation Data (SNPs)
- The NCBI Draft Human Genome
- The Map Viewer
- Mouse and Other Genomes

Working Group: New Algorithms for Inferring Molecular Structure from Distance Restraints

Dates: January 12 - 16, 2004
Location: DIMACS Center, CoRE Building, Rutgers University
Organizer: Michael W Trosset, College of William & Mary
Attendance: 43

An important problem in structural molecular biology is the problem of determining 3-dimensional molecular structure using NMR spectroscopy. One critical step in inferring molecular structure from NMR data involves computing a 3-dimensional configuration of points that is consistent with a specified set of lower and upper bounds on the interpoint distances. This step can be formulated, in various ways, as a numerical optimization problem. This research week allowed several researchers who were then developing new algorithms for such problems to interact and collaborate.

Workshop: Interface Between Biology and Game Theory

Dates: April 5, 2004
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Adam Arkin, Lawrence Berkeley Labs and UC Berkeley; Denise Wolf, Lawrence Berkeley Labs; Vijay Vazirani, Georgia Tech
Attendance: 43

Starting with the pioneering work of John Maynard Smith, game theory has been increasingly used to explain, understand, and predict biological phenomena. In a sense, game theory is even more readily applicable to biology than to economics, for which it was initially intended, because the concept of human rationality, a rather uncomfortable assumption, can be replaced by more robust notions such as evolutionary stability.

In recent years, game theory has been used to explain RNA phage dynamics, viral latency, chromosome segregation subversion in sexual species, E. coli mutant proliferation under environmental stress, and aspects of competitive bacterial ecology. At this workshop, the first of its kind, we saw some of the best work being done on this exciting interface, and also discussed future directions for exploring it.

Short Course: Gene Expression Resources at the National Center for Biotechnology Information (NCBI)

Dates: April 13 - 14, 2004
Location: DIMACS Center, CoRE Building, Rutgers University
Organizer: Paul Ehrlich, BIOMAPS Institute; Mel Janowitz, DIMACS; Tara Matise, Department of Genetics
Attendance: 60

This course was intended for principal investigators, senior scientists, thesis directors, laboratory directors, postdoctoral fellows, graduate students, advanced undergraduates and scientific support staff who would like to incorporate gene expression data into their research.

The training course presented by the NCBI was a 1½ hour lecture and a 2 hour computer workshop that demonstrated how the gene expression resources and tools available at the NCBI could be used to obtain qualitative and quantitative measurements of gene expression in different biological samples and under various experimental conditions.

In this course, workshop participants learned how to:

- Query the Gene Expression Omnibus (GEO) database and the two related Entrez databases, GEO Expressions and GEO DataSets to:
    - identify genes that may be regulated together and be part of a common regulatory pathway (profile neighbors);
    - use sequence similarity searching to find expression patterns for genes which may be members of a multi-gene family (sequence neighbors);
    - perform cross-species comparisons of expressed genes.

- Use the Digital Differential Display tool for the comparative analysis of mRNA abundance in cDNA (EST) libraries;

- Use the xProfiler tool to measure mRNA abundance in SAGE libraries;

- Mine the data in the Online Mendelian Inheritance in Man (OMIM) database and NCBI's variation database, dbSNP, to find polymorphisms that may alter gene expression;

- Integrate and visualize expression-related data (UniGene, SAGE tags, CpG islands and variation) in NCBI's genome browser, MapViewer.

Attendance at the lecture was a prerequisite for participation in the computer session.

For further information on Geo see: http://www.ncbi.nlm.nih.gov/geo/

BioMaPS/DIMACS/MBBC/PMMB Short Course: Transcriptional Regulation from Molecules to Systems and Beyond
Dates: June 21 - 25, 2004
Location: DIMACS Center, CoRE Building, Rutgers University
Organizers: Wilma Olson, Rutgers University; Anirvan Sengupta, Rutgers University
Attendance: 70

This short course on transcription was designed to: (1) enable participants with advanced training in the mathematical, computational, and physical sciences, but with a more limited background in biology, to contribute to research at the interface of the biological, mathematical, and physical sciences, (2) introduce participants with traditional backgrounds in biochemistry, genetics, and molecular biology to the potential value of quantitative approaches in their own work, and (3) provide participants with in-depth training in an important subfield within molecular biology. The participants included graduate students, post-doctoral fellows, faculty members, and biomedical researchers from non-academic organizations.

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

1. Basic introduction to transcription 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 transcription by leading scientists and their group members

Professor Richard Ebright of Rutgers presented the foundation required by non-expert researchers for the understanding of gene transcription. This included a review of the structural components of transcription and a basic description of transcription initiation, regulation, elongation, and termination. The introduction to transcription in bacteria laid the groundwork for materials on more complex eukaryotic systems. These lectures were designed to provide participants who had a limited knowledge of molecular biology with a smooth transition to the understanding and appreciation of cutting-edge research.

In the remaining lectures, participants gained an in-depth view of transcription both from the content of the presentations and the wide range of fields that were represented, including bioinformatics, computational and experimental structural biology, molecular biology, statistical mechanics, and systems biology. Participants from Rutgers University and the University of Medicine and Dentistry of New Jersey-Robert Wood Johnson Medical School (UMDNJ-RWJMS) included faculty members in the Departments of Biochemistry, Chemistry & Chemical Biology, Molecular Biology & Biochemistry, Mechanics, Pharmacology, and Physics. Researchers from Cold Spring Harbor Laboratory, Cornell

University, MIT, Princeton University, Rockefeller University, and University of Wisconsin presented their recent findings.

Each day of the program was devoted to different aspects of transcription following, in general, the pathway for producing a complete, freely diffusing RNA polymer. Thus, the first day included a detailed account of the structural components of transcription, including RNA polymerase and chromatin, as well as an overview of eukaryotic gene expression. Subsequent days focused on: 1) initiation and control of transcription; 2) elongation and termination; 3) regulation, bioinformatics, and genomics; and 4) transcriptional network modeling.

The short course "Transcriptional Regulation from Molecules to Systems and Beyond" was part of the second annual Summer School sponsored by the newly established BioMaPS Institute for Quantitative Biology at Rutgers in collaboration with the Center for Discrete Mathematics and Theoretical Computer Science at Rutgers (DIMACS), the Center for Molecular Biophysics and Biophysical Chemistry at Rutgers (MBBC), and the Program in Mathematics and Molecular Biology (PMMB) based at Florida State University. The Sloan Foundation and the Burroughs-Wellcome Fund provided partial funding of the Summer School.

Workshop: Reticulated Evolution

Dates: September 20 - 21, 2004
Location: DIMACS Center, CoRE Building, Rutgers University
Organizer: Mel Janowitz, DIMACS; Bernard Moret, University of New Mexico; Randy Linder, University of Texas
Attendance: 52

Species evolution has long been modeled as a branching process that can uniquely be represented by a tree topology. In such a topology, each species can only be linked to its closest ancestor, while interspecies relationships such as species hybridization or lateral gene transfer in bacteria are not allowed. With the advent of phylogenetic analysis at the molecular level, there is increasing evidence that such a model is inadequate. This workshop explored the history and latest status of these new models of "reticulate evolution", and was coupled with a smaller working group meeting designed to explore promising avenues for future research.

Working Group: Reticulated Evolution

Dates: September 22, 2004
Location: DIMACS Center, CoRE Building, Rutgers University
Organizer: Mel Janowitz, DIMACS; Bernard Moret, University of New Mexico; Randy Linder, University of Texas
Attendance: 21

This working group meeting was coupled with the workshop on the same subject. Its goal was to initiate promising avenues of research designed to explore new models of "reticulate evolution" that are biologically meaningful, and computationally feasible..

III. Project Findings

<u>Summary of some of the research results obtained by special focus participants</u>

Consensus List Colorings of Graphs and Physical Mapping of DNA. N.V.R. Mahadev (Fitchburg State College) and Fred Roberts (Rutgers) studied a bioconsensus problem relating to graph coloring and investigated the applicability of the ideas to the DNA physical mapping problem. In many applications of graph coloring, one gathers data about the acceptable colors at each vertex. A list coloring is a graph coloring so that the color assigned to each vertex belongs to the list of acceptable colors associated with that vertex. Mahadev and Roberts considered the situation where a list coloring cannot be found. If the data contained in the lists associated with each vertex are made available to individuals associated with the vertices, it is possible that the individuals can modify their lists through trades or exchanges until the group of individuals reaches a set of lists for which a list coloring exists. Mahadev and Roberts described several models under which such a consensus set of lists might be attained. In the physical mapping application, the lists consist of the sets of possible copies of a target DNA molecule from which a given clone was obtained and trades or exchanges correspond to correcting errors in data. They showed that there are graphs and list assignments for which not only is there no list coloring, but almost everyone has to add an acceptable color that was not on their original list in order for a list coloring to exist. In physical mapping, this means that there are situations where essentially every list of copies needs to be expanded.

Immunoglobulin Folds

Boris Galitsky (Visitor), Andrey Dementiev and Sergey Shelepin addressed the contradiction that a rather limited number of residues or classes of amino acids (about 10) determines the fold (Immunoglobulin-like) for the sequence that is about 100 residues long. The Immunoglobulin fold comprises the protein super-families with rather distinguishing sequences with less than 10% identity; their sequence alignment can be accomplished only taking into account the 3D structure. Therefore, it is believed that discovering the additional common features of the sequences is necessary to explain the existence of common folds for these (SCOP) superfamilies. The analysis of pair-wise interconnection between residues of the multiple sequence alignment helped these researchers to reveal the set of mutually correlated positions, inherent to almost every super-family of protein fold. Hence, the set of constant positions plus the set of variable but mutually correlated ones can serve as a basis of having the common 3D structure for distinguishing protein sequences.

Statistical Properties of Contact Maps.

A contact map is a simple representation of the structure of proteins and other chain-like macro-molecules. This representation is quite amenable to numerical studies of folding. Michele Vendruscolo, Rutgers graduate student, Balakrishna Subramanian, Ido Kanter, Eytan Domany and DIMACS permanent member Joel Lebowitz (Mathematics, Rutgers) showed that the number of contact maps corresponding to the possible configurations of a polypeptide chain of N amino acids, represented by (N - 1) step half avoiding walks on lattices, grows exponentially with N for all dimensions D > 1. They carried out exact enumerations in D = 2 on the square and triangular lattices for walks of up to 20 steps and found exact statistical properties of contact maps corresponding to such walks.

Comparing Gene Order in Related Species

At the 2001 workshop on Whole Genome Comparison, Dannie Durand, started collaborating with David Sankoff on statistical tests for locally conserved patterns in gene content in comparative maps. Comparing chromosomal gene order in two or more related species is an important approach to studying the forces that guide genome organization and evolution. Linked clusters of similar genes found in related genomes are often used to support arguments of evolutionary relatedness or functional selection. However, as the gene order and the gene complement of sister genomes diverge progressively due to large scale rearrangements, horizontal gene transfer, gene duplication and gene loss, it becomes increasingly difficult to determine whether observed similarities in local genomic structure are indeed remnants of common ancestral gene order, or are merely coincidences. A rigorous comparative genomics requires principled methods for distinguishing chance commonalities, within or between genomes, from genuine historical or functional relationships. Durand and Sankoff constructed tests for significant groupings against null hypotheses of random gene order, taking incomplete clusters, multiple genomes and gene families into account. They considered both the significance of individual clusters of pre-specified genes, and the overall degree of clustering in whole genomes. This resulted in a Sixth Annual International Conference on Computational Molecular Biology (RECOMB) paper in 2002. They were later invited to submit an extended version of this paper to a special issue of Journal of Computational Biology. As a result of this work, Durand was invited to write a paper in Trends in Genetics. (See Papers.)

Automatic Methods for Discovering Related Gene Sequences in Different Organisms
To understand evolutionary relationships among genes from different organisms is a problem in modeling evolutionary history while solving practical problems related to functional annotation of genes. Akshay Vashist, Casimir A. Kulikowski, and Ilya Muchnik have developed an automatic method for discovering groups of gene sequences present in different organisms that are functionally related through evolution. They have developed a new clustering method, which allows them to build clusters from multi-component types of data. In their case, the data is a large set of genomes in which one has to find clusters that are groups of orthologous genes, focusing on hyper-inter-similarities among genes from different genomes more than the intra-similarities among genes from the same genome. They have found that discovering these groups provides a "strong draft" of the complete picture of orthologous relations among genes in the complete genomes studied. Comparisons of these groups with the well-known semi-automatically extracted clusters of orthologous groups, COG, http://www.ncbi.nlm.nih.gov/COG/, shows strong correlation between these two systems of clusters. For instance, more than 85% of their clusters include genes from at least three different genomes and each of these genes belongs to COGs. These studies demonstrate that the method can be applied for an automatic screening of groups of orthologous genes in analyzing a large collection of genomes from different organisms. Vashist, Kulikowski, and Muchnik are currently working towards application of the method on real biological data.

DNA Bending
For many years John Widom (NWU) worked on the problem of nucleosome positioning, trying to find sequence determinants for preferable positions of nucleosomes along DNA molecules. After discussions between Widom and Alex Vologodskii (Department Of Chemistry, New York University) at one of the Computational Molecular Biology workshops, they decided to investigate the efficiency of cyclization of short DNA fragments with different affinities to nucleosome binding. Unexpectedly, Widom found that very short DNA fragments, about 100 bp in length, can be cyclized much easier than it was expected according to the generally accepted

theory. Careful theoretical and experimental investigation of this phenomenon has become Vologodskii's top priority.

Hiding Messages in DNA Microdots

Attending workshops in this special focus stimulated Carter Bancroft's thinking about the technological uses of DNA. That thinking, in turn, led to the new research idea, DNA-based steganography. Steganography, concealing messages in a microdot, was developed by Professor Zapp and used by German spies in the Second World War to transmit secret information. A microdot is a greatly reduced photograph of a typewritten page that is pasted over a full stop in an innocuous letter. Catherine Taylor Clelland, Viviana Risca, and Carter Bancroft have taken the microdot a step further and developed a DNA-based, doubly steganographic technique for sending secret messages. A DNA encoded message is first camouflaged within the enormous complexity of human genomic DNA and then further concealed by confining this sample to a microdot.

A Genomewide Examination of Molecular Evolution

Cristian I. Castillo-Davis (Department of Statistics, Harvard University) first learned about the software GeneSeqer, which allows one to predict gene structure from genomic DNA using the coding sequence of a related organism, as a graduate student attending the Workshop on Whole Genome Comparison in 2001. He used it to predict genes in the unannotated genome of Caenorhabditis briggsae using genes known in the completely sequenced and annotated genome of Caenorhabditis elegans. This led to his carrying out a genomewide examination of molecular evolution through ontogeny using comparative genomic data in Caenorhabditis elegans and Caenorhabditis briggsae to test the hypothesis that evolutionary changes will be more frequent in later ontogeny than early ontogeny because of developmental constraint. He found that the mean rate of amino acid replacement is not significantly different between genes expressed during and after embryogenesis. However, synonymous substitution rates differed significantly between these two classes. Surprisingly, it was found that genes expressed after embryogenesis have a significantly greater number of duplicates in both the C. elegans and C. briggsae genomes when compared with early-expressed and nonmodulated genes. A similarity in the distribution of duplicates of nonmodulated and early-expressed genes, as well as a disproportionately higher number of early pseudogenes, lend support to the hypothesis that this difference in duplicate number is caused by selection against gene duplicates of early-expressed genes, reflecting developmental constraint. Developmental constraint at the level of gene duplication may have important implications for macroevolutionary change.

Inferring Piecewise Ancestral History from Haploid Sequences.

Russell Schwartz (Department of Biological Sciences, Carnegie Mellon University) developed methods for modeling haplotype patterns in human genetic sequences and applying them to optimization problems involved in finding genetic factors that influence disease risk. This followed up on previous work he had done and it also described novel research he had prepared specifically for a DIMACS workshop.

Aligning Multiple Sequences

Multiple sequence alignment is usually considered as an optimization problem, which has a statistical and a structural component. It is known that in the problem of protein sequence alignment, a processed sample is too small and not representative in the statistical sense, though this information can be sufficient if an appropriate structural model is used. A new structural description of the pairwise alignment results union was developed by Leonid Shvartser, Casimir Kulikowski, and Ilya Muchnik. They showed that if the structure is restored then multiple sequence alignment is achieved. Introduced structure represents the set of local maximums of a quasi-concave set function on a lower semi-lattice, which in turn is a union of the set-theoretical intervals. They developed an algorithm for local maximums search on a proposed structure consisting of an alternation of the Forward and Backward passes. The Backward pass in this algorithm is rigorous while the Forward pass is based on heuristics. Multiple alignment of 5 protein sequences were used as an illustration of the algorithm.

Reconstructing the Spatial Structure of Proteins from Their Amino Acid Sequences
The result of the stay at DIMACS of Vadim Mottl is a new approach to the problem of reconstructing the spatial structure of proteins from their primary structure in the form of amino acid sequences. The new approach was presented at the Working Group Meeting on Informatics for Protein Classification that took place on December 15, 2000. In contrast to the traditional methodology that is aimed at finding a specific spatial structure of each particular protein, the new method exploits the fact that the fold pattern remains basically the same within large groups of evolutionarily allied proteins, so that the "number" of essentially different spatial structures is much less than that of known proteins. Since spatial structures are classified, the estimation of the spatial structure of a given protein reduces to a search over a finite set of classes, i.e. the problem falls into the competence area of pattern recognition. Mottl collaborated with Ilya Muchnik (DIMACS), Casimir Kulikowski (Rutgers Department of Computer Science), Michail Roytberg (DIMACS visitor, Institute for Mathematical Problems of Biology, Russia) and Leonid Shwartser (DIMACS Visitor, NESS A.T. Ltd-TSG, Tel-Aviv, Israel) to create a new featureless method of pattern recognition specially for this purpose. Judgment about the membership of a protein in one of the classes of spatial structures is made immediately on the basis of measuring the proximity of its amino acid chain to those of some other proteins whose spatial structure is known.

On The Inter-Residue Correlation Patterns And Their Role In Classification of Protein Families.
Boris Galitsky (visitor) and Sergey Shelepin built a novel method to calculate and analyze the correlations in mutational behavior between different positions in a multiple sequence alignment. The inter-dependence between the residues for a protein family was represented as a matrix of correlation values obeying the invariance with respect to specific amino acids, the number of sequences representing a family, the length of sequences, residue variability and the uniformity of data set representation. Common and distinguishing properties of the few protein families, including immunoglobulins, were revealed, based on the geometry of correlation matrices. Galitsky and Shelepin analyzed the specific texture of these matrices, inherent to the specific families, and suggested a way to distinguish protein from non-protein set of sequences. The role of correlation matrix technique in classification was investigated and Galitsky and Shelepin suggested that the classification criteria should be based on the residues at the positions with the highest overall correlation with the other positions. Revealing the positions with various correlation strengths helps to reconstruct the phylogeny of protein families.

Research results from graduate students

Predictive Gene/Protein Selection for the Logical Analysis of Cancer Data

Gabriela Alexe's research at Rutgers was oriented to the selection of variables in logical analysis of genomic and proteomic cancer datasets. Usually, these datasets contain several hundreds of positive and negative cases depending on tens of thousands of variables. A central objective of this study was to identify criteria that are able to measure the importance of variables in the logical analysis of microarray datasets, and to use them for the selection of minimal subsets of variables (support sets), based on which accurate classification models can be developed. The main research questions she worked on were:

> (1) Find (minimal) support sets of variables distinguishing the positive cases from the negative ones;
>
> (2) Determine the most important variables for the prediction of the positive/ negative class.

In order to answer these questions, she developed in collaboration with Peter L. Hammer, Bela Vizvari, and Sorin Alexe, a LAD-based method for variable selection. The proposed method is a two-stage procedure. In the first stage, they filtered the attributes based on several independent "separability"-type criteria, and retained a sufficiently large subset (e.g., hundreds of attributes) of the highest ranked ones. Using this subset, in the second stage, they constructed a large collection of high quality patterns, and retained only those attributes that are frequently used for the description of these patterns. Finally, in order to reduce the number of selected attributes to a reasonable size, a minimal support set procedure might be applied. The performance of the attribute selection method within the LAD modeling process was evaluated through multiple computational experiments on several genomics and proteomics publicly available datasets, including the proteomic ovarian dataset (http://clinicalproteomics.steem.com), the microarray breast cancer dataset (http://www.rii.com), and the microarray leukemia dataset (http://www-genome.wi.mit.edu/mpr/data_set_ALL_AML.html). The fact that the LAD classification models based on the attributes selected using the 2-stage method are highly accurate supports the idea that the set of selected attributes could be used in conjunction with other data analysis techniques.

A Novel Machine Learning Classification Method

Akshay Vashist (Rutgers), working with Ilya Muchnik (DIMACS), developed a novel method for machine learning classification focused on validation of accuracy estimates, through an iterative multi-stage process of learning to stratify training objects in an ordinal scale of their probability to produce incorrect predictions. The basic idea of the method is to find "competitive" objects from different classes through cross-validation analysis, join them temporarily in one class, learn the classifier on this biased data, and repeat the process until either the considered class will be clasified 100% correctly or it will disappear. Such an iteration process induces a stratification on the class, dividing it into ordered parts according to how well they are classified correctly. Vashist found that such a multi-stage learning process is critically important when labels of classes in training data have noise (when expert provided labels have high uncertainity). The gene functional annotation problem is a good example of this situation. Vashist applied the developed method for such problems to classify Rice genome genes from chromosome 10 into three classes: (1) protein coding regions, (2) polyproteins, (3) transposable elements.

Simulating Immune Response

Erich R. Schmidt (Princeton) worked on simulating the immune response taking place in the germinal centers. The questions addressed by his research were: mathematical models, computational performance, and results analysis.

Pattern-Based Scoring Systems in Logical Analysis of Data
Logical Analysis of Data (LAD) is a method for building pattern-based classification models, by extracting and aggregating knowledge from a dataset, consisting of positive and negative observations. Patterns or rules (i.e., knowledge) are the central concept of LAD, and they are used to classify new observations in the following way: if the observation is covered only by positive (negative) patterns, it is classified as positive (negative), If no patterns cover the observations, or it remains simply unclassified, and if both positive and negative patterns cover the observations, a discriminant function is constructed and used for classification. The last situation can appear frequently in a dataset and it has been shown that a simple aggregation of patterns, based on “prognostic index”, i.e., the difference between the proportion of the positive patterns covering the observation and that of the negative patterns covering it, can provide highly accurate classification models. A natural question is if other types of aggregation of the knowledge extracted from a dataset can be effective. Sorin Alexe (Rutgers) worked on this question. The first part of his study proposed several methods for the evaluation of the performance of classification models, and proved their equivalence. This defined the goal of finding good discriminant functions (in pattern space). Eight different types of linear discriminant functions were proposed, each of them providing intuitive reasons for the construction of high quality classification models: two constructed using directly the characteristics of the patterns (e.g., prevalence), three based on optimization of linear programs, two on the optimization of non-linear programs, and one based on the correlation between the observation and an “ideal” one. It was proved that the “correlation” type discriminant provides exactly the same classification as one of the “direct” ones. Some of these discriminants provide classification models that are similar to those obtained by support vector machine, linear regression or artificial neural networks, applied to the pattern space rather than to the original datasets. Sorin Alexe, in collaboration with Peter Hammer (Rutgers faculty member), carried out computational experiments using the LAD standard techniques for pattern detection and Matlab solver for the optimization programs. A benchmark of four publicly available datasets was used for performing a 10-time twofold cross-validation procedure, and estimating the model performance (averaging the accuracy of the 10 experiments for each of the datasets). The results showed that one of the “direct” types of discriminants and one of the “linear” types perform better then the other ones.

Clustering with Short Input
Jie Chen (Princeton) and Andrea LaPaugh (Princeton faculty member) developed a clustering algorithm able to use much shorter pieces of text than using whole documents. Besides improving the time factor, they also proposed a new evaluation measurement to evaluate how good the algorithm can cluster using such short input. In contrast with those algorithms existing in the literature, their evaluation measurement is easy to calculate, gives an exact quantitative measurement, and focuses on individual clusters.

Clustering Analysis on Data from Newborns
Jin Ma (Rutgers) performed data analysis and mining on the data set collected from the newborn child-screening project conducted in Bavaria, Germany. The data had about 20,000 objects (children) with 175 treatment features. Results revealed 6 unique groups of features that

distinguished 20,000 data into normal (about 17,000) and abnormal (about 3,000) clusters. Risk clusters were also generated to signal potential abnormal objects. Clustering using strong connected components was also performed.

The Evolution of Epidemics
Jaewook Joo (Rutgers) and Joel L. Lebowitz (Rutgers faculty member) investigated the time-evolution and steady states of the stochastic susceptible-infected-recovered-susceptible (SIRS) epidemic model on one- and two- dimensional lattices. They compared the behavior of this system, obtained from computer simulations, with those obtained from the mean-field approximation (MFA) and pair-approximation (PA). The former (latter) approximates higher order moments in terms of first (second) order ones. They found that the PA gives consistently better results than the MFA. In one dimension the improvement is even qualitative.

IV. Project Training/Development

Twenty-two graduate students worked on small winter and summer research projects related to this special focus.

Eleven of these graduate students were from Rutgers University:

> Gabrielle Alexe, RUTCOR, Summer 2002: Predictive gene/protein selection for the logical analysis of cancer data
>
> Sorin Alexe, RUTCOR, Summer 2002: Pattern-based scoring systems in logical analysis of data
>
> Khaled Elbassioni, Computer Science, Winter 2001-2002: Learning monotone binary functions in products of lattices and its applications in data mining
>
> German Encisco, Mathematics, Summer 2003: Monotonicity
>
> Rohan Fernandes, Computer Science, Summer 2003: Sorting with length-weighted reversals
>
> Jaewook Joo, Physics, Fall 2002 semester and Spring 2003 semester: The effect of the asymmetric exchange on the low dimensional heterogeneous contact process
>
> Chan-Su Lee, CAIP Center, Summer 2004: Analyze protein structures and find rich low dimensional representation preserving similarities between embedded sequences
>
> Zhong Jun Luo, Computer Science, Winter 2001-2002: Sequence of triple helix-forming TFOs binding site can be screened using computer programming
>
> Jin Ma, Computer Science, Winter 2002-2003: Data analysis and mining performed on the data set collected from newborn child-screening project performed in Bavaria.
>
> Sandor Szesmak, RUTCOR, Summer 2000: New approach to data analysis
>
> Akshay Vashisht, Computer Science, Summer 2002: Structural analysis of repeat elements

Four Princeton graduate students worked on research projects related to this special focus.

Jie Chen, Computer Science, Winter 2000-2001: Clustering using anchor texts of web pages and its applications to improve Jon Kleinberg's link analysis algorithm and Winter 2003-2004: Gene clusterability in gene ontology

Carl Kingsford, Computer Science, Summer 2004: Developing a computational method for predicting protein-protein interactions in yeast using evolutionary information

Steven Kleinstein, Computer Science, Summer 2001 and Summer 2002: Lymphatic tissues and germinal centers in immune reactions

Erich Schmidt, Computer Science, Summer 2001: Simulating immune response

Seven graduate students were recipients of the joint DIMACS - Celera Genomics/ Applied Biosystems Graduate Student Award in Computational Molecular Biology. Each presented a talk at a DIMACS workshop. Recipients and their presentations were:

Ziv Bar-Joseph, MIT, Celera Genomics, 2003: Structure of Gene Expression, talk presented at Workshop: Data Mining Techniques in Bioinformatics: October 30 - 31, 2003.

Barry Cohen, SUNY - Stoney Brook, Celera Genomics, 2001: The Space of RNA Encodings of a Target Protein, talk presented at workshop on Whole Genome Comparison: March 2, 2001.

Jessica Fong, Princeton University, Celera Genomics 2004: Presentation on Predicting Protein Interactions.

Matt Menke, MIT, Celera Genomics 2004: Predicting Protein Folds from Sequence Data

Itsik Pe'er, Tel Aviv University, Celera Genomics 2001: Computational Resequencing by Universal Microarrays, talk presented at Workshop on Analysis of Gene Expression Data: October 25, 2001.

Mihaela Pertea, Johns Hopkins University, Celera Genomics 2001: Gene Finding in Eukaryotes, talk presented at Workshop on Integration of Diverse Biological Data: June 21, 2001.

Scott Rifkin, Yale University, Celera Genomics 2002: Structure of Gene Expression, talk presented at Workshop on Complexity in Biosystems: Innovative Approaches at the Interface of Experimental and Computational Modeling: April 9, 2002

Selected research results are described in more detail in the section on Project Findings. Publications and conference presentations are given in the section Papers/Books/Internet.

V. Outreach Activities

Special Focus visitors, graduate students, and senior faculty were available to interact with 2- and 4-year college faculty in the DIMACS “Reconnect” program, and with high school teachers in the DIMACS Connect Institute and the DIMACS Bio-Math Connect Institute (BMCI). This special focus led to a major emphasis on the interface between the mathematical and biological sciences in the high schools. One major outcome of this was the summer program, BMCI, involving 35 high school teachers in

Summer 2004, being exposed to topics in Computational Biology and Bioinformatics. For more on BMCI, see the website http://dimacs.rutgers.edu/dci/2004/. BMCI led to three technical reports and five classroom modules prepared by high school teachers on the following Computational Biology topics:

- Shortest Common Superstring
- Spikes, Speckles, Canyons and Craters: Noise Reduction in MAGIC Tool Images
- Oligo Cross-Hybridization and Microarray Analysis
- BioMatrices
- The Riddle of the Genes: A Biomathematics Module
- Occam's Razor: Combinatorics and DNA
- Genetic Inversion Rearrangement: Are You a Rotated Mouse?
- Your Condition is Conditional

A second major outcome was the plan for a national conference on the linkage between the mathematical and biological sciences in the high schools. See http://dimacs.rutgers.edu/Workshops/Biomath/ for detailed plans for the conference, to be held April 29 – 30, 2005.

VI. Papers/Books/Internet

Books

Bioconsensus, M. F. Janowitz, F.-J. Lapointe, F. R. McMorris and F. S. Roberts (eds.), DIMACS Book Series, **61**, 2003.

W. Day, and F. R. McMorris, *Axiomatic Consensus Theory in Group Choice and Biomathematics*, Frontiers in Applied Mathematics Series of SIAM, 2003.

Papers

S. Bastea, R. Esposito, J.L.Lebowitz and R. Marra, “Hydrodynamics of binary fluid phase segregation,” *Physical Review Letters*, **89** (2002), 235-701.

C.I. Castillo-Davis and D. L. Hartl, “Genome evolution and developmental constraint in Caenorhabditis elegans”, *Mol. Biol. Evol.*, **19** (2002), 728-735.

E. Carlen, M. Carvalho, R. Esposito, J.L.Lebowitz and R. Marra, “Free energy minimizers for a two-species model of segregation and liquid-vapor transition,” *Nonlinearity 2003*, accepted.

T.E. Cloutier and J. Widom, “Spontaneous sharp bending of double-stranded DNA”, *Mol. Cell.*, **14** (2004), 355-362.

D. Durand and D. Sankoff, “Tests for gene clustering”, Sixth Annual International Conference on Computational Molecular Biology (RECOMB), (2002), 144-154.

D. Durand and D. Sankoff, “Tests for gene clustering”, *Journal of Computational Biology*, **10** (2003), 453-482.

D. Durand, “Vertebrate evolution: Doubling and shuffling with a full deck”, *Trends in Genetics*, **19** (2003), 2-5.

G. Eichler, S. Huang, and D.E. Ingber, “Gene expression dynamics inspector (GEDI): A program for integrated analysis of expression profiles”, *Bioinformatics*, **19** (2003), 2321-2322.

G.A. Enciso, "On the stability of a model of testosterone dynamics", *J. Math. Biology*, to appear.

S. Huang and D.E. Ingber, “From stem cell to functional tissue architecture: What are the signals and how are they processed?” in: S. Sell (ed.), *Stem Cell Handbook*, 2003, 45-56.

S. Huang, C. Sultan, and D.E. Ingber, “Tensegrity, dynamic networks and complex systems biology: Emergence in structural and information networks within living cells”, in T.S. Deisboeck, J.Y. Kresh and T.B. Kepler (eds.), *Complex Systems Science in Biomedicine*, Kluwer Academic Publishers, New York, in press.

D.E. Ingber, “Tensegrity II. How structural networks influence cellular information processing networks”, *J. Cell Sci.*, **116** (2003), 1397-1408.

D.E. Ingber, “The mechanochemical basis of cell and tissue regulation, *Mechanics & Chemistry of Biosystems*, **1** (2004), 55-68.

S.P. Imberman, B. Domanski, and H.W. Thompson, “Using dependency/association rules to find indications for computerized tomography in a head trauma dataset,” *International Journal of Artificial Intelligence in Medicine*, (2002), 55-68

S.P. Imberman, “The KDD process and data mining for computer performance professionals, *Journal of Computing Resources*, **107** (2002), 68-77.

N.V.R. Mahadev and F.S. Roberts, “Consensus List Colorings of Graphs and Physical Mapping of DNA,” in M. Janowitz, F-J. Lapointe, F.R. McMorris, B. Mirkin, and F.S. Roberts (eds.), *Bioconsensus*}, DIMACS Volume 61, American Mathematical Society, Providence, RI, 2003.

B. Mirkin and I. Muchnik, “Induced layered clusters, hereditary mappings, and convex geometries”, *Applied Mathematics Letters*, **15** (2002), 293-298.

R. Schwartz, A.G. Clark, and S. Istrail, “Inferring piecewise ancestral history from haploid sequences”, *Lecture Notes in Bioinformatics* (2004), **2983**, 62-73.

Technical Reports

Gabriela Alexe, Sorin Alexe, Yves Crama, Stephan Foldes, Peter L. Hammer and Bruno Simeone, “Consensus Algorithms for the Generation of All Maximal Bicliques”, DIMACS Technical Report Series, 2002-52.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2002/2002-52.html

Gabriela Alexe and Peter L. Hammer, “Spanned Patterns for the Logical Analysis of Data”, DIMACS Technical Report Series, 2002-50.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2002/2002-50.html

Sorin Alexe, Eugene Blackstone, Peter L. Hammer, Hemant Ishwaran, Michael S. Lauer, Claire E. Pothier Snader, “Coronary Risk Prediction by Logical Analysis of Data”, DIMACS Technical Report Series, 2002-11.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2002/2002-11.html

S. Bastea, R. Esposito, J. L. Lebowitz and R. Marra, “Binary Fluids with Long Range Segregating Interaction I: Derivation of Kinetic and Hydrodynamic Equations”, DIMACS Technical Report Series, 2000-31.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2000/2000-31.html

Sorin Bastea, Raffaele Esposito, Joel L. Lebowitz, and Rossana Marra, “Hydrodynamics of Binary Fluid Phase Segregation”, DIMACS Technical Report Series, 2002-51.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2002/2002-51.html

Endre Boros, Khaled Elbassioni, Vladimir Gurvich, Leonid Khachiyan and Kazuhisa Makino, “Dual-Bounded Generating Problems: All Minimal Integer Solutions for a Monotone System of Linear Inequalities”, DIMACS Technical Report Series, 2001-12.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2001/2001-12.html

Endre Boros, Khaled Elbassioni, Vladimir Gurvich and Leonid Khachiyan, “An Inequality for Polymatroid Functions and its Applications”, DIMACS Technical Report Series, 2001-14.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2001/2001-14.html

Endre Boros, Khaled Elbassioni, Vladimir Gurvich and Leonid Khachiyan, “Generating Dual-Bounded Hypergraphs”, DIMACS Technical Report Series, 2002-23.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2002/2002-23.html

Endre Boros, Khaled Elbassioni, Vladimir Gurvich and Leonid Khachiyan, “Extending the Balas-Yu Bounds on the Number of Maximal Independent Sets in Graphs to Hypergraphs and Lattices”, DIMACS Technical Report Series, 2002-27.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2002/2002-27.html

Endre Boros, Khaled Elbassioni, Vladimir Gurvich and Leonid Khachiyan, “On the Complexity of Some Enumeration Problems for Matroids”, DIMACS Technical Report Series, 2003-17.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2003/2003-17.html

Endre Boros, Khaled Elbassioni, Vladimir Gurvich and Leonid Khachiyan, “On Enumerating Minimal Dicuts and Strongly Connected Subgraphs”, DIMACS Technical Report Series, 2003-35.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2003/2003-35.html

E. Boros, K. Elbassioni, V. Gurvich and L. Khachiyan, “A Global Parallel Algorithm for Finding All Minimal Transversals of Hypergraphs of Bounded Edge-size”, DIMACS Technical Report Series, 2004-31.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2004/2004-31.html

Alair Pereira do Lago, “Local Groups in Free Burnside Groupoids”, DIMACS Technical Report Series, 2001-22.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2001/2001-22.html

K.M. Elbassioni, “An Algorithm for Dualization in Products of Lattices”, DIMACS Technical Report Series, 2002-26.
http://dimacs.rutgers.edu/TechnicalReports/2002/2002-26.ps.gz

German A. Enciso and Eduardo D. Sontag, “A Note on a Monotone Small Gain Theorem, with Applications to Delay Systems”, DIMACS Technical Report Series, 2004-29.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2004/2004-29.html

Yuri Goncharov, Ilya Muchnik and Leonid Shvartser, “Simultaneous Feature Selection And Margin Maximization Using Saddle Point Approach”, DIMACS Technical Report Series, 2004-08.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2004/2004-08.html

Jaewook Joo and Joel L. Lebowitz, “Pair Approximation of the Stochastic Susceptible-infected-recovered-susceptible Epidemic Model on the Hypercubic Lattice”, DIMACS Technical Report Series, 2004-13.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2004/2004-13.html

Jaewook Joo and Joel Lebowitz, “Behavior of SIS Epidemics on Heterogeneous Networks with Saturation”, DIMACS Technical Report Series, 2004-14.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2004/2004-14.html

Vadim Mottl, Sergey Dvoenko, Oleg Seredin, Casimir Kulikowski, Ilya Muchnik, “Alignment Scores in a Regularized Support Vector Classification Method for Fold Recognition of Remote Protein Families”, DIMACS Technical Report Series, 2001-01.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2001/2001-01.html

Vadim Mottl and Ilya Muchnik, “Serial and Tree-serial Dynamic Programming with Application to Fact Identification”, DIMACS Technical Report Series, 2002-45.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2002/2002-45.html

Leonid Shvartser, Casimir Kulikowski, Ilya Muchnik, “Multiple sequence alignment using the quasi-concave function optimization based on the DIALIGN combinatorial structures”, DIMACS Technical Report Series, 2001-02.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2001/2001-02.html

Akshay Vashist, Casimir A. Kulikowski, Ilya Muchnik, “Automatic Screening for Groups of Orthologous Genes in Comparative Genomics Using Multiple-component Clustering, DIMACS Technical Report Series, 2004-33.
http://dimacs.rutgers.edu/TechnicalReports/abstracts/2004/2004-33.html

Posters

A. Handzel, Senior Scientist, Beyond Genomics, Inc.
IEEE sponsered Workshop on Genomic Signal Processing and Statistics (GENSIPS):
http://www.cis.jhu.edu/gensips2004/

A. Handzel, Senior Scientist, Beyond Genomics, Inc.
ISMB/ECCB 2004: http://www.iscb.org/ismb2004/posters/ahandzelATbeyondgenomics.com_146.html

Talks and Special Sessions

A. Handzel, “System Biology in Practice: Science, Technology, Challenges”**,** Institute for Systems Research, University of Maryland, May 28, 2004.

A. Handzel, co-organizer, Special Session on Signal Processing in the Post-genomic Era, IEEE Annual Signal Processing Conference (ICASSP 2005), pending approval.

B. Mirkin, Clusters in proteomics and genomics, UCL-Birkbeck Meeting on

Bioinformatics, London, 13 March 2002

V. Mottl, “Alignment scores in a regularized support vector classification method for fold recognition of remote protein families,” Working Group Meeting on Informatics for Protein Classification, December 15, 2000.

F. S. Roberts, "Voting, Metasearch, and Bioconsensus," NYAcademy of Science Distinguished Scientist Lecture, May 2001.

F. S. Roberts, "Voting, Metasearch, and Bioconsensus," AAAS Annual Meeting, Boston, MA, February 2002.

F. S. Roberts, "Consensus List Coloring and Physical Mapping of DNA," International Conference on Ordinal and Symbolic Data Analysis, Irvine, CA, August 2003.

F. S. Roberts, "Consensus List Coloring and Physical Mapping of DNA," DIMACS-DIMATIA-Renyi Meeting on Generalizations of Graph Coloring, October 2003.

F. S. Roberts, "The RNA Detective Game: Finding RNA Chains From Fragments," DIMACS Biomath Connect Institute Meeting, July 2004.

L. Segel, Organizer, Mathematical and Computer Models in Medicine: Disease and Treatment, Santa Fe Institute (SFI), Santa Fe, New Mexico, July 25 - August 6, 2004

R. Schwartz, “Comparative structural and functional genomics using weak prediction methods”, DIMACS Workshop on Protein Structure and Structural Genomics, March 8-9, 2001

R. Schwartz, “Inferring piecewise ancestral history from haploidsequences", DIMACS/RECOMB Satellite Meeting on SNPs and Haplotype Inference, November 21-22, 2002.

L. Shvartser, “Multiple Sequence Alignment Based on the Quasi-concave function Optimization Over a Lower Semi Lattics” in the working group meeting : Informatics of Protein Classification, December 15, 2000.

Graduate student publications

Abramsona, S., Gabriela Alexe, Peter L. Hammer, Doyle Knight, and Joachim Kohn, “Using Logical Analysis of Data (Lad) to Find Physio-Mechanical Data Patterns which Predict Cellular Outcomes,” Rutcor Research Report-40-2002 December, 2002.

Alexe, G., S. Alexe, P.L. Hammer, B. Vizvari. “Feature Selection for the Logical Analysis of Microarray Data,” Third international conference for the Critical Assessment of Microarray Data Analysis (CAMDA 2002), November 14-15, 2002, Durham, North Carolina.

Alexe, G., S. Alexe, D. Axelrod, E. Boros, and P.L. Hammer, “Combinatorial Analysis of Breast Cancer Data from Image Cytometry and Gene Expression Microarrays,” *Journal of Computational Biology*.

Alexe G, Alexe S, Liotta LA, Petricoin E, Reiss M, Hammer PL. “Ovarian cancer detection by logical analysis of proteomic data.” *Proteomics* 2004;4:766-783.

Alexe, G., S. Alexe, P. L. Hammer, D. Weissman. Logical analysis of leukemia microarray data (paper, in preparation).

S. Alexe, E. Blackstone, P.L. Hammer, H.Ishwaran, M.S. Lauer, and C.E. Pothier Snader, "Coronary Risk Prediction by Logical Analysis of Data," *DIMACS Technical Report* 2002-11.

S. Alexe, E. Blackstone, P.L. Hammer, H.Ishwaran, M.S. Lauer, and C.E. Pothier Snader, "Coronary Risk Prediction by Logical Analysis of Data," *Annals of Operations Research*, 2002

S. Alexe, "Pattern-Based Scoring Systems in Logical Analysis of Data" is in preparation.

E. Boros, K. Elbassioni, V. Gurvich, L. Khachiyan, and K. Makino, "Dual-bounded hypergraphs: A survey, in Proceedings of the SIAM Workshop on Discrete Mathematics and Data Mining (DM & DM)," Arlington, VA, April 2002, pp. 87--98, DIMACS Technical Report 2002-23.

E. Boros, K. Elbassioni, V. Gurvich and L. Khachiyan, "Matroid intersections, polymatroid inequalities, and related problems," to appear in *Proceedings of the 27th International Symposium on Mathematical Foundations of Computer Science (MFCS),* Warszawa--Otwock, Poland, August 26-30, 2002, 143-154.

E. Boros, K. Elbassioni, V. Gurvich, L. Khachiyan and K. Makino, "Dual-bounded generating problems: All minimal integer solutions for a monotone system of linear inequalities," to appear in *SIAM Journal on Computing*.

K. Elbassioni, "On dualization in products of forests," in *Proceedings of the 19th International Symposium on Theoretical Aspects of Computer Science (STACS),* (H. Alt and A. Ferreira, eds.),March 14-16, 2002, Antibes, Juan les pins, France, Lecture Notes in Computer Science 2285, pp. 142--153.

K. Elbassioni, "Incremental Algorithms for Enumerating Extremal Solutions of Monotone Systems of Submodular inequalities and Their Applications," Ph.D. Thesis, Department of Computer Science, Rutgers University, New Jersey, 2002.

K. Elbassioni, "An algorithm for dualization in products of lattices," DIMACS Technical Report 2002-26, RutgersUniversity ,

K. Elbassioni, "An algorithm for dualization in products of lattices," to appear in Proc. 10th Annual European Symposium on Algorithms (ESA 2002), 17-21 September, 2002.

J. Joo and Joel L. Lebowitz, "Pair approximation of the stochastic susceptible-infected-recovered-susceptible epidemic model on the hypercubic lattice," Phys. Rev. E, **70** (2004).

Talks given by graduate students

Gabriela Alexe, DIMACS Mixer II, Telcordia, October 22, 2002. Logical Analysis of Biomedical Data.

Gabriela Alexe, DIMACS workshop on Visualization and Data Mining, October 24-25, 2002. Visualizing Association Rules in the LAD Knowledge Space.

Gabriela Alexe, IEEE Computer Society Bioinformatics Conference, Stanford University, Palo Alto, CA, August 14 - 16, 2002. Logical Analysis of Data and
Applications to Bioinformatics.

Gabriela Alexe, CAMDA 2002, November 14-15, 2002, Durham, North Carolina. Feature Selection for the Logical Analysis of Microarray Data.

Gabriela Alexe, INFORMS Annual Meeting 2002, November 19-21, San Jose, CA. Logical Analysis of Proteomic Data.

S. Alexe, Datascope – a new tool for Logical Analysis of Data, DIMACS Mixer Series, September 19, 2002, DIMACS, Rutgers University

S. Alexe, P. L. Hammer, Logical Analysis of Data and Applications to Bioinformatics – Datascope software, IEEE Computer Society Bioinformatics Conference, Stanford University, Palo Alto, CA, August 14 - 16, 2002

Khaled M. Elbassioni, 7th International Symposium on Artificial
Intelligence and Mathematics, Fort Lauderdale, Florida, January 2-4, 2002. Learning monotone binary functions in products of lattices

Khaled M. Elbassioni Integer Programming Conference in honor of
Egon Balas, Carnegie Mellon University, Pittsburgh, PA, June 3-5, 2002. An inequality for polymatroid functions

Steven H. Kleinstein. "Estimating hypermutation rates during immune responses". *DIMACS/BIOMAPS Seminar Series in Quantitative Biology*. May 2004.

Yoram Louzoun and Steven H. Kleinstein. "Germinal Center Structure Evolution and Affinity Maturation". *DIMACS Workshop on Complexity in Biosystems: Innovative Approaches at the Interface of Experimental and Computational Modeling*. April 2002.

Steven H. Kleinstein and Jaswinder Pal Singh. "Quantitative models of germinal center dynamics". *International Conference on Mathematical and Theoretical Biology*. July 2001.

Erich R. Schmidt, Princeton Computer Science, International Conference on Mathematical and Theoretical Biology, Hilo, Hawai'i, July 16-19, 2001. Presented "Towards More Realistic Affinity Maturation Modeling".

Akshay Vashist, Rutgers Computer Science, DIMACS Mixer II, Telcordia, October 22, 2002 Validation of a supervised classification scheme by hidden structure analysis of training data and its application in bioinformatics for gene annotation

VII. Other Products

http://dimacs.rutgers.edu/SpecialYears/2000_2003/
Special Focus on Computational Molecular Biology main web site

http://dimacs.rutgers.edu/Workshops/Bioconsensus
Workshop and Working Group Meeting on Bioconsensus

http://dimacs.rutgers.edu/Workshops/GeneMyers
Distinguished Lecture: Gene Myers, Vice-President, Informatics Research, Celera Genomics: Whole Genome Assemblies of the Drosophila and Human Genomes

http://dimacs.rutgers.edu/Workshops/Informatics
Working Group Meeting: The Informatics of Protein Classification

http://dimacs.rutgers.edu/Workshops/Bioinformatics
Mini-Workshop: System Based Modeling in Bioinformatics

http://dimacs.rutgers.edu/Workshops/WholeGenome
Workshop: Whole Genome Comparison

http://dimacs.rutgers.edu/Workshops/ProteinStructure2
Workshop: Protein Structure and Structural Genomics: Prediction, Determination, Technology and Algorithms

http://dimacs.rutgers.edu/Workshops/TopologyIII
Workshop: DNA Sequence and Topology

http://dimacs.rutgers.edu/Workshops/Integration/
Workshop: Integration of Diverse Biological Data

http://dimacs.rutgers.edu/Workshops/MiningTutorial
Summer School Tutorial on New Frontiers in Data Mining

http://dimacs.rutgers.edu/Workshops/BioconII
Tutorial on Bioconsensus II

http://dimacs.rutgers.edu/Workshops/BioconII
Workshop on Bioconsensus II

http://dimacs.rutgers.edu/Workshops/GeneExpression/
Workshop: Analysis of Gene Expression Data

http://dimacs.rutgers.edu/Workshops/Complexity
Workshop: Complexity in Biosystems: Innovative Approaches at the Interface of Experimental and Computational Modeling

http://dimacs.rutgers.edu/Workshops/Chiaotung
DIMACS-CTS (National Chiao Tung University) Conference on the Interconnections Among Codes, Designs, Graphs and Molecular Biology

http://dimacs.rutgers.edu/Workshops/SNP
Workshop: Computational Methods for SNPs and Haplotype Inference

http://dimacs.rutgers.edu/Workshops/ProteinDomains
Workshop: Protein Domains: Identification, Classification and Evolution

http://dimacs.rutgers.edu/Workshops/Tree/
Working Group: Mathematical and Computational Aspects Related to the Study of The Tree of Life

http://dimacs.rutgers.edu/Workshops/Medicalapps/index.html
Workshop: Medical Applications in Computational Geometry

http://dimacs.rutgers.edu/Workshops/Biomaps/
BioMaPS/DIMACS Tutorial: Introduction to Modern Concepts in Biology for Mathematical and Physical Scientists

http://dimacs.rutgers.edu/Workshops/Techniques
DIMACS Workshop on Data Mining Techniques in Bioinformatics

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

http://dimacs.rutgers.edu/Workshops/Molecular/
Working Group: New Algorithms for Inferring Molecular Structure from Distance Restraints

http://dimacs.rutgers.edu/Workshops/Biogame/
Workshop: Interface Between Biology and Game Theory

http://dimacs.rutgers.edu/Workshops/Reticulated/
Workshop on Reticulated Evolution

http://dimacs.rutgers.edu/Workshops/Reticulated_WG/
Working Group on Reticulated Evolution

VIII. Contributions within Discipline

This special focus was of course 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 documenting this.

“Seems that it will be an easy case to make that the CMB series of special years has had huge impact on the field .... It has been wonderful to have (and organize) workshops, and I have learned many things and met many wonderful colleagues as a result. My students have attended several of the workshops, and have gotten a lot out of them. For example, Elena Zaslavsky attended Wilma Olsen's recent workshop/tutorial on transcription and she found the introductory lectures so useful that she has screened them for our group. Two of my students, Carl Kingsford and Jessica Fong, have also gotten awards from DIMACS. In Carl's case, it is in direct support for research, so eventually that will lead to a paper (on predicting protein-protein interactions).

Personally, the special years on CMB have been invaluable to me. My involvement as a postdoc was critical in making the complete transition into this field (and there are so many of us for which this is true). Plus, I got a broad introduction to computational biology (not just my specific area expertise) and I have found this useful in both research and teaching. Now, I appreciate the training my students are getting...” Mona Singh, Princeton

“My group is definitely delving more and more into problems relating to Complexity (the subject of the DIMACS conference.) We recently obtained grant funding in this area and are beginning collaborations with other people in this area who attended this meeting.” Don Ingber, M.D., Ph.D., Children's Hospital/Harvard Medical School. The grant "Design Principles of Complex Biological Networks" was funded by the Army Research Office.

“The workshops exposed me to various problems and did lead to new research ideas which I plan to pursue in the near future … I very much liked workshops organized under the special focus on

Computational molecular biology. I hope it would continue and be more geared toward problems in molecular evolution, phylogenetics etc.   I particularly liked the Transcriptional regulation "summer school" held in June 2004. There could be more such tutorials which would present overview of the core biological processes by experts so that people in computational and mathematical sciences can better understand and identify real problems. " Akshay Vashist, graduate student, Rutgers

"I have participated in a few DIMACS Computational Molecular Biology workshops and found them very useful. In particular, I met there Prof. John Widom (NWU). Collaboration with Dr. Widom resulted in a remarkable new development in my study of DNA bending. This work has top priority for me now and I am sure that very interesting results will be obtained there." Alex Vologodskii, Department Of Chemistry, New York University

"I participated in the two-week concentrated introductory course that was held June-July 2003. Its timing was especially fortunate for me, as I was in the midst of transition to research in Molecular Systems Biology. The course was well organized and informative, and it allowed me to acquire an important knowledge base for the succeeding period. Shortly afterwards, I began working at a Systems Biology company called Beyond Genomics. Although none of the projects or collaborations that I am involved in, stem directly from the course, it did contribute to my abilities and indirectly to their fruits. These include the presentation of posters about data analysis methods (at the IEEE sponsered Workshop on Genomic Signal Processing and Statistics, 2004, and the ISMB/ECCB 2004.) In addition, I recently gave an invited seminar (System Biology in Practice: Science, Technology, Challenges**)** at the Institute for Systems Research, the University of Maryland." Amir Handzel, Senior Scientist, Beyond Genomics, Inc.

"I am a graduate student in computer science at Princeton University. My advisor is Mona Singh and my thesis work is in computational biology, specifically protein structure prediction and protein-protein interaction prediction. The upcoming year will be my 5th (and final) year.

The workshops that DIMACS has hosted in connection with the Special Focus on Computational Molecular Biology have been a terrific addition to my training in the field. These workshops are a great way for graduate students new to the field to add breadth to their experience at low cost.

While not directly connected with the Special Focus on Computational Molecular Biology, I did present some of my thesis work at the DIMACS Workshop on Geometric Optimization. DIMACS workshops provide a much needed venue for new researchers to present current work.

I also applied for and was awarded a DIMACS Summer Support Award ($1,000) for this (2004) summer. This award was supported through the Special Focus on Computational Molecular Biology. It has supported me for part of a project to predict protein-protein interaction using evolutionary information. Under the assumption that interacting proteins experience similar pressures over time and evolve in a correlated way, we can use similar evolutionary history as an indicator that two proteins interact. The goal of this work is to find an effective way to compare the inferred evolutionary histories of proteins. While this project is ongoing, our method has shown some promising results, beating similar, widely-used methods on a few test sets." Carl Kingsford, graduate student, Princeton

"I would say that attending the DNA-Based Computer meetings II and III certainly helped to stimulate my thinking about the technological uses of DNA. That thinking, in turn, led to the new research idea, DNA-based steganography." Carter Bancroft, Mount Sinai School of Medicine

"I participated in a DIMACS conference at Rutgers called "Workshop on Whole Genome Comparison in 2001." I was a graduate student in evolutionary biology at the time and it opened up a whole new world for me. The conference, through seeing talks and speaking to leaders in the field, catapulted me into bioinformatics and stimulated my desire to learn more. I subsequently finished my Ph.D. in a biology department with no wet-lab component to my thesis.

More directly, I used an algorithm and associated software that I saw at the conference for my own research that resulted in my first bioinformatics-related publication.

Overall, it was a great experience and I don't think that I've been to a conference/workshop in this field as good as the DIMACS one since. Thank you!" Cristian I. Castillo-Davis, Department of Statistics, Harvard University

"I have only been back to DIMACS … for a meeting on SNPs and Haplotypes. That meeting did have a large effect on my subsequent work. It truly brought together many biologists along with CS and math types, and was the first meeting to get everyone involved in that topic talking to each other. I think it was so successful because it was really the first meeting on that topic in either community and so almost all the world leaders in that area were there. The problems I worked on subsequently were greatly influenced by that meeting, and contacts made there were also very helpful." Dan Gusfield, University of California, Davis

"I definitely have been affected by participation in this meeting, and others like it over the past few years. I just recruited a new graduate student and some more senior staff in this area. I also am trying to organize a complex systems initiative at Harvard University." Don Ingber, M.D., Ph.D..Children's Hospital/Harvard Medical School

"Thank you for inviting me to submit my comments about your Special Focus in Computational Biology. I do have a few comments to make, since I believe it was of much influence to my professional career.

Over the last two summers, I have been funded through Dimacs as a graduate student, which has given me a chance to work without the need to teach. As a concrete result of this, I should mention the tech report 2004-29, "A Note on a Monotone Small Gain Theorem, with Applications to Delay Systems", which I prepared together with my advisor, as well as the paper "On the stability of a model of testosterone dynamics", which will appear on the J. Math. Biology.

Throughout the course of my graduate program, I have been able to get first-hand information on biological research problems through conferences at Dimacs, even before I had decided on mathematical biology as my dissertation topic. This has given me substantial interest in a subject I had not considered before. Herein lies the long term influence of your special focus on my career, and I can confidently say, I would have probably worked on something else otherwise." German Andres Enciso, graduate student, Rutgers

"I participated in the workshop on Protein Domains: Identification, Classification and Evolution (February 2003). I think it was one of the best workshops I participated in. The organizers did an excellent job overall, and especially in the choice of the speakers. I had interesting discussions with colleagues who are working on similar problems, who helped answering some of the questions I had regarding specific aspects of the data." Golan Yona, Cornell

"The list of DIMACS workshops gave me perspective that I found useful in understanding the expanding scope of how mathematics can be useful to solve problems in modern biology … I enthusiastically support what DIMACS has done and I hope will continue to do. If more funds become available, I want to send students to DIMACS workshops. It's very well organized and one learns a lot in a short time." Harvey J. Greenberg, Professor of Mathematics, Adjunct Professor of Computer Science & Engineering

(UCD), Adjunct Professor of Preventive Medicine & Biometrics (Medical School), Founding Director, CU Center for Computational Biology, University of Colorado at Denver

“My participation in the Focus was only as an attendee at a two day meeting at the end of 2003. However, this was quite important to me in helping me start a new research area in my group, bringing me up to speed with the state of the art, and making individual connections. I have no publications yet, but have a graduate student working in the area.

The meeting was on statistical genetics, and the subject area is population and statistical genetics. For the layman, this concerns how to analyze genetic data collected on a number of individuals by examining positions in the genome sequence that are known to be variable in general between individuals (around 1 in a thousand positions are of this type - the other 999/1000 are the same across all humans). Because we are all related by common ancestry if you go back far enough, there is a lot of complex structure to the correlations between these variable positions (polymorphisms), and this can be used to help identify which differences cause disease, and also whether some differences have been beneficial to (part of) the population in the past. This is a field where statistical and computational approaches come together, because the "full correct" model for how to handle it statistically is incomputable in practice.” Richard Durbin, Wellcome Trust Sanger Institute

“I was a speaker at a DIMACS Reconnect meeting, the DIMACS Workshop on Protein Structure and Structural Genomics in 2001, and the DIMACS/RECOMB Satellite Meeting on SNPs and Haplotype Inference in 2002. In all cases, I found them a great opportunity to meet with communities of interdisciplinary scientists who do not generally attend the same meetings. The SNP and Haplotype Meeting in particular was influential on me because of its unusual success at getting computer scientists, statisticians, and geneticists together in the same room to discuss their different approaches to a common set of problems. The meeting was so successful that a second iteration was planned at Carnegie Mellon University, which I was partially responsible for organizing, and a third is now planned for the University of Southern California next year. **The original DIMACS meeting appears to have laid the groundwork for one of the most interesting annual meetings in interdisciplinary genetics**.” Russell Schwartz, Assistant Professor of Biological Sciences, Carnegie Mellon University

“Some of my research focuses on medical data mining. Attending workshops such as these always gets one to think along new lines…focus groups such as these with top presenters always has an influence. One always brings back some nugget of useful information after attending.” Susan Imberman, Assistant Professor, Department Computer Science, College of Staten Island

“The problem I started to work on at DIMACS was new to me. In general, I found DIMACS to be a great place to work and the DIMACS ‘formula’ to be very prone to fruitful collaborations.” Gregory Kucherov, INRIA

DIMACS programs spawn new collaborations, often between people in different research areas or even different disciplines. Some of these new collaborations resulted in research discussed in the section on Project Findings. Here are a few selected examples of collaborations that are in their early stages.

The workshop on Protein Domains: Identification, Classification and Evolution (February 2003) led to a collaboration between Golan Yona, Cornell, and Sarah Teichmann. Sarah is studying the repertoire of domain architectures in different organisms using data that Golan Yona generated with her domain prediction algorithm.

Lee Segel, Computer Science/Applied Mathematics, Weizmann Institute of Science, Israel, and Tim Buchman, Surgery and Critical Care Medicine, Washington University, have been collaborating on the multi-component interaction implied by what is called Multiple Organ Dysfunction Syndrome (MODS). Particularly in critical care facilities, MODS often results in successive organ collapse and death. In thinking about MODS, questions arise of the type that Lee has addressed in thinking about the immune system and other complex biological systems. In particular, how does the body, and how might the physician, handle situations where several physiological systems (e.g. parts of the immune system, or heart, lung, kidney etc) each have independent jobs to do, but where the parts interact. Tim was a leading participant in a meeting Lee ran at the Santa Fe Inst, on Math Models in Medicine.

DIMACS visitor Gregory Kucherov, with the host of his stay, Martin Farach-Colton, started joint work on an algorithmic problem motivated by a computational biology application to processing protein sequence data. After studying the related literature, they proposed an algorithmic formalization of the problem. The work is currently in progress.

As a direct result of Bill Day's and Fred McMorris' participation in the Bioconsensus workshop of Oct 2, 2001 and the tutorial they ran immediately before this workshop, Bill Day and Fred McMorris were motivated to begin work on the research monograph "Axiomatic Consensus Theory in Group Choice and Biomathematics" published in the Frontiers in Applied Mathematics Series of SIAM (Nov 2003).

IX. Contributions -- other Disciplines

The entire project involves a mix of disciplines. Contributions to other disciplines are covered in the section on Contributions within Discipline.

X. Contributions -- Human Resource Development

Many of the comments in the section on Contributions within Discipline illustrate the human resource development contributions of this project. There were twenty-one graduate students who worked on small research projects. This is described in detail in the section on Project Training/Development. Below are some comments on the effect of the special focus on some of these students. Many other students attended special focus workshops and tutorials.

"Thanks to your generous funding, I was able to attend the International Conference on Mathematical and Theoretical Biology which took place July 16-19, 2001. The research presented at this conference was very closely related to my own work modeling the immune system. In addition to the sessions on immunology, I attended interesting sessions on cancer and evolution. While at the meeting, I was able to make a number of new contacts. Among them was Dr. Tim Manser, an experimentalist who works on exactly the system that I am trying to model (the germinal center). Talking with him was very informative and he agreed to send me some of his data. I also had a great experience presenting my research in the session on "Disease and Immunology". Although the talk was only 15 minutes, I feel that I was able to get my point across clearly. Afterwards, I had discussions with many people about my work during which a number of interesting points were raised. For example, the model I presented did not include explicit spatial effects. One scientist thought that such effects might have a great impact on the results. This was an important discussion since I am currently working with others to develop a spatially explicit model. All in all, I feel that I learned a lot from attending this conference. Most importantly, I was able to make many important contacts." Steven H. Kleinstein, Princeton, graduate student

"I am now back to Vanderbilt University, working on Bioinformatics at Genetic Medicine Division as an assistant professor. I have been working on and am still working on my proposal which was granted a DIMACS award. I am planning on using the results to apply for a grant from NIH." Zhongjun luo, Genetic Medicine Division, Vanderbilt University Medical Center

“The results obtained during the DIMACS program, by applying the Logical Analysis of Data to biomedical datasets are very promising, and they will be included in my Ph.D. thesis. I am very grateful to DIMACS for the summer award.” Gabriele Alexe, Rutgers, graduate student

XI. Contributions to Resources for Research and Education

XII. Contributions Beyond Science and Engineering