D I M A C S

Center for Discrete Mathematics and Theoretical Computer Science

MATHEMATICAL SUPPORT FOR MOLECULAR BIOLOGY: ANNUAL REPORT - January 1997

DIMACS Center • Rutgers University • P.O. Box 1179 • Piscataway, NJ 08855-1179 • USA
TEL: 908-445-5928 • FAX: 908-445-5932 • EMAIL: center@dimacs.rutgers.edu
WWW: http://dimacs.rutgers.edu/ • TELNET: telnet info.rutgers.edu 90

DIMACS is a partnership of Rutgers University, Princeton University, AT&T Labs, Bellcore, and Bell Labs

Contents

1 Overview

The DIMACS "Special Year" in Mathematical Support for Molecular Biology began in August 1994 as an activity intended to last 25 months. With some $377K of funds from its NSF science and technology center grant, significant additional external funding of some $525K from numerous sources, plus $75K of additional funds from Rutgers, the center was able to plan an extensive series of tutorials and workshops, to support 4 postdoctoral fellows for two year appointments, host several long term visitors, and host a distinguished lecturer series.

The program has been extended into a third year of activity, aided by an additional grant of $100K from the National Center for Human Genome Research (through NSF) and an additional $118K from DIMACS' science and technology center grant. The current year has two postdoctoral fellows, several workshops, and an active seminar series.

A list of special year organizers, the distinguished membership of the special year steering committee, and a description of special year workshops is contained in Section 6.

Theme of the Project

Discrete structures have been used to formulate some of the most fundamental concepts of molecular biology. It is now well-known that information is stored within a cell in long nucleic molecules which can be thought of as long strings of smaller units called nucleotides. Some of the most important problems of modern molecular biology involve combinatorial and algorithmic questions about these strings. Models of proteins, DNA, and RNA molecules as sequences form a basic tool of molecular biology.

Throughout its history, molecular biology has interacted with methods of discrete mathematics (DM) and theoretical computer science (TCS). The first nucleic acid sequence was determined in 1965 by R.W. Holley and his co-workers at Cornell University (it contained 77 bases). The method used was the fragmentation stratagem, which involved the combinatorial problem of reconstructing an unknown sequence from fragments obtained by certain enzyme decompositions. Graph-theoretical methods, using eulerian chains, were developed by Hutchinson to give algorithms for implementing the fragmentation stratagem. Of course, these methods are not used any more, and indeed were used only for a short time before other, more efficient methods were adopted. However, they played an important role in the development of molecular biology. In this special year, we are investigating similar uses of DM/TCS in molecular biology.

Goals of the Project

The special year has a number of specific goals: To expose a large community of discrete mathematicians/theoretical computer scientists to problems of molecular biology that are fundamentally DM/TCS problems; provide an opportunity for intensive collaboration to discrete mathematicians/theoretical computer scientists who have already discovered these problems; introduce some outstanding young people to the field in a very concentrated way; forge lasting partnerships between mathematical scientists and biological scientists; involve biological scientists in a primary role in defining the agenda of computational biology; and introduce molecular biologists to a community of mathematical scientists who are interested in helping them with their problems.

The success of our special year shows that we can connect different communities, that we can find ways to speed the transition from research in discrete mathematics and theoretical computer science to biological usefulness, and that we can influence the research of scientists in discrete math/theoretical computer science, including some very talented young people, by exposing them to real problems from another discipline.

Perhaps a recent unsolicited comment from biologist George Estabrook of the University of Michigan, following his participation in the DIMACS workshop on Mathematical Hierarchy and Biology, best sums up our impact. He wrote: "My participation was very valuable to me, indeed, it was the biggest intellectual 'kick in the pants' I've had in the past decade. I hope I was able to help some of the other particpants see some things in new ways too. I hope DIMACS Center will be able to continue to present small, intense, interactive, interdisciplinary workshops like this in the future."

2 Collaborations between Communities

In many ways, this special year was not that different from DIMACS' earlier special years. It involved a workshop program, a visitor program, and a postdoc program. However, in one way it was very different. Crucial goals of the special year were real outreach from the mathematical sciences community to the biological sciences community and the establishment of serious, lasting partnerships among mathematical and biological scientists. This is in line with one central goal of science and technology centers (STC's) such as DIMACS, the goal of creating "linkages that involve significant intellectual exchange with other appropriate communities and institutions beyond the sponsors."

Partnerships with Other Institutions

A key to the success of the year has been the establishment of partnerships with other institutions. Here is a summary of some of the new partnerships established through the special year. First is with *the Waksman Institute for Molecular Genetics at Rutgers*, whose director Joachim Messing is a chair of the year. Many Waksman faculty have been involved in mentoring, tutoring, organizing, many events have been held at Waksman, and Waksman has made available its Molecular Biology Computing Laboratory. A second partnership has been with *the Center for Molecular Biotechnology*, our sister STC at the University of Washington. Leroy Hood, Director of the center, is on the steering committee for the year and played a central role in convincing us that the year was a good idea. We have exchanged scientific visits with that center, and in particular there have been visits to DIMACS by Hood, Maynard Olson, Phil Green, and their postdocs. A third partnership has been with *the Genome Center and the Department of Computer Science at the University of Pennsylvania.* One mini-workshop was held through the auspices of the Genome Center, our seminar series has been held at Penn occasionally, a second seminar series (on phylogeny) was initiated there and held in collaboration with DIMACS, DIMACS and Penn collaborated in a conference in the Spring of 1996, a workshop is planned at Penn in the Spring of 1997, and Sampath Kannan, David Searls, and Tandy Warnow from Penn are on the special year steering committee, Searls representing the Genome Center. Computational biology will be a central piece of the future joint activities between Penn and DIMACS, which are starting with a joint postdoc program. *The Center for Theoretical and Applied Genetics at Rutgers* has been an active partner in the year as well. Its Director, Robert Vrijenhoek, and its Associate Director, Peter Smouse, are involved in our seminar series, have served on workshop committees, and are participating in joint research. Postdocs from CTAG are involved and faculty from there have posed problems for mathematical scientists. We have formed active and growing partnerships with *the Departments of Molecular Biology and of Ecology and Evolutionary Biology at Princeton.* Scientists there have been busy formulating problems for mathematicians and computer scientists, we visited those departments as part of our tutorial program, and those visits have led to major collaborations (details below). *The Rutgers Center for Molecular Biophysics and Biophysical Chemistry* is a partner in several activities. Its director, Wilma Olson, is on the steering committee for the special year, has also mentored for several years in our summer reu program, a research experiences program for undergraduates, and is an organizer of the upcoming workshop DNA Topology II, which has co-sponsorship of her center. Other faculty from the center have expressed an interest in getting involved in our reu program as well. In addition, Helen Berman from the center is going to play a role in the upcoming DIMACS special year on massive data sets, and has interacted with us on nucleic acid database problems. *The Program in Mathematics and Molecular Biology*, based at UC - Berkeley, is also a partner in the DNA Topology II workshop. A growing partnership between DIMACS and the *Mt. Sinai School of Medicine* led to a joint program during the summer of 1996, featuring collaborations in DIMACS' summer reu program and in DIMACS Young Scholars Program for talented high school students, and is leading to plans for a joint postdoc program.

Biotech and Pharmaceutical Companies

Biotech and pharmaceutical companies have also been involved heavily in the special year. Participating companies include CuraGen, Seq Limited, Roche Molecular Systems, Hoffman LaRoche, Pharmagenics, SmithKline Beecham, Johnson and Johnson, Biosym Tech., Ciba Geigy, Merck, Perkin-Elmer, Genome Therapeutics, and The Institute for Genomic Research. These companies have provided financial support, speakers, committee members, organizers, attendees, and data and feedback for our algorithm implementation challenge. To give just one example of industrial participation: In our miniworkshop on mathematical problems in drug discovery, held in September 1995, the majority of speakers and participants were from industry. There was financial sponsorship from Biosym Tech., Ciba Geigy, Merck Research, and Smith Kline Beecham. This miniworkshop was designed to acquaint pharmaceutical and computational chemists with the latest developments in such fields as discrete and computational geometry and to introduce discrete mathematicians and theoretical computer scientists to the problems of the field coming to be known as "combinatorial chemistry," a field that so far consists largely of heuristic algorithms based upon chemical intuition. The workshop led to lots of discussion and to some promising beginning contacts, for example one between a Chiron Corp. chemist and DIMACS mathematician who have common interests and ideas about the use of large-scale cluster analysis in drug design.

Biological Leadership

DIMAS has among its 150 permanent members at Rutgers and Princeton Universities, AT&T Labs, Bell Labs, and Bellcore many of the world's leading mathematicians and computer scientists. They have been joined in this activity by prominent biologists who have been playing a leadership role in all special year activities. On the steering committee for the year are Joshua Lederberg, Leroy Hood, Charles Cantor, Michael Bulmer, and other biological scientists. Each workshop has had two principal organizers, a mathematical scientist and a biological scientist. Each workshop organizing committee has had additional biologists.

Algorithm Implementation Challenge

The biological/biotech participation in our algorithm implementation challenge was noteworthy. The challenge, fourth in the line of such challenges, was devoted to "shotgun sequencing" and "genome rearrangements." Participants were challenged to test their algorithms against both random and real data sets. A workshop describing the challenge participants' results was held in September 1995. Ellson Chen, a "cutting edge biological sequencer," defined real data sets and provided intermediate biological feedback. Participants were highly motivated to continue coordinated work on testing their algorithms on common benchmark data. The addition of further, large, and difficult data sets was agreed upon. A major result was that scientists from industry not only participated actively, but helped with the benchmark data sets by giving away their raw data at a very early stage – something rather unusual. We received data from Applied Biosystems, a Divison of Perkin-Elmer, Genome Therapeutics and The Institute for Genomic Research. These are among the leading genome sequencing companies. Since participants from both academia and industry tested their programs in the same way, there was a lively discussion. It has become apparent that current experimental techniques lead to a need for extremely powerful software and this observation is now strongly influencing ongoing research.

Biological Visitors

In addition to visitors from the mathematical sciences, many biological sicentists have been special year visitors. Prominent biological visitors include Leroy Hood, Herbert Hauptman, Charles Cantor, Temple Smith, Gary Stormo, and Maynard Olson. We have also gotten inquiries and requests to visit from less prominent biologists around the world. Many biological scientists have been speakers in our weekly seminar series.

All indications are that this has been a remarkably successful outreach effort. We feel that it also underlines our achievement of a major STC goal: Demonstrating the value of the center mode of funding over individual research grants: This kind of interdisciplinary program could not be accomplished through individual research grants.

3 Research Collaborations between Mathematical and Biological Scientists

We have had many significant research accomplishments during the special year. Perhaps most important are the results of or beginnings of significant interdisciplinary collaborations.

Computer Modeling of Ultraselfish Genes in Mice. Lee Silver, a Princeton biologist, and Dannie Durand, a DIMACS member from Bellcore who became a special year visitor, have been working on the t-haplotype regions on the chromosomes of mice. Silver and Durand met through the DIMACS tutorial. They began modelling work on the effect of environmental and hereditary factors on the t-haplotype region, which has unusual sperm killing properties. Working with Silver, Durand developed a Monte Carlo simulator; the results of their simulations upset an explanation of t-haplotype persistence which had been widely accepted for thirty years and have suggested new hypotheses to be tested through laboratory and field work. (See details in Section 5.)

Prediction of Functional Domains on a Herpesvirus Gene. Lynn Enquist, a Princeton biologist, DIMACS member Dannie Durand from Bellcore, and DIMACS postdocs Vineet Bafna and R. Ravi, collaborated on the prediction of functional domains on a herpesvirus gene. Enquist sought help with a sequence comparison problem involving the herpesvirus gene. He hoped to find similar sequences in other species and that that would help him identify the function of the virus proteins. But searches through the GenBank data base were not very revealing. The interaction led to negative results: Bafna, Durand, and Ravi did not find new comparison algorithms for better searches nor did they discover that the methods used in the searches were faulty. However, they did learn a lot about interaction with biologists. Moreover, the negative results have pointed us and the biologists in the direction of seeking similarities in the 3-dimensional structures of the proteins rather than in their amino acid sequences. It has pointed us in the direction of new research questions about protein structure. As a side note: This collaboration led to Ravi supervising an undergraduate senior thesis at Princeton.

Diabetes. Fredda Ginsberg-Fellner, a physician at Mt. Sinai in New York City, Larry Shepp, DIMACS member from AT&T Labs, Martin Farach, DIMACS member from Rutgers, and Mick Noordewier, DIMACS member from Rutgers Computer Science and Waksman, have worked on diabetes and auto-immune diseases. Shepp had the idea that the existence of auto-immune diseases itself might provide the long-sought window into understanding how DNA and function are related. It is now possible to perform an experiment to test the hypothesis that there is a similarity between the triggering virus and some part of the DNA in those individuals susceptible to these diseases. It is hoped that by sequencing the DNA of the viruses and the presumed relevant parts of the human genome, a match can be found at the DNA level rather than at the protein level. Shepp got Ginsberg, Farach, and Noordewier interested. Ginsberg presented a lecture on this topic at DIMACS. This work had progressed to the point where it was planned that Ginsberg would apply for a grant from the Juvenile Diabetes Foundation to support work on acquiring the sequence data and Shepp, Farach, and Noordewier would search for the matches and design further experiments. However, Ginsberg's illness has put a damper on these plans.

Phylogenetic Tree Reconstruction. David Swofford, a Smithsonian biologist, and DIMACS permanent member Martin Farach from Rutgers have collaborated on phylogenetic tree reconstruction. Farach had developed a theoretical algorithm for tree reconstruction, he and Swofford met at our workshop on Phylogeny, and Swofford returned home and coded up the algorithm for practical implementation ("a theoretician's dream").

Phylogenetic Trees and Minimum Spanning Trees. Peter Smouse, Rutgers biologist, Richa Agarwala, DIMACS postdoc, and Martin Farach have collaborated on phylogenetic trees and minimum spanning tree problems. When analyzing data, Smouse came up with the problem: Given a distance matrix M, find an edge-weighted spanning tree in which the distance $D(i,j)$ in the tree between i and j is such that $\sum[D(i,j) - M(i,j)]$ is minimized. Not surprisingly, they showed that the problem is NP-complete, and they produced lower bounds for it.

DNA Computing. Princeton biologist Laura Landweber and DIMACS permanent member Dick Lipton from Princeton, who met through the special year, have started a serious collaboration on DNA computing. They have submitted a joint research proposal and are designing experiments to test out some of the theory in Landweber's wet lab. This collaboration will probably be incorporated in a future DIMACS special focus on DNA Computing. (See details in Section 5.)

Models of Duplication. Several visits to DIMACS by Martin Vingron led to discussions between him and long-term DIMACS visitor Boris Mirkin on the subject of models of duplication in molecular evolutionary biology. These discussions led to a two month visit by mathematician Mirkin to Vingron's lab at the German National Cancer Institute in 1996, a DIMACS visit by biologist Robert Page, and a six month visit by Mirkin to Vingron's lab in 1996-97. (See details in Section 5.)

4 Postdoc Program

The special year featured an intense training program for postdocs. This started with a three week tutorial during the summer of 1994, open to many members of the mathematical sciences community, and providing an introduction both to fundamental topics in molecular biology and to the field of computational biology. Our first group of postdocs was appointed for two years. They all developed outstanding collaborations with biological scientists during their stay at DIMACS (some of these were described in Section 3 and others are described in Section 5.) All went on to good positions, two in positions outside the mathematical sciences. Our first group of postdocs and their current positions are:

- Richa Agarwala, 1994-96, currently Systems Analyst, R.O.W. Sciences, Inc. (on a full-time subcontract to NIH)
- Vineet Bafna, 1994-96, currently Senior Computational Biologist, SmithKline Beecham Pharmaceuticals
- S. Muthukrishnan, 1994-96, currently member of technical staff, AT&T Labs
- R. Ravi, 1994-95, currently Assistant Professor, Carnegie-Mellon University

Our current postdocs are:

- Mona Singh, 1995-97
- Scott Decatur, 1996-97

Although technically not a postdoc, DIMACS permanent member Dannie Durand from Bellcore participated in many parts of the postdoc program and in the process made a major career change, deciding to specialize in computational biology. Starting in the year 1995-96, she became a joint visitor at DIMACS and the University of Pennsylvania, where she is pursuing her interest and training in computational biology. Durand's collaboration with biologists is described elsewhere in this report.

5 Selected Research Results

The Special Year led to many significant research accomplishments. We give selected examples in this section.

Computer Modeling of Ultraselfish Genes in Mice. The *t-haplotype* is a mutant chromosomal region in mice that is passed on to progeny more often than the "normal" version of this region. This "ultraselfish" behavior is quite surprising because under Mendelian genetics, offspring inherit paired genetic segments with equal probability. Males carrying two t-haplotypes are sterile and most forms of the t-haplotype also carry a recessive lethal mutation. The evolution and persistence in nature of such a deleterious gene is therefore equally surprising and a number of explanations have been proposed in the literature. Since the complex social structure of mice makes it difficult to develop mathematical models of the population genetics of the t-haplotype, Dannie Durand, DIMACS permanent member from Bellcore and later a special year visitor, developed a Monte Carlo simulator to test these hypotheses in collaboration with Lee Silver of the Department of Molecular Biology at Princeton. The results of these simulations upset an explanation of t-haplotype persistence which had been widely accepted for thirty years and have suggested new hypotheses to be tested through laboratory and field work. This collaboration is continuing.

DNA Computing. Laura Landweber, a Princeton biologist, and Dick Lipton, a DIMACS permanent member from Princeton, who met through the special year, have started a serious collaboration on DNA computing. The collaboration has led to the idea that DNA computation can solve certain problems fundamental to DNA itself – an idea also espoused by others such as Pevzner. These problems include, for example, the analysis of DNA sequence composition, DNA diagnostics, and mutation detection. All use a simple encoding principle in which DNA molecules anneal to their target sequences, with each DNA probe encoding a specific tagged sequence. The idea is to use molecular selection with DNA probes to interpret existing DNA sequences. This is similar to how RNA enzymes and functional RNA molecules act in cells on target RNA substrates, through use of standard base-pairing. Up until now, the field of DNA computers has focused on only a small number of "classical" problems or calculations which can be more easily performed on a conventional computer or on paper. By expanding the role of DNA computers to solve molecular biological problems, Landweber and Lipton hope to provide a bridge for this field into the main stream of technology. More importantly, they expect that their initial ideas will define a class of problems for which DNA computers are ideally suited, and even better suited than a digital computer: Problems inherently involving DNA. This work has progressed to the stage where Landweber and Lipton are ready to test it by experiment in Landweber's wet lab.

Models of Duplication. For the last 20 years, an important idea in molecular evolutionary biology has been that at least some of the differences in patterns of divergence in single-gene-family based phylogenies must be explained by duplications of the genes involved (paralogous evolution). In this framework, a likely evolutionary species' phylogeny is supposed to be most parsimonious, that is, involving a minimum number of the duplications and corresponding loss-of-a-copy events. (This is very much different from the parsimonious trees currently of wide interest: The latter minimize the number of between-symbol differences in the molecular sequences.) The duplication idea has previously had no explicit expression except for some very technical constructions based on mapping the nodes (leaf clusters) of one tree into their least common ancestor (lca) nodes in another tree, constructions having no explicit biological meaning. An explicit graph-theoretic model of the gene duplication in a species tree (involving duplicate-copy labeling of the descent species) was suggested by Boris Mirkin, a long-term DIMACS visitor. His model led to an explicit definition of corresponding loss events which allowed him to develop a set theoretic framework for defining a collection of between-tree difference measures as counts of various duplication-based events needed for explanation of the difference between two trees. In the framework of this model, the technical suggestions about counting duplication based events made in the earlier literature were partly proved correct and partly disproved. A conjecture of Mirkin's about the equivalence of a model-based measure of difference with a combinatorial one (counting the total number of the intermediate nodes between the lca images of all child-parent pairs in the gene tree) was proved correct independently by special year visitor Martin Vingron and Oliver Eulenstein and by L. Zhang. This also led to an invitation to Mirkin to visit the Department of Theoretical Informatics of the National Cancer Center of Germany in Heidelberg, where he continued this work in collaboration with Vingron. Mirkin and Vingron are investigating the relationship between Mirkin's model and yet another cladistic-minded approach being developed in Glasgow, UK by biologist Robert Page, a special year visitor, and his associates, with an aim toward a larger area of application (host-parasite co-evolution, biogeographical evolution). The conjecture is that there is an equivalence between Mirkin's model and that approach. This conjecture was investigated during a recent DIMACS visit by Page. Mirkin and Vingron are also developing a model based algorithm for combining different gene-trees into a most parsimonious species tree, one needing the least number of the duplication-based events explaining all differences.

Approximating Numerical Taxonomy. A common problem in phylogeny is to fit an $n \times n$ distance matrix D by a tree metric T. DIMACS Postdocs Richa Agarwala, Vineet Bafna, and Babu Narayanan, DIMACS permanent member Martin Farach from Rutgers, special year visitor Mikkel Thorup, and collaborator Mike Paterson presented an $O(n^2)$ algorithm for finding an additive tree T such that the difference between the matrix of tree distances and the distance matrix D is at most three times the optimal. This is the first algorithm for the problem with a performance guarantee. They also showed that it is NP-hard to find such a tree T with the difference at most $9/8$ times the optimal. Later, Farach and Sampath Kannan, also a special year visitor, used this work as a building block in a clever algorithm for computing maximum likelihood trees, which is likely to have a great impact on the field. Experiments undertaken by Rutgers graduate student Jaime Cohen and Farach have shown that these distance methods construct better trees in practice than do the most popular heuristics used by biologists.

Phylogenetic Tree Reconstruction Under Stochastic Models of Nucleotide Substitution. Biologists seek to reconstruct evolutionary trees from aligned genetic sequences for an increasing number n of species. How fast must the sequence length N (a measure of the cost of getting input data) grow, as a function of n, in order to accurately recover the underlying tree with probability $1 - \epsilon$, if the sequences evolve according to simple stochastic models of nucleotide substitution? This was the problem considered by special year visitors Mike Steel, Laszlo Szekely, and Peter Erdos. They showed that for a certain model, a reconstruction method exists for which the sequence length N can grow surprisingly slowly with n (sublinearly for a wide range of parameters), and even as a power of $\log n$ in a narrow range (which roughly meets the lower bound from information theory). In contrast, a more traditional technique (maximum compatibility) provably requires N to grow faster than linearly in n. Their results were based on a new, and computationally efficient approach, for reconstructing phylogenetic trees from aligned DNA sequences.

Phylogenetic Tree Reconstruction from Quartet Splits. A significant problem in phylogeny is to reconstruct a slb tree from few valid quartet splits of it. It is well known that every slb tree is determined by its set of all valid quartet splits. Special year visitors Peter Erdos, Mike Steel, Laszlo Szekely, and Tandy Warnow strengthened this result by showing that its local (i.e., small diameter) quartet splits infer by a dyadic inference rule all valid quartet splits, and hence determine the tree. They also presented a polynomial time algorithm to recover the tree.

Sequencing by Hybridization. Sequencing by hybridization (SBH) is a promising alternative to the classical DNA sequencing approaches. However, the resolving power of SBH is rather low; with 64 Kb sequencing chips, unknown DNA fragments only as long as 200 bp can be reconstructed in a single SBH experiment. To improve the resolving power of SBH, Broude and others have recently suggested positional SBH (PSBH), allowing one to measure (with additional experimental work) approximate positions of every l-tuple in a target DNA fragment. Special year visitors Sridhar Hannenhalli, Pavel Pevzner, and Steve Skiena, together with Herbert Lewis, have studied the positional eulerian path problem motivated by PSBH. In this problem, the input is an eulerian graph in which every edge has an associated range of integers and the goal is to find an eulerian path such that the range of the ith edge in the path contains i. They showed that the positional eulerian path problem is NP-complete even when the maximum outdegree of any vertex in the graph is 2. However, they found polynomial algorithms to solve a special case of PSBH where the range of the allowed positions for any edge is bounded by a constant.

Computing the Similarity between RNA Strings. RNA strings are strings over a four-letter alphabet A, C, G, U with a secondary structure of base-pairing between $A - U$ and $C - G$ pairs in the string. Edges are drawn between two bases that are paired in the secondary structure and these edges have traditionally been assumed to be noncrossing. The noncrossing base-pairing naturally leads to a tree-like representation of the secondary structure of RNA strings. DIMACS postdocs Vineet Bafna, S. Muthukrishnan, and R. Ravi have addressed several notions of similarity between two RNA strings that take into account both the primary sequence and secondary base-pairing structure of the strings. They have found efficient algorithms for exact matching and approximate matching between two RNA strings. They defined a notion of alignment between two RNA strings and devised algorithms based on dynamic programming. They discovered a method for optimally aligning a given RNA sequence with unknown secondary structure to one with known sequence and structure, thus attacking the structure prediction problem in the case when the structure of a closely related sequence is known.

6 Special Year Programs

Special Year Organizers:

- Fred Roberts (Chair), Rutgers
- Joachim Messing(Chair), Waksman Institute - Rutgers
- Lawrence Shepp (Co-Chair), AT&T Bell Labs
- Michael Waterman (Co-Chair), University of Southern California
- Martin Farach (Assistant Chair), Rutgers
- Sampath Kannan (Publicity Chair), University of Pennsylvania

Special Year Steering Committee: David Axelrod, Rutgers; Michael Bulmer, Rutgers; Charles Cantor, Boston University; Daniel Gusfield, UC - Davis; Leroy Hood, Center for Molecular Biotechnology, Washington; Richard Karp, Washington; Eric Lander, MIT - Whitehead Institute; Eugene Lawler, UC - Berkeley; Joshua Lederberg, Rockefeller; Joseph Naus, Rutgers: Michiel Noordewier, Rutgers: Wilma Olson, Rutgers: David B. Searls, University of Pennsylvania; William Sofer, Rutgers; Tandy Warnow, University of Pennsylvania; Peter Winkler, AT&T Bell Laboratories; and Andrew Yao, Princeton.

Workshops:

We provide detailed information here about workshops held since September 1995. During the period August 1994 to August 1995, we held a tutorial series on mathematical support for molecular biology, workshops on mapping and sequencing, sequence alignment, and phylogeny, an international conference on HIV sequence variation and statistical methods, and mini-workshops on combinatorial structures in molecular biology, DNA topology and structure, global minimization of nonconvex energy functions, sequence-based methods for protein folding, DNA based computers, and geometrical methods for conformational modeling. Details about these events can be found in the DIMACS web pages and in an earlier annual report on this project.

- **4th DIMACS Algorithm Implementation Challenge: Two Problems in Computational Biology: Fragment Assembly and Genome Rearrangements**
  - Dates: Sept 14 - 15, 1995
  - Location: Rutgers University
  - Attendance: 55
  - Organizers: Michael Waterman, University of Southern California; Martin Vingron, German National Research Center for Computer Science (chair); Ellson Chen, Applied Biosystems; Sorin Istrail, Sandia National Laboratory; David Johnson, AT&T Bell Laboratories; John Kececioglu, UC - Davis; Joachim Messing, Rutgers; Joseph Nadeau,

Jackson Laboratory; Pavel Pevzner, Pennsylvania State University; Peter Rice, Sanger Center

There still is a large gap in Computational Molecular Biology between biological application and algorithmic work aimed at answering biological questions. The goal of this implementation challenge was to help bridge this gap. *Sequence assembly* is the problem of reconstructing a DNA sequence knowing only the sequences of many overlapping fragments of it. *Genome rearrangements* is the problem of comparing two orders of genes under the assumption that one order is transformed into another by a series of re-arrangements. For both problems there exist algorithms and software that in the case of fragment assembly are widely used. The Challenge was a year long program that began with these problems; participants developed a variety of projects to test algorithmic approaches, gather useful data sets, and build libraries of data structures to use experimental methods to study these problems.

Ellson Chen, a "cutting edge biological sequencer," defined real data sets and provided intermediate biological feedback. Participants are highly motivated to continue coordinated work on testing their algorithms on common benchmmark data. The addition of further, large, and difficult data sets was agreed upon. A major result was that scientists from industry not only participated actively, but helped with the benchmark data sets by giving away their raw data at a very early stage – something rather unusual. We received data from Applied Biosystems, a Division of Perkin-Elmer, and Genome Therapeutics and anticipate more from the Institute for Genomic Research. These are among the leading genome sequencing companies. Since participants from both academia and industry tested their programs in the same way, there was a lively discussion. It has become apparent that current experimental techniques lead to a need for extremely powerful software and this observation is now strongly influencing ongoing research.

- **Mini-workshop on Mathematics of Drug Discovery**
  - Dates: Sept. 21-22, 1995
  - Location: Rutgers University
  - Attendance: 75
  - Organizers: Tim Havel, Harvard and I. G. Kuntz, UC - San Francisco

The search for new drugs to combat disease is an endeavor of growing importance, especially as organisms once thought conquered begin to develop immunity to existing compounds. For many years, the chief strategy involved finding compounds by screening large numbers of naturally occurring substances for activity. The active compounds were then identified, and synthesized in hundreds or thousands of variations on their basic structure until something acceptable was found. This is often described as the "shotgun" approach.

Attempts to predict biological activity from chemical structure have led to a veritable alchemy of correlations, which continue today under the name of "quantitative structure activity relations", or QSAR. With the advent of reliable methods of determining spatial molecular structure and inexpensive high-speed computers, the field of 3-D structure-based drug design was born. To date, however, the lack of a simple and reliable method of predicting the binding affinity of a "ligand" to a "receptor" protein of known structure (molecular complementarity) has limited its success. More recently, methods based on "combinatorial chemistry" and "test-tube evolution" have been developed. These promise to make it possible to screen large numbers of compounds simultaneously and/or systematically search for those with improved activity. In each of these diverse approaches, significant computational and mathematical problems exist which limit their applicability and effectiveness. This mini-workshop explored recent developments in such fields as discrete and computational geometry, statistics, data analysis and systems theory, which may prove applicable to any or all of the above approaches, while at the same time helping mathematicians and computer scientists to identify the relevant problems.

The majority of speakers and participants were from industry. There was financial sponsorship from Biosym Tech., Ciba Geigy, Merck Research, and SmithKline Beecham. This mini-workshop was designed to acquaint pharmaceutical and computational chemists with the latest developments in such fields as discrete and computational geometry and to introduce discrete mathematicians and theoretical computer scientists to the problems of the field coming to be known as "combinatorial chemistry," a field that so far consists largely of heuristic algorithms based upon chemical intuition. The mini-workshop led to lots of discussion and to some promising beginning contacts, for example one between a Chiron Corp. chemist and a DIMACS mathematician who have common interests and ideas about the use of large-scale cluster analysis in drug design.

- **Mini-workshop on Gene Finding & Gene Prediction**
  - Dates: Oct. 14-15, 1995
  - Location: University of Pennsylvania
  - Attendance: 100
  - Organizers: David Searls, Genome Center, University of Pennsylvania; Jim Fickett, Los Alamos; Michiel Noordewier, Waksman Institute, Rutgers; and Gary Stormo, Colorado

The Workshop on Gene-Finding and Gene Structure Prediction was concerned with the increasingly important activity in computational biology of discovering protein-encoding genes in otherwise uncharacterized primary sequence data. This has traditionally been done in genomic sequences by discriminating likely coding regions based on a variety of statistical analyses and by detection of landmark sequences such as splice junctions. Recent approaches have involved combination of such evidence using rule-based and/or connectionist architectures, and have dealt in a variety of ways with the combinatorial problem of exon assembly (dynamic programming, clustering, etc.) The recent profusion of expressed-sequence data and related techniques has also raised new issues and opportunities. This workshop explored topics such as compositional measures of exonic tendency (including approaches founded in statistics, information theory, and signal processing), the effects of genome heterogeneity, the role of models of biological signals and processes, dealing with incomplete and error-prone sequence data, algorithmic and probabilistic techniques, and similarity-based gene prediction. Problems of interest include detecting coding sequences and assembling gene models from large-scale genomic sequence, collections of expressed sequence fragments, and sets of putative exons from a region. Practical issues of interest included dataset and performance metric standardization, annotation of genome databases, and software interoperability. Major funding for this mini-workshop was provided by SmithKline Beecham.

- **Second Sandia National Laboratories Workshop on Computational Molecular Biology**
  - Dates: March 4-6, 1996
  - Location: Albuquerque, NM
  - Attendance: 135

- Organizers: Sarina Bromberg, David Greenberg, William E. Hart, Sorin Istrail, and Cynthia Phillips (all of Sandia National Laboratories)

The workshop, jointly sponsored by DIMACS, focused on open problems and research directions in Computational Molecular Biology. In addition to talks given by active researchers in this field, several plenary lectures were presented by speakers of highest distinction in computational molecular biology. Plenary speakers were Ken Dill, Richard Karp, Martin Karplus, Jonathon King, Irwin Kuntz, and Michael Waterman.

- **Workshop on Computational Biology** organized in collaboration with the University of Pennsylvania as part of the 50th Anniversary of ENIAC
  - Dates: May 17-19, 1996
  - Location: Princeton University
  - Attendance: 90
  - Organizers: Tandy Warnow, University of Pennsylvania; Craig Benham, Mt. Sinai; Sue Davidson, University of Pennsylvania; Warren Ewens, University of Pennsylvania; Martin Farach, Rutgers; Sampath Kannan, University of Pennsylvania

The conference was based around sessions focusing on important research areas within the field of computational biology. The special research topics included: Structure Prediction, Tandem Repeats, Databases, Sequencing and Mapping, and Genome Rearrangements. There was also a special session on interdisciplinary education in computational biology, in which workshop organizers and students in various training programs throughout the country participated.

- **2nd DIMACS Workshop on DNA Based Computers**
  - Dates: June 10-12, 1996
  - Location: Princeton University
  - Attendance: 110
  - Organizers: Eric Baum, NEC; Dan Boneh, Princeton; Peter Kaplan, Princeton; Richard Lipton, Princeton; John Reif, Duke; Nadrian Seeman, New York University

The workshop hosted papers in all areas that relate directly to computing with DNA, including: Algorithms, applications, techniques, architectures, practical obstacles to DNA based computers, proposed solutions to such obstacles, computational processes in vivo, and relevant ideas regarding biological evolution.

- **Workshop on Mathematical Hierarchy and Biology**
  - Dates: Nov. 13-15, 1996
  - Location: Rutgers University
  - Attendance: 75
  - Organizers: Martin Farach, Rutgers; Boris Mirkin, DIMACS; F.R. McMorris, Louisville; Fred Roberts, Rutgers; Andrey Rzhetsky, Pennsylvania State University

Viewing hierarchies as combinatorial objects underlies many important areas of current scientific investigation. Examples of areas where this is the case are biology, evolutionary studies and taxonomy, general classification theory, data/knowledge bases, cognitive models and linguistical structures. The mathematical study of hierarchies is being pursued in many disciplines, such as computer science, statistics, biology, psychology, and, of course, mathematics.

Topics of presentations included the representation and enumeration of various types of hierarchies, the comparison of hierarchies, constructing hierarchies from various kinds of data, modeling of hierarchical behavior, etc.

- **Computational and Physical Aspects of Protein Folding and Related Problems** as part of the semi-annual Statistical Mechanics Conference at Rutgers.
  - Dates: Dec. 15-17, 1996
  - Location: Rutgers University
  - Attendance: 120
  - Organizers: Joel L. Lebowitz, Rutgers; Shura Grosberg, MIT; Mona Singh, DIMACS; Scott Decatur, DIMACS

  This special session reviewed physical aspects of biological problems, with an emphasis on protein folding. There were talks on random walks on microtubules, DNA and recognition, why proteins look like proteins, the statistical mechanics of protein folding, cellular protein folding, combinatorial methods for protein design, thermodynamics vs. kinematics in protein folding, correlations and heterogeneity of model protein and heteropolymer energy landscapes, conformations of charged heteropolymers, freezing transition of compact polyampholytes, spectral analysis of biological sequences, specificity and affinity of biomolecular interactions, modeling cellular guts with colloidal soups, cell dynamics of model proteins, and physics from jelly fish.

Workshops Planned for the Rest of 1996-97:

- **Mini-Workshop on Protein Structure**
  - Winter 1997: Dates to be determined
  - Location: To be announced
  - Organizers: Mona Singh, DIMACS; Scott Decatur, DIMACS
- **Workshop on DNA Topology II**
  - Date: April 3-4, 1997
  - Location: Rutgers University
  - Organizers: Wilma Olson, Rutgers; Bernard Coleman, Rutgers; Tamar Schlick, New York University; Irwin Tobias, Rutgers
- **Second DIMACS Workshop on Phylogeny**
  - Spring/Summer 1997: Dates to be determined
  - Location: To be announced
  - Organizer: Martin Farach, Rutgers
- **3rd DIMACS Workshop on DNA Based Computers**
  - Dates: June 1997
  - Location: University of Pennsylvania
  - Organizers: Harvey Rubin, University of Pennsylvania (chair); Len Adleman, University of Southern California; Carter Bancroft, Mt. Sinai; Anne Condon, Wisconsin; Richard Lipton, Princeton; David Wood, Delaware