DIMACS Center
Rutgers University

DIMACS Exploratory Postdoctoral Program in Computational Epidemiology

Final Report

June 2005

Ia. Participants in the program

Senior Leadership:

PI: Fred Roberts, DIMACS
Organized the project and collaborations with the postdocs.

Postdocs:

Graham Cormode

Patrick DeLeenheer

Mentors:

Simon Levin, Princeton University

S. Muthukrishnan, Rutgers University

Eduardo Sontag, Rutgers University

Ib. Participating Organizations

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

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

NEC Laboratories America: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning.

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

Princeton University: Collaborative Research
Partner organization of DIMACS. Individuals from the organization participated in the program planning, and mentoring.

Avaya Labs: Collaborative Research

Partner organization of DIMACS. Individuals from the organization participated in the program planning.

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

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

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

1c. Other Collaborators

Simon Levin, Princeton University

II. Project Activities

Mathematical methods have become important tools in analyzing the spread and control of infectious diseases and in the analysis of alternative interventions in the case of newly emerging diseases, naturally occurring epidemics of traditional diseases, and deliberate attacks using biological weapons. Much is to be gained by extending the mathematical tradition by making use of the powerful tools of modern computer science. This project was based on the belief that partnerships among computer scientists, mathematical scientists, biologists, and epidemiologists can lead to important new contributions to the usefulness of such methods, and in particular on the belief that opening lines of communication between computer scientists and mathematicians interested in problems of epidemiology but coming at them from completely different perspectives can stimulate important new developments.

In this project, we supported the work of two postdoctoral fellows at DIMACS, the Center for Discrete Mathematics and Theoretical Computer Science. One, Patrick De Leenheer, comes from a mathematical tradition, and the other, Graham Cormode, from a computer science tradition, and both are interested in epidemiological applications. We involved them in the activities centered around the DIMACS Special Focus on Computational Epidemiology (http://dimacs.rutgers.edu/SpecialYears/2002_Epid/). They each were given the opportunity to work on problems they are interested in, in conjunction with developing research groups at DIMACS. De Leenheer worked with Professors Eduardo Sontag at Rutgers University and Simon Levin at Princeton University; he worked on control theory models of virus dynamics and chemostat models with Sontag and worked with the flu modeling group organized by Levin. Cormode worked with Professors S. Muthukrishnan and Fred Roberts at Rutgers and in particular worked on applying powerful methods of modern data mining to the developing DIMACS efforts in data cleaning of public health data, biosurveillance, and bioterrorism sensor location. However, we also encouraged the interaction of De Leenheer and Cormode with other

DIMACS researchers and visitors. Each postdoc also explored work related to the other's special interests, with De Leenheer considering ways to use control theory in biosurveillance and Cormode considering ways to use methods of data mining in flu modeling.

The postdocs participated in a variety of workshops that allowed them to present their work to a broad community. They also acted as mentors to undergraduates doing research.

III. Project Findings

Chemostat Models Applied to a Single Species

Patrick De Leenheer initiated several collaborations during his stay at DIMACS. All of them involved his post-doc advisor, Eduardo Sontag from Rutgers. The first result was obtained from joint work with senior project organizer Simon Levin from Princeton and Christopher Klausmeier from Georgia Tech. They considered chemostat models (a chemostat is a biological reactor in which one or more species compete for one or more nutrients) and studied the dynamics of a single species, limited by two nutrients, assuming that nutrient uptake and growth are decoupled. For a broad class of uptake and growth functions they showed that a nontrivial equilibrium might exist. Moreover, if it exists it is unique and globally stable, generalizing a previous result by Legović and Cruzado. The resulting publication is listed in the Papers section of this report.

Chemostat Models Applied to Multiple Species

A second result on chemostat models followed from collaboration of Patrick De Leenheer with David Angeli from the University of Firenze, Italy, visiting at DIMACS, and Eduardo Sontag from Rutgers. They considered again a chemostat model, but this time an arbitrary number of species was assumed to compete for a single nutrient. Also, here they idealized the conversion of nutrients into new biomass, in contrast to the previously described work. A distinct feature of this model however -one that sets it apart from the traditional chemostat models- is that crowding effects are taken into consideration. The model can be rewritten as a negative feedback interconnection of two systems that are monotone (as input-output systems). Moreover, these subsystems behave nicely when subject to constant inputs. This allows the use of a particular small-gain theorem that has recently been developed for feedback interconnections of monotone systems. Both global stability and coexistence results have been obtained (the latter holds if crowding effects are large enough). Their coexistence result should be compared to the classical chemostat theory where coexistence does not occur. The resulting publication is listed in the Papers section of this report.

Monotone Systems

Patrick De Leenheer, David Angeli, University of Firenze, Italy, and Eduardo Sontag, Rutgers, were invited to write a tutorial paper for the 16th Mathematical Theory of Networks and Systems conference, held in Leuven, Belgium, Summer, 2004. This publication is listed in the Papers section of this report. They reviewed some of the main features of monotone systems, highlighting some application areas (mainly in biology, chemistry). A convergence result for solutions of monotone systems was given with a very short and self-contained proof. Ordinarily such proofs are rather lengthy and require knowledge of a number of fundamental results from the theory of monotone systems. As an immediate application they considered the behavior of a general class of chemical reaction networks with monotone but otherwise arbitrary reaction kinetics. The main result is that every solution converges to a steady state. In addition they showed that the inclusion of diffusion effects in the models does not alter this conclusion.

Predator-Prey Systems

Patrick De Leenheer, David Angeli, University of Firenze, Italy, and Eduardo Sontag, Rutgers, obtained an almost global convergence result for Lotka-Volterra systems with predator-prey interactions. Since the early work of Lotka and Volterra, predator-prey systems have continued to attract significant attention. It is well known that these systems may exhibit oscillatory behavior; the best known is the classic Lotka-Volterra predator-prey system. These systems can be written as (negative) feedback systems. The subsystems of the feedback loop are monotone control systems, possessing particular input-output properties. De Leenheer, Angeli, and Sontag used a small-gain theorem, adapted to a context of systems with multiple equilibrium points, to obtain the desired almost global convergence result, which provides sufficient conditions to rule out oscillatory or more complicated behavior that is often observed in predator-prey systems.

Sequence Information Searching

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

Interaction Network Monitoring.

Graham Cormode and Muthu Muthukrishnan, Rutgers, studied the problem of monitoring massive sequences of interactions (person-person or animal-animal, etc.). The challenge of monitoring massive amounts of data generated by interaction networks has led to the interest in data stream processing. They studied streams of edges in massive interaction multigraphs, defined by (source, destination) pairs. The goal is to compute properties of the underlying graph while using small space (much smaller than the number of participants), and to avoid bias introduced because some edges (interactions) may appear many times, while others are seen only once. They gave results for three fundamental problems on multigraph degree sequences: estimating frequency moments of degrees, finding the heavy hitter degrees, and computing range sums of degree values. Space bounds for the summarizing algorithms are significantly smaller than storing complete information. In an experimental study, such summaries are seen to be highly effective, enabling massive multigraph streams to be effectively summarized to answer queries of interest with high accuracy using only a small amount of space.

Sensor Networks

Graham Cormode, Martin Farach-Colton, Rohan Fernandes, Miguel Mosteiro and Muthu Muthukrishnan, all from Rutgers, worked on a variety of mathematical optimization problems relating to the placement and organization of sensor networks with particular application to tracking the spread and transmission of disease and other epidemic-like phenomena. They have results bounding the time of communication of events in these networks, and have shown how to compute maximum likelihood estimators for readings with errors between multiple sensors. (Farach-Colton and Muthukrishnan were senior faculty in the project, Fernandes a project graduate student and Mosteiro a graduate student in Farach-Colton's group.)

Sensor networks, that is, networks of very inexpensive sensors distributed at random, are likely to become an important tool for epidemiological data gathering, for example when the sensors detect bioactive agents. The cheaper the sensors, the easier they will be to deploy. But cheap sensors have many limitations -- low transmission power, no global positioning systems, low memory. Thus, they don't know where they are when they get thrown out of an airplane, and they can only talk to a few of their nearby neighbors.

Cormode, Farach-Colton, Fernandes, Mosteiro and Muthukrishnan have studied how to find highly connected low-degree subnetworks of sensor networks in sensors randomly distributed in 2-space. They have preliminary results that show that such networks must exist, but their algorithm for constructing them relies on sensors knowing where they are.

The second thrust of their work has been in algorithms for having sensors compute their position from knowledge of the approximate distance to their close neighbors. They have shown that sensor networks, whose distance measurements form distance threshold graphs (that is, you know the distances to nearby neighbors but have no information of distances to distant neighbors) have interesting rigidity properties in two dimensions, which means that the problem of finding the position of each sensor given noisy distance measurements is highly constrained.

Sensor Networks -- Efficient Monitoring

Graham Cormode, in conjunction with Minos Garofalakis and Rajeev Rastogi of Bell Labs, and S. Muthukrishnan of Rutgers, studied the problem of effectively conducting distributed monitoring tasks within a large sensor network. Emerging large-scale monitoring applications require continuous tracking of complex aggregates and data-distribution summaries over collections of physically-distributed streams. Thus, effective solutions have to be simultaneously space efficient (at each remote site), communication efficient (across the underlying communication network), and provide continuous, guaranteed-quality estimates. They proposed novel algorithmic solutions for the problem of continuously tracking complex holistic aggregates in such a distributed-streams setting -- the primary focus is on approximate quantile summaries, but the approach is more broadly applicable and can handle other holistic-aggregate functions (e.g., “heavy-hitters” queries). They presented the first known distributed-tracking schemes for maintaining accurate quantile estimates with provable approximation guarantees, while simultaneously optimizing the storage space at each remote site as well as the communication cost across the network. The algorithms employ a combination of local tracking at remote sites and simple prediction models for local site behavior in order to produce highly communication- and space-efficient solutions. They performed extensive experiments with real and synthetic data to explore the various tradeoffs and understand the role of prediction models in their schemes. The results clearly validated the approach, revealing significant savings over naive solutions as well as the analytical worst-case guarantees.

Constructing the First Known Data Stream Algorithm for Estimating Dominance of Multiple Signals.

Graham Cormode and S. Muthukrishnan, Rutgers, considered streams of multiple signals (i,a_i,j) where the i's correspond to the domain, the j's index the different signals and i,a_i,j gives the value of the j'th signal at point i. They addressed the problem of determining the dominance norms over the multiple signals, in particular the max-dominance norm defined as $\sum_i \max_j a_i, j$. Besides finding many applications, such as in estimating the “worst case influence” of multiple processes in IP traffic analysis or electrical grid monitoring, or finding the dominant signal among many from the type of bioterrorism sensors that future technology may provide, or analyzing continuous data feedback human body monitors being explored by Muthukrishnan, this norm is a natural measure: it generalizes the notion of union of data streams and may be alternately thought of as estimating the L1 norm of the upper envelope of multiple signals. Cormode and Muthukrishnan constructed the first known data stream algorithm for estimating max-dominance of multiple signals. The algorithm is simple and implementable; its analysis relies on using properties of stable random distributions with small parameter alpha, which may be a technique of independent interest. They also showed that other dominance norms -- min-dominance ($sum_i min_j a_i,j$), count-dominance ($|i|a_i>b_i|$) or relative-dominance ($sum_i a_i/max 1,b_i$) -- are all impossible to estimate accurately with sublinear space.

Cluster Analysis

Graham Cormode collaborated with James Abello, Dmitriy Fradkin, David Madigan, Ofer Melnik, and Ilya Muchnik, all of Rutgers, on applying data mining techniques to epidemiology. Cormode specialized in the area of cluster analysis. The notion of ‘clusters’ is a very natural one, and occurs frequently in discussions of epidemiology. We hear about ‘cancer clusters’, areas where the number of reported cancer cases within an area or group of people exceeds the expected amount. Such clusters lead to investigation of possible carcinogens or explanation for greater susceptibility amongst certain groups. Cluster Analysis is an unsupervised learning technique that takes large collections of data points and attempts to identify clusters of similar points. More formally, it tries to create clusters to optimize various mathematical properties, such as minimizing the maximum spread of each cluster, or minimizing the sum of the spreads. A variety of algorithms have been proposed to create clusters from a data set, including k-means, hierarchical clustering, and expectation maximization. Cormode developed several new applications of cluster analysis and provided a survey article on this technique, which is listed in Papers.

IV. Project Training/Development

This project centered around training two postdoctoral fellows who had relevant scientific interests but no major prior exposure to epidemiological applications.

V. Outreach Activities

VI. Papers/Books/Internet

Books

Abello, J., and Cormode, G. (eds.), *DIMACS Computational Methods in Epidemiology, AMS-DIMACS Volume Series*, American Mathematical Society, in press, 2005.

Papers

Abello, J., Cormode, G., Fradkin, D., Madigan, D., Melnik, O., and Muchnik, I., “Selected data mining concepts,” *DIMACS Computational Methods in Epidemiology, AMS-DIMACS Volume Series*, Abello, J., and Cormode, G. (eds.), American Mathematical Society, in press, 2005.

Angeli, D., De Leenheer, P., and Sontag, E.D., “A small-gain theorem for almost global convergence of monotone systems,” *Systems and Control Letters*, **52** (2004), 407-414.

Angeli, D., De Leenheer, P., and Sontag, E.D., “Monotonicity and convergence in chemical reaction networks,” *Proceedings of 43rd IEEE Conference on Decision and Control*, Bahamas, 2004.

Cormode, G., Garofalakis, M., and Muthukrishnan, S., “Holistic aggregates in a networked world: Distributed tracking of approximate quantiles,” *Proceedings of ACM SIGMOD International Conference on Management of Data,* 2005.

Cormode, G., Korn, F., Muthukrishnan, S., and Srivastava, D., “Diamond in the rough: Finding hierarchical heavy hitters in multi-dimensional data,” *Proceedings of SIGMOD*, (2004), 155-166.

Cormode, G., Korn, F., Muthukrishnan, S., Johnson, T., Spatscheck, O., and Srivastava, D., “Holistic udafs at streaming speeds,” *Proceedings of SIGMOD*, (2004), 35-46.

Cormode, G., Korn, F., Muthukrishnan, S., and Srivastava, D., “Effective computation of biased quantiles over data streams,” *Proceedings of 21st International Conference on Data Engineering (ICDE)*, 2005.

Cormode, G. and Muthukrishnan, S., "Space efficient mining of multigraph streams," *Proceedings of ACM Principles of Database Systems*, 2005.

Cormode, G. and Muthukrishnan, S., "Summarizing and mining skewed data streams," *Proceedings of SIAM Conference on Data Mining*, 2005.

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

De Leenheer, P., Angeli, D., and Sontag, E.D., "Crowding effects promote coexistence in the chemostat," DIMACS Technical Report 2003-44, 2003.

De Leenheer, P., Angeli, D., and Sontag, E.D., "A tutorial on monotone systems - with an application to chemical reaction networks," *Proceedings of the 16th International Symposium on Mathematical Theory of Networks and Systems (MTNS2004),* July 5-9, 2004, Leuven, Belgium, to appear.

De Leenheer, P., Angeli, D., and Sontag, E.D., "A tutorial on monotone systems - with an application to chemical reaction networks," DIMACS Technical Report 2004-15, 2004.

De Leenheer, P., Angeli, D., and Sontag, E.D., "Monotone chemical reaction networks," DIMACS Technical Report 2004-16, 2004.

De Leenheer, P., Angeli, D., and Sontag, E.D., "On predator-prey systems and small gain theorems," *Mathematical Biosciences and Engineering*, **2** (2005), 25-42.

De Leenheer, P., Angeli, D., and Sontag, E.D., "Monotone chemical reaction networks," *Journal of Mathematical Chemistry*, to appear.

De Leenheer, P., Angeli, D., and Sontag, E.D., "Crowding effects promote coexistence in the chemostat," *Journal of Mathematical Analysis and Applications*, submitted.

De Leenheer, P., Levin, S.A., Sontag, E.D, and Klausmeier, C.A., "Global stability in a chemostat with multiple nutrients," *Journal of Mathematical Biology*, submitted.

De Leenheer, P., Levin, S.A., Sontag, E.D, and Klausmeier, C.A., "Global stability in a chemostat with multiple nutrients," DIMACS Technical Report 2003-40, 2003.

De Leenheer, P., Li,, B., and Smith, H.L., "Competition in the chemostat: Some remarks," *Canadian Applied Mathematics Quarterly*, to appear.

De Leenheer, P., and Malisoff, M., "A small-gain theorem for monotone systems with multi-valued input-state characteristics," submitted.

De Leenheer, P., and Sontag, E.D., "A note on the monotonicity of matrix Riccati equations," DIMACS Tech Report 2004-36, 2004.

De Leenheer, P., and Sontag, E.D., "A note on the monotonicity of matrix Riccati equations," *IEEE Transactions on Automatic Control*, to be submitted

Talks

Cormode, G., "Zeroing in on the L0 metric," DIMACS Workshop on Discrete Metric Spaces and Their Algorithmic Applications, Princeton University, August 20 - 23, 2003.

Cormode, G., "Simple fast data stream summaries with count-min sketches," DIMACS Mixer Series, AT&T Labs - Shannon Laboratory, Florham Park, NJ, September 25, 2003.

Cormode, G., "Algorithmic embeddings for comparing large text streams," DIMACS Theoretical Computer Science Seminar, Rutgers University, September 29, 2003.

Cormode, G., "How hard are computer games?," DIMACS Theoretical Computer Science Seminar, Rutgers University, February 16, 2004.

Cormode, G., Garofalakis, M., and Muthukrishnan, S., "Holistic aggregates in a networked world: Distributed tracking of approximate quantiles," ACM SIGMOD International Conference on Management of Data, 2005.

Cormode, G., Korn, F., Muthukrishnan, S., and Srivastava, D., "Diamond in the rough: Finding hierarchical heavy hitters in multi-dimensional data," SIGMOD, 2004.

Cormode, G., Korn, F., Muthukrishnan, S., Johnson, T., Spatscheck, O., and Srivastava, D., "Holistic udafs at streaming speeds," SIGMOD, 2004.

Cormode, G., Korn, F., Muthukrishnan, S., and Srivastava, D., "Effective computation of biased quantiles over data streams," 21st International Conference on Data Engineering (ICDE), 2005.

Cormode, G., and Muthukrishnan, S., "High speed data analysis at MassDAL," Poster presentation, DIMACS / Rutgers CS State of New Jersey Symposium on Homeland Security Research, Rutgers University, October 29, 2003.

Cormode, G. and Muthukrishnan, S., "Space efficient mining of multigraph streams," ACM Principles of Database Systems, 2005.

Cormode, G. and Muthukrishnan, S., "Summarizing and mining skewed data streams," SIAM Conference on Data Mining, 2005.

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

Leenheer, P., “Crowding effects promote coexistence in the chemostat,” DIMACS Mixer Series, Princeton University, December 4, 2003.

Leenheer, P., “Crowding effects promote coexistence in the chemostat,” Department of Mathematics Colloquium, University of Florida, January 30, 2004.

Leenheer, P., “Crowding effects promote coexistence in the chemostat,” Department of Mathematics Colloquium, University of Iowa, February 18, 2004.

Leenheer, P., “Crowding effects promote coexistence in the chemostat,” Department of Mathematics Colloquium, Queen’s University, February 23, 2004.

Leenheer, P., “Growth on 2 nutrients in the chemostate: An application of monotone systems theory,” PACM Dynamical Systems/Nonlinear Science Seminar, Princeton University, April 2, 2004.

Leenheer, P., “An introduction to monotone systems,” Applied Mathematics Seminar, University of Florida, October 26, 2004.

Reports

DIMACS Working Group Report on Data Mining and Epidemiology, March 18-19, 2004
Report Authors: James Abello and Graham Cormode, DIMACS
Available as DIMACS Technical Report 2004-37, 2004.
http://dimacs.rutgers.edu/Workshops/WGDataMining/material/ (password protected, login name: wgepid, password: mining)

VII. Other Products

VIII. Contributions within Discipline

Graham Cormode and Patrick De Leenheer, in association with members of the BioMaps research group at Rutgers, organized the joint DIMACS/BIOMAPS Seminar Series on Quantitative Biology and Epidemiology during the 2003-2004 academic year. In total, there were almost forty talks given in the first year of this joint venture. The audience came from a wide variety of disciplines, including Biology, Mathematics, Medicine, Physics, Chemistry, Computer Science and more, including regular participants from Rutgers, The University of Medicine and Dentistry of New Jersey (UMDNJ), Robert-Wood Johnson Medical School (RWJMS), Princeton and beyond. The program of this series, including titles and speakers, is given here.
http://dimacs.rutgers.edu/SpecialYears/2002_Epid/seminars.html

Graham Cormode served as one of the organizers of the second meeting of the DIMACS working group on Data Mining and Epidemiology, March 18-19, 2004. This meeting brought together researchers from a variety of areas: epidemiologists, statisticians, data miners and others. The aim was to foster greater communications between these disparate groups, to help data mining experts understand concrete epidemiological questions, and to help epidemiologists understand better the results possible from using data mining techniques. Several collaborations have already resulted from interactions at the meeting and after, and publications from talks given at the meeting are in progress. A report of this meeting, detailing some of the questions posed as challenges to the area, is available as DIMACS Technical Report 2004-37, 2004.

Epidemiology is an observational science that concerns itself with finding and explaining patterns of health and disease in populations, usually of humans, but also populations of animals, insects and plants. Data mining is an active area of research interested in finding algorithms for describing latent patterns in often very large data sets. This Working Group on Data Mining and Epidemiology has the objective of fostering collaboration between these two disciplines. In March of 2004 it organized a two-day meeting at DIMACS to bring these two groups together in a format designed to initiate such collaborations.

Each day there were a small number of presentations from epidemiologists, statisticians and computer scientists. They gave a detailed description of a specific study analyzed by methods current in today's epidemiological practice. The studies were selected to represent a range of typical techniques such as exploratory factor analysis, multiple logistic regression, proportional hazards models, generalized additive models and other current methods. The epidemiologists concentrated on the epidemiological aspects of the study, with some relatively brief explanations of the main ideas by the statisticians. Computer scientists or discrete mathematicians commented on algorithmic issues and alternate approaches to the underlying questions. This led into a moderated audience-wide discussion of each study.

The goal was to identify some fruitful areas for applying new techniques from theoretical computer science, discrete mathematics, and statistics and to allow new collaborations among participants to be initiated. This working group brought together with epidemiologists experts interested in exploring data mining and algorithmic techniques potentially useful in epidemiology from a range of viewpoints. In order to facilitate a synthesis of perspectives, the working group was organized around half-day sessions, each with one or two talks followed by active discussion. The five primary topics of inquiry were:

Disease Dynamics, Epidemiological Studies and the Interpretation of Data
Genetic, Pharmaceutical and Environmental Epidemiology
Elements of Statistical Learning and Models
Regression Methods, Support Vector Machines and Cluster Analysis
Association Rules and Lattice/Graph Theoretical Methods

Through short presentations and discussions that included all participants, the workshop aimed to foster research cooperation among Epidemiologists, Computer Scientists and Mathematicians. The goal was to identify fundamental epidemiological problems that can benefit from efficient computational, statistical and mathematical models that can aid in the processing and understanding of combined epidemiological, and immunological data. The working group meeting included a problem section and the resulting discussion is documented in the meeting report.

An important outcome of the meeting was the identification of data mining methods appropriate for the early indication of plant diseases. This has led to collaboration between DIMACS and the National Plant Diagnostic Network headquartered at UC Davis.

The working group also resulted in a DIMACS special volume, edited by James Abello and Graham Cormode, with the objective of documenting and disseminating the results of the collaboration between these two disciplines. The volume includes topics explored by the working group, which emphasized the presentation of specific epidemiological problems and collaborative discussion of ways to use data mining methods, informed by modern tools of computer science, statistics, and mathematics, to gain epidemiological insights. Many of the papers included in the volume give detailed descriptions of specific studies analyzed by methods current in today's epidemiological practice. There is a section on open problems from the interface of Epidemiology, Computer Science, Mathematics and Statistics. There are papers that clearly describe fundamental epidemiological problems that can benefit from efficient computational, statistical and mathematical models that can aid in the processing and understanding of combined epidemiological and immunological data. Included are research, expository and survey papers.

IX. Contributions -- other Disciplines

The topics of this project are inherently interdisciplinary. The team of postdocs and mentors involved in the project represents a variety of disciplines. It is anticipated that a major outcome of the project will be the continued enhancement of the lines of communication between mathematical and biological scientists as well as the lines of communication between computer scientists and mathematicians wishing to work on epidemiological problems. See the section on Contributions within Discipline for activities that brought together researchers from a broad variety of disciplines.

XI. Contributions to Resources for Research and Education

XII. Contributions Beyond Science and Engineering

XIII. Contributions To Human Resources Development

Patrick De Leenheer mentored undergraduate Andrew Hodges from Manchester College on his project “Mathematical models incorporating HIV mutation data” as part of the 2004 DIMACS REU program. De Leenheer and Hodges constructed mathematical models for describing the dynamics of the virus, its target cells (CD4+ T-cells), and infected T-cells in the blood stream of an infected individual. They assumed the existence of 2 strains of viruses and associated infected T-cells. By means of local stability analysis and through simulations they determined that the mutation rate has a significant effect on which strain will be dominant. (Mutation occurs often and on a rapid time-scale in HIV.) They found that a strain that may be dominant if there is no mutation might be cleared if the mutation rate is sufficiently high.

Graham Cormode mentored undergraduate Shiri Azenkot from Pomona College on her project "String edit distance with moves" as part of DIMACS' 2004 REU Program. The goal of this project was to see how the string edit distance with moves problem can be applied to genetic sequences. Cormode and Azenkot implemented a new algorithm that approximates the edit distance with moves between two strings (introduced by Graham Cormode and S. Muthukrishnan) and conducted some tests to evaluate the performance of the algorithm in practice. Shiri Azenkot gave a talk on this work at the Seventh Annual Nebraska Conference for Undergraduate Women in Mathematics, University of Nebraska, February 2005. Shiri had the following to say about her research experience. "The REU was an amazing experience. I learned a lot in the specific area in which I researched, and about research in general. It motivated and inspired me to go to graduate school in computer science. The REU was also great fun. I especially enjoyed the trip to Prague--it was an amazing opportunity. Thank you for everything."

Graham Cormode also mentored undergraduate Yinmeng Zhang from Carnegie Mellon University on his project "Zipf analysis of algorithms for approximating quantiles." Cormode and Zhang considered specifically the Count-Min Sketch, a sketch algorithm that answers point queries such as how frequent was a specified element and which elements appeared more than some threshold percent of the time. The algorithm uses space O(1/epsilon log(1/delta)) given no assumptions about the input. However, in practice something is often known about the distribution of the input; a Zipfian distribution is common. A Zipfian distribution looks like 1/q^z, where the frequency of the occurrence of the qth ranked item is proportional to 1/q^z. Note that z=0 gives the uniform distribution. However, when z is large, a few elements account for the majority of the distribution, while most elements have negligible effect. Intuitively, this might allow for savings in space. In fact, for z>1, Cormode and Zhang found the CM sketch only requires O(1/epsilon^z log(1/delta)) space.