Achievements of the DIMACS “Special Year” on Mathematical Support for Molecular Biology August 1994 to August 2000

The DIMACS “Special Year” on Mathematical Support for Molecular Biology commenced in 1994-1995 with a very intense level of activity and continued at a lower level until August 2000. Through that special “year,” we focused the interests of a large number of computer scientists and mathematicians on the field of molecular biology, created genuine partnerships between mathematical and biological scientists, and trained outstanding young people to work in computational biology. At the time it was launched, there was a need to review and coordinate a variety of topics and methods, and to introduce many people to the field. Computational biology prior to the special year was largely focused on a few problems such as alignments, physical mapping, and reconstructing phylogenies. Now, many of these topics emphasized in the special year have become “standard” topics of research in computational biology and the field has become much broader in its focus. Other topics have begun to be investigated in a much more far-reaching way using methods of computer science and mathematics, 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) have been brought to the attention of a much larger number of computer scientists so that they can formulate and investigate the right questions concerning new technologies.

As a result, a whole host of new problems at the interface between the computational/ mathematical sciences and the biological sciences have arisen, stimulated 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. This was the motivation for DIMACS to launch a second special focus on Computational Molecular Biology, starting in September 2000.

The DIMACS Special Year on Mathematical Support for Molecular Biology included an extensive series of tutorials and workshops, a program of interdisciplinary training of a group of outstanding postdoctoral fellows, a visitor program, graduate and undergraduate student research, a distinguished lecture series, and an algorithm implementation challenge. The success of the 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 in particular and mathematical science in general to useful biological applications, and that we can influence the research of individuals in the mathematical sciences, including some very talented young people, by exposing them to real problems from another discipline.

1. Partnerships

A key to the success of this activity was the establishment of partnerships with other institutions, such as the Waksman Institute for Molecular Genetics at Rutgers, the Rutgers Center for Molecular Biophysics and Biophysical Chemistry, the Center for Theoretical and Applied Genetics at Rutgers, the Center for Molecular Biotechnology at the University of Washington, the Genome Center and the Department of Computer Science at the University of Pennsylvania, the Departments of Molecular Biology and of Ecology and Evolutionary Biology at Princeton, the Program in Mathematics and Molecular Biology originally at UC-Berkeley and now at Florida State University, and the Biomathematical Sciences Department at the Mt. Sinai School of Medicine. Participating biotech and pharmaceutical companies included BIOSYM Tech., Ciba Geigy, CuraGen, Genome Therapeutics, Hoffman LaRoche, the Institute for Genomic Research, Johnson and Johnson, Merck, Perkin-Elmer, Pharmagenics, Roche Molecular Systems, Seq Limited, and SmithKline Beecham.

2. Research Accomplishments

Here, we give details of just a few of the many research accomplishments that can be traced to the work of those participating in the special year on Mathematical Support for Molecular Biology or forming interdisciplinary collaborations that resulted from the special year

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 so-called "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, computer scientist Dannie Durand (a permanent member of DIMACS who went through our special year postdoc program as a 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 suggested new hypotheses to be tested through laboratory and field work (Durand, et al., 1997). This collaboration, which started during the special year tutorial, has continued. It has led Durand to change her field from computer science to biology, to take positions in Biology Departments at Princeton and Carnegie Mellon, and to get her first grant for biological sciences research.

<u>Models of the Progression of Cancerous Tumors</u>. Cancer progression is marked by large-scale over-duplication or disappearance of certain genes, a phenomenon known as a copy number aberration (CNA). Using hybridization methods, laboratory researchers have been able to detect and isolate CNAs, which has led to the discovery of cancer-related genes. It is a long-established fact that CNAs do not appear randomly. The Vogelstein model explained how colorectal cancer is marked by the appearance of a sequence of events appearing in a typical order. Rutgers mathematics graduate student Richard Desper presented a new model for the appearance of CNAs, based upon classical percolation models. Desper showed how a tree-based percolation model, the signal propagation tree (SP tree), generates a probability distribution similar to the well-known Cavender Farris (CF) tree. He modified the CF-reconstruction algorithms to fit SP trees and showed that the new algorithms are consistent, as well as calculating the rate of convergence to the true tree. He has since gone on to work on a variety of models of tumor progression. See Simon, et al. (2000), Desper, et al. (1999), Desper, et al. (to appear), Bissig, et al. (1999), Kainu, et al. (2000). After getting his Ph.D. in Mathematics, Desper went to work as a postdoctoral fellow at the German Cancer Research Center DKFZ and is now a research fellow at the National Center for Biotechnology Information at the National Library of Medicine.

<u>Constructing Optical Maps by Physical Mapping</u>. Permanent member and former postdoc S. Muthukrishnan from AT&T Labs and Laxmi Parida (Muthukrishnan and Parida, 1997) initiated the algorithmic study of physical mapping of DNA with the emerging technology of Optical Mapping (OM) pioneered by David Schwartz. In electrophoretic approaches to physical mapping, information about the relative ordering of the fragments comprising the DNA molecule is lost, thus leading to difficult computational problems of composing the fragments into a physical map depicting their relative order. In contrast, the relative ordering of the pieces is readily obtained in OM. However, OM faces serious technological challenges as it has low resolution and is fault-prone. Muthukrishnan and Parida took a combinatorial approach to the problem of constructing physical maps from the erroneous data generated by OM. They obtained a heuristic algorithm that rapidly and accurately (within 3% error) constructs the physical map from input data with immense experimental errors and imprecision (even with only 10% expression of a restriction site in the molecules.) Their strong experimental results indicate that, although OM has immense experimental imprecision, the errors appear to be "local" and hence more easily manageable than the ones in other approaches where the errors appear "global."

<u>Group Testing with Sequences</u>. Suppose a large population of items is partitioned into two classes, positives and negatives. We wish to locate positive items, but to examine each item is prohibitively expensive. However, we can group the items into subsets and can test if a subset contains at least one positive item. We can then identify all positive items through a number of group tests. The theory of group testing arose in connection with testing millions of World War II military draftees for syphillis (Dor'fman, 1943). The subject is very relevant to large-scale schemes for testing of blood for viruses such as HIV. The modern theory of group testing (Du and Hwang, 2000) is heavily influenced by combinatorial methods, in particular by the methods of combinatorial designs. Group testing arises in connection with the mapping of genomes. We have a long list of molecular sequences, form a library of subsequences (clones), and test whether or not a particular sequence (a probe) appears in the library by testing to see which clones it appears in. We do this by pooling the clones into groups, since a clone library can have thousands, even hundreds of thousands, of clones. Motivated by the economics of maintaining a clone library by robotic means, permanent member Martin Farach-Colton, special year

visitors Sampath Kannan and Manny Knill and permanent member and former postdoc S. Muthukrishnan from AT&T Labs designed and solved many novel group testing problems by formulating them as coding theory questions and using classical coding theory results. (See Farach, et al., 1997.)

Shotgun Sequencing/Genome Rearrangements. DIMACS ran an "Algorithm Implementation Challenge" during the special year in which the biological/biotech participation 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. Scientists from industry not only participated actively, but shared their raw data at a very early stage -- something rather unusual. Some of the alignment procedures presented consistently suggested an alternative and more plausible reconstruction for one of the test cases provided by a sequencing company, leading to a rethinking of alignment procedures. It has become apparent that current experimental techniques lead to a need for extremely powerful software and this is now strongly influencing ongoing research.

Sequence Reconstruction through Sequencing by Hybridization. Problems of sequence reconstruction led to some interesting results by special year visitors Richard Arratia and Bela Bollobas, Don Coppersmith, and Greg Sorkin. The original question they considered arises in Sequencing By Hybridization (SBH), which is the attempt to reconstruct a long sequence from the knowledge of which q-tuples it contains, without being given the knowledge of the order in which those q-tuples occur. For random sequences, through deBruijn graphs, this reduces to a question of enumerating euler circuits. Arratia, et al. (to appear) showed that the interlacement graph, which gives only the relation between pairs of repeated pairs of (q-1)-tuples, determines the number of euler circuits

Positional Sequencing by Hybridization. To improve the resolving power of SBH, Broude and others have 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, in collaboration with W. Feldman,and Herbert Lewis, 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 (Hannenhalli, et al., 1996) 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. Stronger results building on these were obtained more recently by visitors A. Ben-Dor and Ron Shamir, in collaboration with I. Pe'er and R. Sharan; they developed new reconstruction algorithms and tightened certain hardness results for PSBH (see Ben-Dor, et al., 1999).

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 (Bafna, Muthukrishnan, and Ravi, 1995) 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 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. Many aspects of this work have been followed up extensively in the community, e.g., by researchers Tiao Jiang and John Kececioglou,

Parametric Alignment. When using sequence alignment methods to study DNA or amino acid sequences, there is often considerable disagreement about how to weight matches, mismatches, insertions and deletions, and gaps. The most commonly used alignment software packages require the user to specify fixed values for those parameters. This choice can greatly affect the biological significance of the resulting alignment. Parametric sequence alignment is a tool that efficiently explores such penalty variation and avoids the problem of choosing fixed parameter settings by computing the optimal alignment as a function of variable parameters for weights and penalties. Special year visitor Dan Gusfield and his collaborators K. Balasubramanian and D. Naor (Gusfield, Balasubramanian, Naor, 1994) designed efficient algorithms for this problem and implemented these in a widely used software package.

Combinatorial Approaches to Gene Finding. The problem of identifying the introns and exons constituting a gene is an important one. Traditional approaches to this problem have typically been statistical. Visitor Pavel Pevzner, in collaboration with Mischa Gelfand, developed a different approach, using a new dynamic programming algorithm. Experimental evaluation of this algorithm has shown that it is very successful at prediction. (See Gelfand, Mironov, and Pevzner, 1996.)

DNA Computing. A model of computation that has generated considerable excitement in recent years is the DNA computer. The goal is to use biological processes undergone by DNA and other biomolecular sequences to perform useful and difficult computations. Laura Landweber, a Princeton biologist, and Richard Lipton, a Princeton computer scientist and permanent member of DIMACS, started a collaboration during the special year on Mathematical Support for Molecular Biology, and the work rapidly progressed from the theory to the wet lab. (This collaboration led to a DIMACS Special Focus on DNA Computing during the period 1997-1999.) In an important breakthrough, Landweber, Lipton, and their colleagues designed a computer built from RNA to solve a chess problem, the problem of determining the positions knights can occupy on a chessboard such that no knight can attack any other knight on the board. Though not NP-complete for a defined board, this problem is from the class of NP-complete “satisfiability” problems. Anthony Cukras, Dirk Faulhammer, Lipton, and Landweber developed a theory of RNA computing and a method for solving the knight problem. Using only biological molecules and enzymes as tools, they developed and successfully executed an algorithm for solving this variant of the knight problem on a 3x3 chessboard using a 10-bit combinatorial pool and a protocol involving sequential R-NASE H ‘destroy' operations. The results of preliminary experiments revealed that the protocol recovers far more correct solutions than expected at random, and further work led to the successful completion of the largest molecular computation in any lab to date. See Faulhammer, et al., 2000). (See Knight, 2000 for a comment on this work.)

Structural Motif Recognition, With Applications to the Histidine Kinase Protein Family. Biologists are interested in predicting protein structure because the structure of a protein provides insight into its biological function. Experimental approaches to determining a protein's structure are difficult and time-consuming. As a result, much research has focused on applying mathematical and computational techniques to predicting protein structure from one-dimensional sequence data. A first step to tackling this complex problem is a solution to the *structural motif recognition problem*: given a known, commonly occurring three-dimensional substructure, or motif, determine whether this motif occurs in a given amino acid sequence. Special year visitor Bonnie Berger and DIMACS postdoc Mona Singh developed an iterative algorithm that uses randomness and statistical techniques to improve existing methods for recognizing protein structural motifs (Berger and Singh, 1997). The algorithm is particularly effective in situations where large numbers of sufficiently diverse examples of the motif are not known. These are precisely the situations that pose significant difficulties for previously known methods. They implemented their algorithm and showed that it is able to substantially improve performance on the coiled coil motif. In collaboration with biologists at the Whitehead Institute and Genentech Corporation, Berger, Singh, with Peter Kim, James Berger, and Andrea Cochran, applied their program to the histidine kinase protein family. (See Singh, Berger, and Kim, 1999, Singh, et al., 1998.) Most members of the protein family are sensors that allow microorganisms to adjust to varying conditions in their environments. Their program identified a common but previously undescribed coiled-coil region in the histidine kinase family. Their work also suggested mechanisms by which the predicted coiled coil may mediate kinase regulation and signal transduction. There are some very interesting medical applications, as some dangerous bacteria, such as those causing pertussis and salmenollosis, need histidine kinase proteins to be fully virulent.

Protein Folding. Postdoc Richa Agarwala, Serafim Batzoglou, Vlado Dancik, postdoc Scott Decatur, permanent member Martin Farach-Colton (Rutgers, CS), and special year visitors Sridhar Hannenhalli and Steven Skiena worked on the problem of determining the three-dimensional folding of a protein given its one-dimensional amino acid sequence. They used the HP model for protein folding proposed by Dill in 1985, which models each protein as a chain of amino acid residues that are either hydrophobic or polar, and makes the assumption that hydrophobic interactions are the dominant initial driving force for protein folding. Although it is not known how to find the optimal fold in polynomial time, Hart and

Istrail in 1995 gave polynomial time approximation algorithms for folding proteins on the cubic lattice under the HP model. Agarwala, et al. (1997) examined the choice of lattice by considering its algorithmic and geometric implications and argued that a triangular lattice is a more reasonable choice. They presented a set of folding rules for a triangular lattice and analyzed the approximation ratio which they achieved. In addition, they introduced a generalization of the HP model to account for residues having different levels of hydrophobicity. After developing the biological foundation for this generalization, they showed that in the new model they were able to achieve similar constant factor approximation guarantees on the triangular lattice as were achieved in the standard HP model. While the structures derived from their folding rules are probably still far from biological reality, it is hoped that having a set of folding rules with different properties will yield more interesting folds when combined.

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 permanent member Joel Lebowitz (Mathematics, Rutgers) showed (Vendruscolo, 1999) 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 lattice, 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.

New Approaches to Phylogenetic Tree Reconstruction. One of the great challenges of biomathematics is to develop computationally feasible algorithms to reconstruct phylogeny from biomolecular sequences. Given a set of species with corresponding segments of aligned DNA sequences, the problem is to select the true phylogenetic tree from the set of all possible evolutionary trees. Two complex mathematical issues of phylogeny reconstruction are the computational complexity of the reconstruction algorithm and the sequence length needed for correct reconstruction. Special year visitors Peter Erdos, Mike Steel, Laszlo Szekely and Tandy Warnow initiated analytic studies of the sequence length needed for phylogeny reconstruction. The idea for this research emerged at a phylogeny workshop at DIMACS in 1995. Erdos, Steel, Szekely and Warnow (Erdos, et al. 1999a, Steel and Szekely 1999, Steel and Szekely,2001) introduced information theoretic lower bounds, which hold for all tree reconstruction methods, either deterministic or randomized. They also showed that particular tree reconstruction methods---like the optimality criterion called maximimum compatibility---may require substantially longer sequences. Steel and Szekely (1999) and Steel and Szekely (2001) also showed that the Maximum Likelihood Method, as it is applied in phylogeny, is not optimal even on average, not to speak about the worst case, contrary to opinions which abound in the literature. The paper of Steel and Szekely (2001) is the first giving an explicit sequence length guarantee for the Maximum Likelihood Method that statisticians consider their method of choice. Erdos, et al. (1997a,b, 1999a,b) designed new tree reconstruction methods which lead to the following conclusion: the trivial-looking information theoretic lower bounds are almost sufficient to reconstruct phylogenetic trees under standard stochastic models.

Maximum Agreement Subtrees. Visitor David Swofford, a Smithsonian biologist, and Martin Farach-Colton, a DIMACS permanent member and Rutgers computer scientist, collaborated on algorithms for finding maximum agreement subtrees --- a useful tool in phylogeny construction. Swofford coded up and included Farach-Colton's algorithm for this problem in his popular package PAUP (Swofford, 1998).

Models of Duplication. An important idea in molecular evolutionary biology is 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. A likely evolutionary species' phylogeny is thought to involve a minimum number of duplications and corrsponding loss-of-a-copy events. An explicit graph-theoretic model of the gene duplication in a species tree was developed by special year visitor Boris Mirkin. In the framework of this model, suggestions in the literature about counting duplication based events 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 was proved correct independently by special year visitor Martin Vingron and Oliver Eulenstein and by L. Zhang. This led to an invitation to Mirkin to visit the Division of Bioinformatics of the National Cancer Center of Germany, where he continued this work in collaboration with Vingron. (See Mirkin, Muchnik, and Smith, 1995, Eulenstein, Mirkin, and Vingron, 1998.)

Approximating Numerical Taxonomy. A common problem in phylogeny is to fit an n x n distance matrix D by a tree metric T. Postdocs Richa Agarwala, Vineet Bafna, Babu Narayanan, permanent member Martin Farach-Colton (Rutgers, CS), special year visitor Mikkel Thorup, and Mike Paterson presented an efficient algorithm (quadratic in n) 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 was the first algorithm for the problem with a performance guarantee relative to the optimal solution. (See Agarwala, et al., 1996.) Later, Farach-Colton and special year visitor Sampath Kannan (Farach and Kannan, 1999) used this work as a building block in a clever algorithm for computing maximum likelihood trees. Experiments undertaken by Rutgers graduate student Jaime Cohen and Farach-Colton showed that these distance methods construct better trees in practice than do the most popular heuristics used by biologists (Cohen and Farach, 1997).

Evolutionary Tree Reconstruction: Better Methods for Solving Parsimony and Compatibility. One of the most promising approaches to evolutionary tree reconstruction is to find the "maximum parsimony" tree, while in linguistics the use of the "maximum compatibility" method has been very useful. However, these problems are NP-hard and current approaches to solving them amount to heuristic searches through the space of possible tree topologies (a search which can, on large trees, take months to complete). Former postdoc Maria Bonet, special year visitors Mike Steel and Tandy Warnow, and postdoc Shibu Yooseph (Bonet, et al., 1998) developed a new technique, called optimal tree refinement, for reconstructing very large trees. Their technique is motivated by recent experimental studies which have shown that certain polynomial time methods reliably return contractions of the true tree. They studied the use of this technique in solving parsimony and compatibility, and obtained efficient, polynomial time algorithms.

Evolutionary Tree Reconstruction Using Hybrid Methods. Accuracy in reconstructing evolutionary trees for species sets is often measured by comparing the topologies of the reconstructed tree and the model tree. One of the major debates in the field is whether large evolutionary trees can be even approximately reconstructed from biomolecular sequences of realistically bounded lengths (up to about 2000 nucleotides) using standard techniques (polynomial time distance methods, and heuristics for NP-hard optimization problems). Using both analytical and experimental techniques, postdoc Daniel Huson, S. Nettles, K. Rice, special year visitor Tandy Warnow, and postdoc Shibu Yooseph (Huson, et al., 1999) showed that on large trees, the two most popular methods in systematic biology, neighbor-joining and maximum parsimony heuristics, as well as two promising methods introduced by theoretical computer scientists, are all likely to have significant errors in the topology reconstruction of the model tree. They also found a new general technique for combining outputs of different methods (thus producing hybrid methods), and showed experimentally how one such hybrid method has better performance than its constituent parts.

The Disk-Covering Method for Tree Reconstruction. The reconstruction of evolutionary trees containing significant amounts of divergence is especially challenging. Postdoc Daniel Huson, S. Nettles, L. Parida, special year visitor Tandy Warnow, and postdoc Shibu Yooseph developed a new tree reconstruction method, the Disk-Covering Method (DCM), which can be used to recover accurate estimations of the evolutionary tree for otherwise intractable datasets (see Huson, Nettles, and Warnow, 1999). DCM obtains a decomposition of the input dataset into small overlapping sets of closely related taxa, reconstructs trees on these subsets (using a "base" phylogenetic method of choice), and then combines the subtrees into one tree on the entire set of taxa. Because the subproblems analyzed by DCM are smaller, computationally expensive methods such as maximum likelihood estimation can be used without incurring too much cost. At the same time, because the taxa within each subset are closely related, even very simple methods (such as neighbor-joining) are much more likely to be highly accurate. The result is that DCM-boosted methods are typically faster and more accurate as compared to "naïve" use of the same method.

Nearly Tight Bounds on the Learnability of Evolution. Evolution is often modeled as a stochastic process which modifies DNA. One of the most popular and successful such processes are the Cavender-Farris (CF) trees, which are represented as edge weighted trees. The phylogenetic tree reconstruction problem can be stated as follows: given k samples drawn from a CF tree, output a CF tree which is close to the original by some metric on the space of probability distributions. Rutgers Mathematics graduate student Richard Desper showed that a very simple, polynomial time algorithm converges to the true tree at a rate which differs from the optimum by a constant. He did this by analyzing upper and lower bounds for the convergence rate of learning very simple CF trees, and then showed that the learnability of each CF tree is sandwiched between two such simpler trees. (See Desper and Vingron, 2000.) Each CF tree naturally defines a random variable, and the gold-standard for reconstructing such trees is the Maximum Likelihood Estimator of this variable. This approach is notoriously computationally expensive. Graduate student visitor Andris Ambainis, M. Farer, permanent member Martin Farach-Colton, and special year visitor Sampath Kannan (see Ambainis, et al., 1999) showed that a very simple algorithm, which is a variant on one of the most popular algorithms used by practitioners, converges to the true tree at a rate which differs from the optimum by a constant. They did this by analyzing upper and lower bounds for the convergence rate of learning very simple CF trees, and then showing that the learnability of each CF tree is sandwiched between two such simpler trees.

The Linear-Cost Subtree-Transfer Distance between Phylogenetic Trees. Different phylogenetic trees for the same group of species are often produced either by procedures that use diverse optimality criteria or from different genes in the study of molecular evolution. Comparing these trees to find their similarities and dissimilarities (i.e., distance) is thus an important issue in computational molecular biology. Several distance metrics including the nearest neighbor interchange (nni) distance and the subtree-transfer distance have been proposed and extensively studied in the literature. Special year visitor Bhaskar DasGupta, Xin He, visiting graduate student Tao Jiang, special year visitor Ming Li and John Tromp (DasGupta, et al., 1999) considered a natural extension of the subtree-transfer distance, called the linear-cost subtree-transfer distance, and found efficient approximation algorithms for this distance as well as its relationship to the nni distance. The linear-cost subtree-transfer model seems more logical than the (unit-cost) subtree-transfer model in some applications. They showed that the linear-cost subtree-transfer distance is in fact identical to the nni distance on unweighted phylogenies. They found an efficient algorith to compute an optimal linear-cost subtree-transfer sequence between unweighted phylogenies that is quadratic in n as long as the linear-cost subtree-transfer distance is small. They showed that computing the linear-cost subtree-transfer distance between two weighted phylogenetic trees is difficult, i.e., NP-hard, provided one allows multiple leaves of a tree to share the same label (i.e. the trees are not necessarily uniquely labeled). Finally they showed that there is an efficient approximation algorithm for computing the linear-cost subtree-transfer distance between weighted phylogenies with performance ratio 2, i.e., that is guaranteed to come within a factor of 2 of the optimal.

3. Books

Among the highlight publications of the special year on Mathematical Support for Molecular Biology are books and special issues of journals. Until seven years ago, no adequate textbooks were available for computational biology. The DIMACS special year played an important role in the development of two pioneering textbooks in the field. The long visits to DIMACS by Daniel Gusfield and Michael Waterman helped shape their books Gusfield (1998) and Waterman (1995). Some important books coming out of the special year and appearing in the DIMACS book series provide both an overview of the topic of computational biology (Farach-Colton, Roberts, Vingron, and Waterman, 1999), and the latest thinking about specific topics through conference proceedings (Du, Pardalos and Wang, 2000, Lipton and Baum, 1996, Landweber and Baum, 1999, Rubin and Wood, 1999, Mirkin, McMorris, Roberts, and Rzhetsky, 1997, Pardalos, Shalloway, and Xue, 1996, Winfree and Gifford, 2000). Other proceedings of DIMACS workshops were published in special issues of journals; for instance, volume 2 (No.2) (1995) of the *Journal of Computational Biology* includes the proceedings of our Workshop on DNA Mapping and Sequencing and volume 74 (1998) of the *Biophysical Journal* contains papers presented at our Workshop on DNA Topology.

4. Human Resources Development: Postdoctoral Training, Graduate Student and Undergraduate Student Programs

A major goal of the special year on Mathematical Support for Molecular Biology was to introduce outstanding young people with a background in mathematics or computer science to the problems and challenges of molecular biology. We consider our role in this regard to be one of the most significant accomplishments of the special year, and an outside evaluation of the program concurred (see Section 6). We focused on young people with a background in the mathematical sciences rather than on young people with a background in the biological sciences (with exceptions). The expense of the program, especially the postdoctoral training program, and our desire to involve a sufficient number of young people to achieve a “critical mass,” led us to decide to concentrate on individuals from just one side. The proposed program will address both sides.

The special year on Mathematical Support for Molecular Biology featured an intense training program for postdocs. The postdocs, coming with a background in mathematics or computer science, spent either a year or two at DIMACS, were exposed to topics in molecular biology through a tutorial program and contacts with biological scientists, and were introduced to the leaders in the field of computational biology through the special year workshop and visitor programs. Our emphasis on interdisciplinary activity enabled the postdocs to develop outstanding collaborations with biological scientists during their stay at DIMACS, and they have gone on to excellent positions, many playing leadership roles in computational biology. A list of special year postdocs and their present positions is provided in Appendix 1. The postdoctoral fellows were heavily involved in the research of the special year and contributed to many of its important research accomplishments. Other DIMACS postdocs not associated with this special year also became interested in computational biology through it and contributed interesting scientific results. Their names are also listed in Appendix 1.

Many graduate students from Rutgers and Princeton and other institutions world-wide participated in the program. Students worked on such topics as phylogeny reconstruction, models of tumor progression, physical mapping, searching through databases of molecular sequences, modeling the immune system, design and test of DNA computing lab experiments, proximity searching in protein databases, nearest neighbor searches in phylogenetic tree reconstruction, and computational tools for analysis of proteins based on x-ray crystallographic experiments. Hundreds of graduate students from around the world participated in the special year through the DIMACS program which offered travel support to attend workshops. In addition, DIMACS hosted extended visits by a number of graduate students, who came to work with DIMACS permanent members, visitors, and postdoctoral fellows.

Every summer DIMACS has hosted outstanding undergraduates from around the country, working in directed research on topics connected with the special year. The research areas have included neighbor-joining methods of phylogenetic tree reconstruction, ribbon models for DNA supercoiling, approximation algorithms for protein folding, protein structure prediction, DNA knots, sequence comparison, and DNA symmetric structure.

5. Workshops in the Special Year on Mathematical Support for Molecular Biology

A total of 24 workshops and conferences were held during the special year between 1994 and 2000. For more information about them, see Appendix 2. Programs are available at our website http:dimacs.rutgers.edu , which also contains workshop summaries in the DIMACS Annual Reports.

6. External Report on the Impact of the Special Year on Mathematical Support for Molecular Biology

In 1998, DIMACS commissioned a report on the impact of the special year on Mathematical Support for Molecular Biology, and the report (Gusfield, Kannan, and Pevzner, 1998), says that “The special year has been enormously successful in meeting its goals” and that “DIMACS has played an important role in exciting interest in computational biology and informatics.” Here are some of the highlights of that report:

- DIMACS played a significant role in building bridges between disciplines: “The special year helped launch a number of cross-disciplinary collaborations. Prior to the special year research in computational biology had been rather compartmentalized. The work of computer scientists did not usually find its way into the biologist's toolkit. In addition, computer scientists were working on a small set of cleanly formulated problems without a full knowledge of their biological relevance. It was therefore important to greatly encourage collaborative research.”
- The training program of young people was “the foremost accomplishment” of the special year: “perhaps the most important function of DIMACS has been in training young computer scientists in computational biology so that they go on to establish collaborations later in their career.”
- The importance of establishing partnerships between mathematical scientists and biological scientists was recognized through the involvement in a mathematical sciences center of many prominent biologists: “The success of this effort at training computer scientists in biology depended in large part on the participation of biologists. Many prominent biologists both at Rutgers and from elsewhere in the country participated in the workshops and distinguished lectures organized during the special year.”

- The many research results emanating from the special year show that the “research program was an unqualified success. A few of the tangible consequences are listed below:
    - The field of DNA computing which was at a nascent stage at the time of the first workshop, received a significant boost from it and from subsequent workshops organized with DIMACS participation at Princeton University and at the University of Pennsylvania. This workshop has now become an annual event and the field has a number of active researchers from computer science, biology, biochemistry and other fields.
    - Early work in computational biology prior to 1994 was largely focused on a small set of problems such as alignments, physical mapping, and reconstructing phylogenies. The special year has stimulated work on other important problems such as gene finding and motif recognition, protein and RNA folding, protein structure prediction, and linkage analysis. In addition computer scientists have begun to turn their attention to specific problems motivated by particular biological scenarios. This is no small feat considering that the natural proclivity of computer scientists is to work at the greatest level of abstraction possible. A large part of this change was initiated by the talks given by prominent biologists at various DIMACS fora.
    - New biotechnologies such as sequencing by hybridization, optical mapping, and EST (expressed sequence tags) have been quickly brought to the attention of computer scientists so that they have been able to formulate and research the right questions concerning these technologies.
    - An important part of algorithmic research in computational biology will be experimental validation of the efficacy of the algorithms, since theoretical analysis often yields overly pessimistic answers. This has been facilitated by the setting up of benchmark biological data at least in some problem domains. Thus algorithm designers can test how well their algorithms perform on real instances.”

References

Agarwala, R., Bafna, V., Farach, M., Narayanan, B., Paterson, M., and Thorup, M. (1996), "On the approximability of numerical taxonomy: Fitting distances by tree metrics," *Proc. of the 7th Annual ACM-SIAM Symposium on Discrete Algorithms* (SODA).

Agarwala, R., Batzoglou, S., Dancik, V., Decatur, S., Farach, M., Hannenhalli, S., Muthukrishnan, S., and Skiena, S. (1997), "Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP Model," *J. Computational Biology*, RECOMB Special Issue, **4**, 275-296.

Ambainis, A., Desper, R., Farach, M., and Kannan, S. (1997), "Nearly tight bounds on the learnability of evolution", *38th Symposium on Foundations of Computer Science* (FOCS '97), Miami, Florida.

Arratia, R., Bollobas, B., Coppersmith, D., and Sorkin, G.B. (to appear), Euler circuits and DNA sequencing by hybridization, *Discrete Applied Mathematics.*

Bafna, V., Muthukrishnan, S., and Ravi, R. (1995), "Computing the similarity between RNA strings," *Proc. 6th Symposium on Combinatorial Pattern Matching, Lecture Notes in Computer Science*, **937**, Springer-Verlag, 1-16.

Ben-Dor, A., Pe'er, I., Shamir, R., and Sharan, R. (1999), "On the complexity of positional sequencing by hybridization," *Lecture Notes in Computer Science*, **1645**, 88-98.

Berger, B., and Singh, M. (1997), "An iterative method for improved protein structural motif recogntion," *J. Comput. Biol.*, **4**, 261-273.

Bissig, H., Richter, J., Desper, R., Meier, V., Schraml, P., Schäffer, A., Sauter, G., Mihatsch, M., and Moch, H. (1999), "Evaluation of a clonal relationship between primary and meta-static renal cell carcinoma by comparative genomic hybridization", *Am. J. Pathol.*, **155**, 267-274.

Bonet, M., Steel, M., Warnow, T., and Yooseph, S. (1998), "Faster algorithms for solving parsimony and compatibility," *The Journal of Computational Biology*, **5**, 409-422.

Cohen, J., and Farach, M. (1997), "Numerical taxonomy on data: Experimental results," *Proceedings of First Annual ACM International Conference on Computational Molecular Biology*, RECOMB, Santa Fe, NM.

DasGupta, B., He, X., Jiang, T., Li, M., and Tromp, J. (1999), "On the linear-cost subtree-transfer distance between phylogenetic trees, *Algorithmica*, **25**, 176-195.

Desper, R., Jiang, F., Kallioniemi, O.-P., Moch, H., Papadimitriou, C.H., and Schäffer, A.A. (1999), "Inferring tree models for oncogenesis from comparative genome hybridization", *Journal of Computational Biology*, **6**, 37-52.

Desper, R., Jiang, F., Kallioniemi, O.-P., Moch, H., Papadimitriou, C.H. and Schäffer, A.A. (to appear), "Distance-based reconstruction of tree models for oncogenesis", *Journal of Computational Biology.*

Desper, R., and Vingron, M. (2000), "Tree fitting: an algebraic approach using profile distances", *Fourth Annual International Conference on Computational Molecular Biology* (RECOMB '00), Tokyo, Japan.

Dorfman, R. (1943), "The detection of defective members of a large population," *Annals of Math. Stat.*, **14**, 436-440.

Du, D-Z, and Hwang, F.K. (2000), *Combinatorial Group Testing and its Applications*, 2nd ed., World Scientific, Singapore.

Du, D-Z., Pardalos, P.M., and Wang, J. (2000), *Discrete Mathematical Problems with Medical Applications,* DIMACS Series, **55**, American Mathematical Society, Providence, RI.

Durand, D., Ardlie, K., Buttel, L., Levin, S.A., and Silver, L.M. (1997), "Impact of migration and fitness on the stability of a lethal *t*-haplotype polymorphism in house mice: A computer study," *Genetics*, **145**, 1093-1108.

Erdos, P.L., Steel, M.A., Székely, L.A., and Warnow, T. J. (1997a), "Local quartet splits of a binary tree infer all quartet splits via one dyadic inference rule," *Computers and Artificial Intelligence*, **16**, 217-227.

Erdos, P.L., Steel, M.A., Székely, L.A., and Warnow, T. J. (1997b), "Constructing big trees from short sequences," in P. Degano, R. Gorrieri, and A. Marchetti-Spaccamela (eds.), *ICALP'97, 24th International Colloquium on Automata, Languages, and Programming* (Silver Jubilee of EATCS), Bologna, Italy. Lecture Notes in *Computer Science* (1997), Springer-Verlag, 1256, 1-11.

Erdos, P.L., Steel, M.A., Székely, L.A., and Warnow, T. J. (1999a), "A few logs suffice to build (almost) all trees I," *Random Structures and Algorithms*, **14**, 153-184.

Erdos, P.L., Steel, M.A., Székely, L.A., and Warnow, T. J. (1999b), "A few logs suffice to build (almost) all trees II," *Theor. Comp. Sci*, **221**, 77-118.

Eulenstein, O., Mirkin, B., and Vingron, M. (1998), "Duplication-based measures of difference between gene and species trees," *Journal of Computational Biology*, **5**, 135-148.

Farach, M., and Kannan, S. (1999), "Efficient algorithms for inverting evolution," *J. of the Association for Computing Machinery*, **46**, 437-449.

Farach, M., Kannan, S., Knill, E., and Muthukrishnan, S. (1997), "Group testing problems with sequences in experimental molecular biology," *Proc. of Sequences '97 (Conference on Compression and Complexity of Sequences*, Positano, Salerno, Italy).

Farach-Colton, M., Roberts, F.S., Vingron, M., and Waterman, M. (1999) (eds.), *Mathematical Support for Molecular Biology*, DIMACS Series, **47**, American Mathematical Society, Providence, RI.

Faulhammer, D., Cukras, A.R., Lipton, R.J., and Landweber, L.F. (2000), "Molecular Computation: RNA Solutions to Chess Problems," *Proc. Nat. Acad. Sci.*, **97**, 1385-1389.

Gelfand, M., Mironov, A., and Pevzner, P. (1996), "Gene recognition via spliced alignment," *Proc. National Academy of Sciences*, **93**, 9061-9066.

Gusfield, D. (1998), *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, Cambridge University Press.

Gusfield, D., Balasubramanian, K., and Naor, D. (1994), "Parametric optimization of sequence alignment," *Algorithmica*, **12**, 312-326.

Gusfield, D., Kannan, S., and Pevzner, P. (1998), "Evaluation of the DIMACS Special Year on Mathematical Support for Molecular Biology", http://dimacs.rutgers.edu/About/Reports/SYMB.html

Hannenhalli, S., Feldman, W. Lewis, H., Pevzner, P., and Skiena, S. (1996), "Positional sequencing by hybridization computer applications in the biological sciences (CABIOS)," **12**, 19-24.

Huson, D., Nettles, S., Rice, K., Warnow, T., and Yooseph, S. (1999), "The Hybrid tree reconstruction method," *The Journal of Experimental Algorithmics*, **4**, Article 5.

Huson, D., Nettles, S., and Warnow, T. (1999), "Disk-Covering, a fast converging method for phylogenetic tree reconstruction," *Journal of Computational Biology*, **6**, 369-386.

Knight, J. (2000), "Game of life: Biocomputers take their first step with a little chess puzzle," *New Scientist*.

Landweber, L.F., and Baum, E.B. (1999) (eds.), *DNA Based Computers II*, DIMACS Series, **44**, American Mathematical Society, Providence, RI.

Lipton, R.J., and Baum, E.B. (1996) (eds.), *DNA Based Computers*, DIMACS Series, **27**, American Mathematical Society, Providence, RI.

Mirkin, B., McMorris, F.R., Roberts, F.S., and Rzhetsky, A. (1997) (eds.), *Mathematical Hierarchies and Biology*, DIMACS Series, **37**, American Mathematical Society, Providence, RI.

Mirkin, B., Muchnik, I., and Smith, T.F. (1995), "A biologically consistent model for comparing phylogeny trees," *Journal of Computational Biology*, **2**, 493-507.

Muthukrishnan, S., and Parida, L. (1997), "Towards constructing physical maps by optical mapping: An effective, simple, combinatorial approach," *Proc. of the First ACM Conference on Computational Molecular Biology* (RECOMB 97), Santa Fe, NM, 202-219.

Pardalos, P.M., Shalloway, D., and Xue, G. (1996) (eds.), *Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding*, DIMACS Series, **23**, American Mathematical Society, Providence, RI.

Rubin, H., and Wood, D.H. (1999) (eds.), *DNA Based Computers III*, DIMACS Series, **48**, American Mathematical Society, Providence, RI.

Simon, R., Desper, R., Papadimitriou, C., Peng, A., Alberts, D., Taetle, R., Trent, J., and Schäffer, A. (2000), "Chromosome abnormalities in ovarian adenocarcinoma III: Using breakpoint data to infer and test mathematical models for oncogenesis," *Genes, Chromosomes, and Cancer*, **28**, 106-120.

Singh, M., Berger, B., and Kim, P.S. (1999), "LearnCoil-VMF: Computational evidence for coiled-coil-like motifs in many viral membrane-fusion proteins," *J. Mol. Bio.*, **290**, 1031-1041.

Singh, M., Berger, B., Kim, P.S., Berger, J.M., and Cochran, A.G. (1998), "Computational learning reveals coiled coil-like motifs in histidine kinase linker domains," *Proc. Natl. Acad. Sci. USA*, **95**, 2738-2743.

Steel, M.A., and Székely, L.A. (1999), "Inverting random functions," *Annals of Combinatorics*, **3**, 103-113.

Steel, M.A., and Székely, L.A. (2001), "Inverting random functions II: Explicit bounds for parametric and non-parametric MLE, with applications", preprint.

Swofford, D. L. (1998), *PAUP* 4.0 - Phylogenetic Analysis Using Parsimony (and Other Methods)*, Sinauer Assoc., Sunderland, MA.

Vendruscolo, M., Subramanian, B., Kantor, I., Domany, E., and Lebowitz, J. (199), "Statistical properties of contact maps," Physical Review, 59, 977-984.

Waterman, M. (1995), *Introduction to Computational Biology*, Chapman and Hall.

Winfree, E., and Gfford, D.K. (2000) (eds.), *DNA Based Computers V*, DIMACS Series, **54**, American Mathematical Society, Providence, RI.

Appendix 1: Postdoctoral Fellows in the DIMACS Special Focus on Mathematical Support for Molecular Biology

Special Year Postdocs and their Current Position:

Richa Agarwala, National Center for Biotechnology Information, NIH
Vineet Bafna, Celera Genomics
Scott Decatur, Grantham, Mayo, van Otterloo & Co. LLC
Dannie Durand*, Assistant Professor of Biological Sciences, Carnegie Mellon University
S. Muthukrishnan, AT&T Labs Research
R. Ravi, Associate Professor of Operations Research and Computer Science, Carnegie Mellon University
Mona Singh, Assistant Professor of Computer Science, Princeton University
Elizabeth Sweedyk, Clare Boothe Luce Assistant Professor of Computer Science, Harvey Mudd College
Shibu Yooseph, Celera Genomics

*Durand was not formally a postdoc, but she went through the postdoc program as a visitor.

Other Postdocs not Formally Associated with the Special Year but Who Interacted with the Program:

Maria Bonet, Lenguajes y Sistemas Informáticos, Universidad Politécnica de Cataluña
Hui Chen, Global Modelling and Analytics, Credit Suisse First Boston
Daniel Huson, Celera Genomics
Babu Narayanan, Nortel Networks

Appendix 2: Workshops in the DIMACS Special Year on Mathematical Support for Molecular Biology

Workshop on Combinatorial Methods for DNA Mapping and Sequencing
October 6, 1994 - October 9, 1994
DIMACS Center, CoRE Building, Rutgers University, Piscataway, NJ
Attendance: 130
Organizers:
Pavel Pevzner, Pennsylvania State University
Charles Cantor, Boston University

Mini-Workshop on Combinatorial Structures in Molecular Biology
November 11, 1994
DIMACS Center, CoRE Building, Rutgers University, Piscataway, NJ
Attendance: 70
Organizers:
Jerrold R. Griggs, Chair, University of South Carolina
Timothy F. Havel, Harvard University
Ron Shamir, Tel Aviv University
Temple F. Smith, Boston University
Peter M. Winkler, Bellcore

Workshop on Sequence Alignment
November 10 - 12, 1994
Princeton University, Princeton, NJ
Attendance: 74
Organizer:
Alberto Apostolico, Purdue Univesity

Workshop on DNA Topology and Structure
December 9, 1994
DIMACS Center, CoRE Building, Rutgers University, Piscataway, NJ
Attendance: 40
Organizers:
Craig Benham, Mt. Sinai School of Medicine
Bill Bauer, Stony Brook

Phylogeny Workshop
February 6 - 8, 1995
Princeton University, Princeton, NJ
Attendance: 99
Organizer:
Simon Tavare, University of Southern California

Workshop on Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding
March 20 - 21, 1995
DIMACS Center, CoRE Building, Rutgers University, Piscataway, NJ
Attendance: 52
Organizers:
Panos M. Pardalos, University of Florida
David Shalloway, Cornell University
Guoliang Xue, University of Vermont

Mini-Workshop on Sequence-Based Methods for Protein Folding
March 24, 1995

DIMACS Center, CoRE Building, Rutgers University, Piscataway, NJ
Attendance: 45
Organizer:
Bonnie Berger, MIT

Mini-workshop: DNA Based Computers
April 4, 1995
Princeton University, Princeton, NJ
Attendance: 201
Organizers:
Dick Lipton, Princeton University
Eric Baum, NEC Research Institute

Workshop on HIV Sequence Analysis
May 3 - 5, 1995
DIMACS Center, CoRE Building, Rutgers University, Piscataway, NJ
Attendance: 76
Organizers:
Dennis Pearl, Ohio State
Tom White, Roche Molecular Systems

Mini-Workshop on Geometrical Methods for Conformational Modeling
August 2 - 3, 1995
DIMACS Center, CoRE Building, Rutgers University, Piscataway, NJ
Attendance: 43
Organizers:
Leonidas J. Guibas, Stanford University
Tim Havel, Harvard University

Algorithm Implementation Challenge Workshop
September 14 - 15, 1995
DIMACS Center, CoRE Building, Rutgers University, Piscataway, NJ
Attendance: 55
Advisory Committee
Ellson Chen, Applied Biosystems
Sorin Istrail, Sandia National Laboratory
David Johnson, AT&T Bell Laboratories
John Kececioglu, UC Davis
Joachim Messing, Rutgers University
Joseph Nadeau, Jackson Laboratory
Pavel Pevzner, Pennsylvania State University
Peter Rice, Sanger Center
Martin Vingron, German National Research Center for Computer Science (Coordinator)
Michael Waterman, University of Southern California

Mini-Workshop on the Mathematics of Drug Discovery
September 21 - 22, 1995
DIMACS Center, CoRE Building, Rutgers University, Piscataway, NJ
Attendance: 75
Organizers:
Tim Havel, Harvard Medical School
Irwin Kuntz, University of California - San Francisco

Mini-Workshop on Gene Finding & Gene Prediction
October 13 - 14, 1995
Genome Center at the University of Pennsylvania, Philadelphia, PA
Attendance: 100

Organizer:
David Searls, Univ. of Pennsylvania

2nd Sandia National Laboratories Workshop on Computational Molecular Biology
March 4 - 6, 1996
Sandia Labs, Albuquerque, NM
Attendance: 135
Organizer:
Sorin Istrail, Sandia Labs

Workshop on Computational Biology as part of the 50th Anniversary for ENIAC
May 17 - 19, 1996
Princeton University, Princeton, NJ
Attendance: 90
Organizers:
Craig Benham, Mt. Sinai School of Medicine
Sue Davidson , University of Pennsylvania
Warren Ewens, University of Pennsylvania
Martin Farach, Rutgers University
Sampath Kannan, University of Pennsylvania
Tandy Warnow (Chair), University of Pennsylvania

2nd Annual Workshop on DNA Based Computers
Princeton University, Princeton, NJ
June 10 - 12, 1996
Attendance: 110
Program Committee:
Eric Baum, NEC
Dan Boneh, Princeton University
Peter Kaplan, Princeton University
Richard Lipton, Princeton University
John Reif, Duke University
Nadrian Seeman, New York University

Workshop on Mathematical Hierarchies and Biology
November 13 - 15, 1996
DIMACS Center, CoRE Building, Rutgers University, Piscataway, NJ
Attendance: 75
Organizer:
F.R. McMorris, University of Louisville
Boris Mirkin, Rutgers University
Fred Roberts, Rutgers University
Andrey Rzhetsky, Pennsylvania State University

Workshop on Computational and Physical Aspects of Protein Folding and Related Problems
December 15 - 17, 1996
Hill Center, Rutgers University, Piscataway, NJ
Attendance: 120
Organizers:
Joel L. Lebowitz, Rutgers University
Alexander Grosberg, MIT
Mona Singh, Princeton University
Scott Decatur, Rutgers University

DIMACS/PMMB/MBBC Workshop on DNA Topology II
April 3- 4, 1997
DIMACS Center, CoRE Building, Rutgers University, Piscataway, NJ

Attendance: 35
Organizers:
Wilma Olson (Chair), Rutgers University
Bernard Coleman, Rutgers University
Tamar Schlick, NYU
Irwin Tobias, Rutgers University

Third International Workshop on DNA Computation
June 23-25, 1997
University of Pennsylvania, Philadelphia, PA
Attendance: 123
Organizers:
Harvey Rubin (Chair), University of Pennsylvania
Len Adleman, University of Southern California
Anne Condon, University of Wisconsin
Richard Lipton, Princeton University
David Wood, University of Delaware
Carter Bancroft, Mount Sinai School of Medicine

Joint DIMACS/Universidad de Buenos Aires Tutorial on Computational Biology
Date: July 21-26, 1997
University of Buenos Aires, Buenos Aires, Argentina
Attendance: 73
Organizers:
Dannie Durand, University of Pennsylvania
Martin Farach, Rutgers University
R. Ravi, Carnegie Mellon University
Mona Singh, Princeton University

mmCIF* Software Developers' Workshop
October 17-19, 1997
DIMACS Center, CoRE Building, Rutgers University, Piscataway, NJ
* mmCFI stands for macromolecular Crystallographic Information File
Attendance: 50
Organizers:
Helen M. Berman, Rutgers University
Paula Fitzgerald, Merck Research Laboratories
John Westbrook, Rutgers University

Minisymposium on Advances in Phylogenetics Inference Methods
March 20, 1998
DIMACS Center, CoRE Building, Rutgers University, Piscataway, NJ
Attendance: 25
Organizers:
Daniel Huson, Princeton University
Shibu Yooseph, Rutgers University

Symposium on Estimating Large Scale Phylogenies: Biological, Statistical, and Computational Problems
June 26 - 28, 1998
Princeton University, Princeton, NJ
Attendance: 63
Organizers:
Tandy Warnow, University of Pennsylvania
Junhyong Kim, Yale University
Ken Rice, SmithKline Beecham

Workshop on Combinatorial Clustering and Multi-Domain Protein Structure Analysis
June 26 - 27, 1998
DIMACS Center, CoRE Building, Rutgers University, Piscataway, NJ
Attendance: 52
Organizers:
Sylvia Spengler, Lawrence Berkeley National Lab
Manfred D. Zorn, Lawrence Berkeley National Lab
Inna Dubchak, Lawrence Berkeley National Lab
Fred Roberts, Rutgers University
Ilya Muchnik, Rutgers University
Casimir A. Kulikowski, Rutgers University

9th Annual Conference on Combinatorial Pattern Matching
July 20 - 21, 1998
Busch Campus Student Center, Multipurpose Room, Rutgers University, Piscataway, NJ
Attendance: 60
Organizers:
Martin Farach, Rutgers University
Udi Manber, University of Arizona

Workshop on Advances and Opportunities at the Biology/Math/Computation/Physical Sciences Interface
March 6 - 7, 1999
DIMACS Center, CoRE Building, Rutgers University, Piscataway, NJ
Attendance: 149
Principal Organizers:
Wilma Olson, Rutgers University
Fred S. Roberts, Rutgers University

Workshop on Protein Structure and Function
November 19, 1999
DIMACS Center, CoRE Building, Rutgers University, Piscataway, NJ
Attendance: 36
Organizers:
Israel M. Gelfand, Rutgers University

Workshop on Discrete Mathematical Problems and Medical Applications
December 8 - 10, 1999
DIMACS Center, CoRE Building, Rutgers University, Piscataway, NJ
Attendance: 35
Organizers:
Ding-Zhu Du, University of Minnesota
Panos M. Pardalos, University of Florida
Jie Wang, University of North Carolina

June 2001