Statement of Research

MINGZHOU (JOE) SONG

Knowing what hides behind observations is always fascinating, both spiritually and mundanely. One day on campus, a neuroscience professor asked me, "I have a neuronal signal data set of 43 features. Can you help me compute clusters with a measure of confidence?" During a microarray conference in Toronto, a molecular biologist inquired if I could model the gene regulatory network with his microarray data sets of thousands of cow genes expressed under different conditions. I will not be surprised if more people in a variety of disciplines are interested in using algorithms to gain insights from their particular data sets. Automatic data acquisition instrumentation in various fields is rapidly acquiring large data sets. Researchers can no longer inspect each piece of data individually. In cases where sample size is huge or dimension is high, standard procedures in statistical computing software such as SPlus and the R language become inefficient.

My broad research interest stands at the interface between statistics and computer science. Statistical science determines what should be computed from the data, while computer science pursues efficient algorithms to achieve the computation. My research focus is to develop efficient statistical learning algorithms to effectively compute a mathematical representation of the underlying mechanism from the time series of chaotic and complex data vectors. The underlying mechanism is delineated by three dependencies among the random variables in the mathematical representation – the structural one, the temporal one and the statistical one – which I refer as *three dependencies framework* (3DF). Observed data manifest the mechanism perturbed by environmental noises. Properties of both may evolve over time.

The structural dependency is traditionally studied in the area of functional estimation or regression models; the temporal dependency is tackled in the area of stochastic processes; and the statistical dependency is covered in the area of multivariate analysis. Although each area is already rich, many data sets require one to examine all the dependencies simultaneously for pattern discovery. To make 3DF possible in a balanced way, I build computational models that best explain observed data with minimum complexity. Existing mathematical models handle only special cases in 3DF. For example, a dynamic Bayesian network describes statistical and temporal dependencies among random variables, but it is not expressive enough to reveal possible structural dependencies. Another example is the Markov logic, which combines first-order logic and Markov networks, the former being the structural dependency and the latter being the statistical dependency in the 3DF. Multivariate stochastic processes, on the other hand, subsume temporal and statistical dependencies in the 3DF. The stochastic differential/difference equations can reveal structural and temporal dependencies in 3DF. However, none of these existing models encompasses all three dependencies. 3DF is ambitious, but in my belief, it will become computationally feasible for some challenging real-world problems in the coming decades. My overall strategy is to embed refined local models in a coarse global description. This is justified by the complexity of the 3DF. It is unaffordable to treat everywhere equally in the world I am trying to model. Delicacy should only be attained at necessary places.

My previous cross-disciplinary graduate training in engineering, computer science and statistics laid out the foundation for me to design and realize computational models in 3DF. I will start with

special cases which have not been resolved well. Then I will generalize my results for more difficult problems. I am intrigued by applications in biomedical areas due to personal experiences and the quantitative research trends in biology and medicine. My attachment to biomedical applications started when I first participated in research as an undergraduate student. At the time, I worked on a single photon emission computer tomography project. Since then, I have made a variety of algorithm research endeavors for biomedical applications, which I will summarize in the following sections.

Past Contributions

During my doctoral study under the supervision of Professor Robert Haralick, I designed a new method to reconstruct an optimal 3-D left ventricle surface model by integrating both the low-level image evidence and the high-level prior shape knowledge. Through a pixel class prediction mechanism, my approach avoided not only the unreliable edge detection or image segmentation problem, but also the pixel correspondence problem. In my experiments, the average projection distance errors of epicardial and endocardial surfaces are 3.2 ± 0.85 mm and 2.6 ± 0.78 mm, respectively. Among different methods reported in the literature, my results were the closest to the 2 mm error standard for clinical practice.

The conception of 3DF was triggered by a piece of my dissertation research – optimal quantization. It is a nonparametric technique to approximate the probability density function by means of many parameters. The space is partitioned into hyper-rectangles determined by jointly quantizing the ranges of each variable to maximize a global quantizer performance measure. The measure is a weighted combination of average log likelihood, entropy and correct classification probability. It ensures that the resulting probability density function representation is compact and efficient. On data generated from not very well separated Gaussian mixture models, optimal quantization yielded results superior to those of the expectation maximization algorithm. Not well separated mixture models cause the expectation maximization algorithm to converge to the true parameters extremely slow. Since it does not rely upon the distributions, optimal quantization performs better in such scenarios.

I did my first industrial research project in computational biology when I worked as a consultant for Informatics Research of Celera Genomics. I improved by 20 fold the computational efficiency of the conserved exon algorithm for comparative gene finding in human and mouse genomes. I accelerated the program using exon length statistics, banded dynamic programming and code optimization. I also designed a Markov model to predict open reading frames.

My first independent academic research project is on spike sorting, which detects neuronal signals from background noises. By optimal quantization, I performed spike sorting under a Bayesian framework. I used the grid density resulted from optimal quantization to represent distribution of the shapes of spike signals. I carried out experiments on simulated and real spike signals. The advantage of my approach is to learn spike shape variation directly from data, while previous methods hardwire shape knowledge into their algorithms. The digitized spike signals were provided by Plexon, a Dallas-based company.

Present Work

Currently, my research projects include: sorting and clustering techniques for neuronal signal analysis, analyzing patterns of repetitive elements in mammalian genomes, mathematical modeling of gene regulatory network from temporal expression data, and microscopic image analysis. I am mentoring two doctoral students on above projects. I also run a journal club to discuss recent advances in computational biology.

I am working on a probability density based data stream clustering approach, which requires only the newly arrived data but not the entire historical data, to be saved in memory. The idea roots on a theorem of density updating and it works naturally with Gaussian mixture models. I implement it through the expectation maximization algorithm and a cluster merging strategy by multivariate statistical tests for equality of covariance and mean. This algorithm is applied to clustering neuronal spikes received in real time. Each neuron typically maps to a unique cluster of spikes. Thus the clusters will identify neuron activities, providing the building blocks for studying brain functions.

I maintain a close collaboration with Dr. Stéphane Boissinot, professor of biology. We are analyzing the statistical patterns exhibited by features such as recombination rate, guanine and cytosine content, and distance to genes in the evolution of long interspersed nuclear elements (LINE1s), a repetitive element in mammalian genomes. LINE1s abundantly exist in non-coding regions of a genome. Biologists believe that an LINE1 may affect expressions of certain genes and that some law governs its evolution. Our initial study has shown that even some simple patterns differ significantly between LINE1 and randomly sampled genomic sequences. We are writing a manuscript to publish our findings in a genomics journal. Evidently, modeling LINE1 falls into the 3DF.

I have recently started to investigate two methods to construct mathematical models for gene regulatory network: stochastic difference equations and the stochastic generalized logical network. Both methods can capture quantitative relationships among genes at a scale reasonable to the resolution of the data. I will study the performance of both methods using gene expression data sets from the yeast cell cycle and the responses to environmental stresses.

A common belief has been that neurons in the adult brain would stop proliferation after birth. However, researchers have found that expression of debrin, a protein sufficient to promote proliferation in neurogenesis, is increased in brain cells of some adults. Researchers use 2-D or 3-D images taken from brain slices to observe cells stained for the expressed debrin. However, counting the number of debrin marked cells in an image is hampered by cell clutters in images. To expedite quality analysis, I am building a statistical pixel appearance model and a statistical cell shape model. Both models are learned from cell examples from images. To detect cells in an image, I find all shapes that best explain the entire image using these two models. This strategy avoids segmentation of the images, and addresses the clutter problem in a natural and unified way. This tool will facilitate the researchers to study the function and deregulation of debrin. Potentially, my software can contribute to diagnosis for severe diseases due to dysfunctional nervous systems such as Alzheimer; it may also assist in the elucidation of the mechanism of diseases such as mood disorder.

Future Directions

Efficient Quantization. Although my 1-D optimal quantization algorithm guarantees the best solution, it runs in quadratic time complexity. The running time is acceptable for some applications. However, it is impractical for data sets in gigabytes. Originally, I use dynamic programming to obtain the partition that best explains the data with the maximum log likelihood but no overfitting. Strategies to improve its computational efficiency depends on the type of data being quantized. For integral data within the range of less than one million, they can be hashed into equal width bins before being quantized. The time complexity would be quadratic on the range of data, not the size of data. This can be performed on 8-16 bit gray scale images with no practical implications. When the range of data is more than one million, I plan to build a binary tree to partition the data until the subsets of data at each node are small enough for optimal quantization. Although each binary split can be done optimally in linear time, I would like to scrutinize the overall optimality to see if a solution can be guaranteed within a certain error range of the optimal one. For some features in human genome, the size of one data set can reach several gigabytes. Naturally, an immediate application is to represent the distributions of individual genomic features. The equal bin width histograms obtainable from UCSC Genome Browser, a routine web resource for genomics, are blind to the data and present unrealistic density estimation for some bins. With optimal quantization, the new histograms will be highly consistent with data.

Modeling Evolution of Repetitive Elements. I will continue working on modeling LINE1 evolution with my biology colleague. As the first step, I will design an algorithm to collectively inspect multiple features related to LINE1. I will use a Bayesian network to represent their statistical dependencies. In the second step, we will explore any structural dependencies among the features. The final goal is to incorporate the evolutionary history of LINE1, i.e., the temporal dependency, to build a complete 3DF model for LINE1s.

Data Stream Clustering. A data stream clustering algorithm differs from a static clustering algorithm in that the former does not store the entire historical data. Space and time complexity concerns result in such a requirement for data stream clustering algorithms. My research focus is on finding sufficient statistics to abstract the historical data. I will also develop strategies to update the sufficient statistics from newly arrived data. The statistics employed by most data stream clustering algorithms include first and second moments, and medians. First and second moments can be updated efficiently, but they are not statistically sufficient in a stream whose data are not Gaussian distributed. My strategy is to maintain a collection of all relevant statistics, which sometimes may include the outliers not fitting into any clusters. A further goal is to detect temporal and structural dependencies in the data stream. Altogether, this approach gives a complete description of the data stream in the 3DF. I will use this work to solve the clustering part of the spike sorting problem.

Modeling Gene Regulatory Networks. The ultimate goal of modeling gene regulatory network is to accurately predict gene expressions given a condition. The challenge is that only snapshots of the system at low resolution time points are available. The inputs, or signal molecules, to the network are usually unknown. My current strategy is to find associations among genes by investigating many experiments simultaneously without knowing the actual inputs. The demonstrated statistical patterns will shed light on how one gene may affect others under different situations. In my future

endeavor, I will combine different data sources such as protein interaction data and gene knock out experiment data as they come out, as well as gene expression data. In the stochastic difference equations (SDCE) model, gene expression levels are continuous over discrete time. SDCE have a direct link with stochastic differential equations (SDE) model and can be based on established SDE results. Working in discrete time, the SDCE will be more convenient to apply. It might also be true that SDCE are easier to obtain than SDE. I will evaluate the effectiveness of certain SDCE models. A technical challenge is to find an efficient procedure to obtain SDCE parameters. As an alternative, I will augment the generalized logical network by adding stochastic components to capture the non-deterministic nature of gene regulatory network. Then I will search the set of stochastic logical equations and estimate the stochastic component from given temporal gene expression data. I will extrapolate these methods to study other biochemical networks such as protein interaction network.

Summary of Overall Goals. An immediate goal, not mentioned previously, is to establish a computing cluster of 64 to 128 nodes to perform large scale modeling tasks. For the next few years, my objective is to abstract a general description for 3DF from the insightful thoughts gained from the experience of ongoing projects. My long term goal is to design and implement efficient and effective modeling algorithms in the most general 3DF. I will keep working on applications of my methodologies in molecular biology and neuroscience. Meanwhile, I will continue seeking other opportunities to apply my research in solving major life science problems.

Statement of Teaching Philosophy

MINGZHOU (JOE) SONG

If you give a man a fish, he will have a single meal.
If you teach him how to fish, he will eat all his life.
- Chinese proverb

As an educator, I have the following teaching goals: to help every student in my class to learn effectively, to be respected for maintaining a high academic standard, and to develop various curricula for different educational objectives of my students.

Practicing Principles in Teaching

In my typical undergraduate classroom, there are about 20 students. I usually observe three types of students: one type always actively asks or answers questions; another one participates but less active; the last type seemingly wanders. The majority belongs to the second group. This suggests that I can impose strong educational impact should I tailor the course towards the average preparation level of the second group and actively seek their participation in class. For the first group, I will introduce them my research and invite those interested to play a role in my lab. I will assign them small pieces of a research project so that they learn how to conduct scientific investigations. In this way, I have found and supervised two undergraduate students for research. I presented the work with one of them in the Joint Statistical Meetings in 2003. The recently awarded Howard Hughes Medical Institute grant allows me to interact with more undergraduate students. For the third group, I will talk to them privately and check whether they have difficulties in either preparation, learning techniques, or any other problems. For the less prepared, I either suggest them to review the prerequisite or to take a more appropriate course.

Master's program students are very diverse in many aspects such as educational goal and background, programming skill, registration status (full/half-time), work experience, communication style and cultural background. This diversity is partially due to the previous information technology industry boom and partially due to the multicultural nature of North America. Facing such a gamut of students, my experience tells me that it is never enough to practice proactive communication. In the research or software project courses, I require students to summarize their project progresses every week and discuss any difficulties in class. My classes are often closed early on by the department chair so that enough students can be registered in other sections. From the teaching evaluation, the majority of students felt they enjoyed and learned from my classes.

Almost all students in my doctoral level course are highly motivated, especially when the course focuses on an exciting field such as computational biology. My major attention is paid to making the material state-of-the-art and to guiding students to carry out projects of their own choice with few restrictions. In this way I help them prepare early on for their dissertation research.

While tailoring my teaching towards different students, I practice certain principles. I believe some elements are *always* crucial in general education such as creativity, critical thinking, active

learning, and academic integrity. Although it is debatable that creativity could ever be taught, it is certainly achievable to foster a learning environment to encourage students to express their original thoughts. I critically analyze student ideas, not simply rejecting unjustifiable ones. I instill critical thinking skills to students by cultivating a challenging attitude towards established authorities such as the instructor and the textbook. Students have better understanding of the materials after they have questioned them. I practice active learning approach inside and outside classrooms. For example, I assigned extra credits to students who actively engaged in online course discussions. To maintain academic integrity is becoming challenging due to the convenient availability of information from the Internet. I develop guidelines to use the Internet as a research tool at the beginning of the class. I regard it the professor's responsibility to detect plagiarized work. Some homework turn-in software can evaluate the originality of a piece of work by Internet search and comparison.

Interdisciplinary and Multidisciplinary Exposure

In many areas including computer science, classroom activities are influenced by outside world. To face challenges from the job market, I expand teaching towards two dimensions: inter- or multi-disciplinary education and communications skills. Outsourcing of jobs from the United States to overseas is a reality. Out of the 140 million jobs in the US Labor Force, 400,000 to 500,000 information-technology-processing jobs have been off-shored in the last few years.[1] Although it is yet a significantly large figure, outsourcing will be dramatically affecting students in engineering fields if the current trends continue. However, no other country is more a leader of a wide spectrum of industries and research areas than the United States. Students educated by taking advantage of this multidisciplinary strength will be much more competitive than most students graduated elsewhere around the world. To meet the immediate needs to exchange information with practitioners in other disciplines, I emphasize both technical and people-oriented communication skills. What I have accomplished at Queens College includes developing a new bioinformatics curriculum, participating in strategic planning for a new Master's program in bioinformatics and serving on the Howard Hughes Medical Institute Undergraduate Education Program. The Howard Hughes program has one focus on developing new undergraduate curricula in bioinformatics. These considerations have led me to teach computer science concepts in scenarios of important inter- or multi-disciplinary fields. The followings are some examples:

Example 1. When I taught the undergraduate Design and Analysis of Algorithm, I assigned a project to represent the human genome sequences with Markov models. This project not only reinforced the concept of priority queue, but also exposed students to areas outside computer science such as genomics and statistics.

Example 2. In a graduate level Software Practicum course, I led my students to design microscopic image analysis software. The software extracts cell structures of interest from microscopic images. This project enabled the students to learn the basic concepts of cell biology.

Example 3. In a graduate level Research Practicum in Bioinformatics course, I first introduced important algorithms in bioinformatics. Then I let students choose their own project topics. Students

[1] Business Week Online, February 23, 2004.

completed projects such as gene expression clustering, motif finding, protein classification, forensic DNA analysis, *etc*.

Teaching in a Diverse Culture

American classrooms are becoming more and more international. Having been an international student and currently teaching in a culturally rich college in New York City, I understand many practical issues in a diverse classroom. The rule that I have obeyed is to respect the diversity of students and listen to their needs. For example, some international students may have never heard of Tic-Tac-Toe, a popular western game, even though the same game exists under a different name in their culture. I have learned to explain the game and to make less cultural assumptions on my students.

Teaching Experience and Interests

Since I joined Queens College in February 2002, I have lectured the following undergraduate courses : Design and Analysis of Algorithms, Data Communications, Introduction to Bioinformatics, and Artificial Intelligence. I have taught the following graduate courses: Introduction to Bioinformatics, Topics in Computational Biology, Artificial Intelligence, Research Practicum (Bioinformatics), and Software Practicum (Microscopic Image Analysis Software). I was a teaching assistant for several electrical and computer engineering courses at University of Washington, including Continuous Time Linear Systems, Introduction to Computer Networks, Digital Signal Processing, and Digital Signals and Filtering.

My background and experience enable me to develop new curricula for the following courses in bioinformatics: Perl Programming in Bioinformatics, Computer Science for Biologists, Object Oriented Databases for Biological Data, Microarray Data Analysis, Algorithm Design in Molecular Biology, Biostatistics, and Biological Signal Analysis. I am also capable of tailoring materials in computer science, electrical engineering, and statistics to the needs of bioinformatics students.