CANDIDATE STATEMENT – CYNTHIA GIBAS

My research is focused on development of computational methods to support genomic biology. From genome comparison to design and interpretation of microarray experiments, the work I do has been of interest to collaborators in areas as diverse as plant systematics, plant genomics, and microbiology. These methods are adaptable to a wide range of problems and have the potential to significantly impact how large-scale experiments are designed and interpreted.

Optimizing hybridization-based assays

DNA oligonucleotide microarrays are widely used in the life sciences to measure gene expression, and in several other common applications. However, there are still significant unresolved issues in microarray design and analysis. Microarrays rely on the tendency of single-stranded polynucleotides to bind strongly to a molecule with a complementary sequence. Microarray probes are designed to uniquely complement each predicted mRNA message that might be detected in the system under study. Arrays can be used to monitor changes in transcription over time, or in response to a particular stimulus from the environment or interaction with another organism. Changes in transcript levels are detected as changes in scanned signal intensity each spot where a target molecule binds to a unique probe.

Many scientists treat the results from microarray experiments as if they are quantitative, and yet, there are many aspects of the experiment that have not been modeled. Complete hybridization of each target molecule to its intended probe is taken for granted, in a milieu where many competing binding scenarios may be equally valid. Theoretically, probe and target molecules are capable of binding not only to each other, but to other partially complementary molecules that compete for binding sites. Probe and target can also form internal structure that can block formation of the DNA duplex, and target molecules in a complex mixture can interact with each other. Naturally occurring sequence mismatches can reduce binding affinity. To interpret microarray data in a truly quantitative way, we must be able to model these effects and separate them out from legitimate differences in transcript level.

Microarrays are now so commonly used that providing new methods and standards for their use impacts broadly across many fields of molecular biology research. Long oligonucleotide microarrays, the class of arrays that we are studying, are increasingly commonly used, and yet it is not even known whether the parameters used to model hybridization reactions are valid for sequences over 50 nucleotides in length. Effects of mismatches and competition for binding on long oligonucleotide arrays have not been systematically modeled.

My research group has begun work on the development of automated methods for probe and target selection and microarray data analysis that will take into account the biophysical properties of molecules involved in the array experiment. An R01 proposal to support this work has received a promising score from an NIH study section and was also recommended for funding at NSF. My first Ph.D. student, Vladyslava Ratushna, has completed a preliminary study of internal structure formation in RNA and cDNA target mixtures (Ratushna et al 2004, submitted), modeling the folding behavior of a group of target transcripts from Brucella suis. Our findings suggest that even at relatively high temperatures and in mixtures that have been sheared into smaller fragments, a significant portion of the target molecule is occupied in stable secondary structure, which can interfere with hybridization. Empirical tests are now required to determine whether placing probes that bind to these regions of stable secondary structure will interfere with hybridization of the target to the array. In order to conduct that study, I have teamed with a collaborator (Dr. Jennifer Weller, GMU) who has access to an established custom array facility and appropriate laboratory resources. This study will serve as a paradigm for future modeling and empirical validation. As we determine whether our biophysical models are predictive, we will develop software that incorporates those models as criteria for probe and target selection.

In another collaboration with a group at Penn State (Claude dePamphilis and the Floral Genome Project) we have designed a 60mer array for California poppy, including a large number of probe pairs that will allow us to observe the effect of including varying degrees of intentional mismatch and consensus overlap on detected hybridization.

On the same general research theme, I have recently submitted a proposal to NSF (Division of Biological Infrastructure) to research, and develop a data model for, inclusion of biophysical information in microarray data reporting standards. This project will complement the R01 described above.

Genome Comparison and Genome-based Diagnostics

Genome comparison is vital for many classes of scientific problems, from detecting possible virulence determinants in pathogenic bacteria, to tracking genomic rearrangements and understanding genome evolution. It is also a necessary part of the experimental design process for the new high-throughput experiments. A genome comparison can be used to identify genes, parts of genes, and sequence fragments that distinguish one species from another, and is a first line of approach in developing DNA-based diagnostic tests.

Despite the utility of comparative approaches in genomics, and the progress that has been made in this area, there is still a need for new methods for mining and querying the results of comparative analysis. Most of the current research in this area is either based only on comparison of coding sequence content (e.g. NCBI’s COG), or focused on algorithms for rapid alignment of complete genome sequences (e.g. MumMER, AVID). However, the former approach ignores interesting data on several scales, from variable regions within gene families to the importance of feature order, and the latter is only valid for certain classes of closely related genomes. Synteny – conservation of gene order among closely related genomes – breaks down as more distantly related genomes are compared, but in such situations there are still questions to be answered about local organization within the genome, definition and detection of common and unique subsets, and identification of diagnostic target regions within feature boundaries.

My first graduate student, David Sturgill (M.S. 2003), prototyped a comparative genomics software application called Genomosaic (Gibas et al., 2003). In Genomosaic, we leave behind the simple models of the genome as a long sequence or a string of genes, instead representing genomes as a complex mosaic of overlapping and connected features defined by different means. We had surveyed existing comparative genomic methods, and discovered that a biologist would have to work with diverse data from several kinds of sequence analysis in order to obtain an answer to a practical question – i.e., “where are the features that can be used as diagnostics?”

We developed a data model to contain the output from several different types of genome analysis, and bring them together to identify useful features for diagnostics and comparative studies. We have used the Genomosaic prototype in a true three-way comparison of Brucella genomes, identifying a group of diagnostic targets. These have since been validated in the lab (Sturgill et al, 2004, submitted) and have been included on a Brucella expression microarray developed in collaboration with researchers at the Virginia-Maryland Regional College of Veterinary Medicine (Drs. Stephen Boyle and Nammalwar Sriranganathan) and at USDA (Dr. Shirley Halling). The Genomosaic prototype, as well as a prototype visualization tool (Kaluzska and Gibas, 2004, in press), will be used as the basis for the bioinformatics component of a multi-investigator proposal to sequence several chloroplast genomes, with PI Khidir Hilu (Virginia Tech). I will also continue to pursue funding independent of that project to support further development of Genomosaic.

Teaching Interests

Teaching is an important component of any academic program, and I have put serious effort into teaching activities in my previous appointments. At University of Illinois as a graduate student, my first substantive teaching experience (1993-4) was to develop and teach lab exercises in a newly created undergraduate course – “The Physical Basis of Life”. At Virginia Tech I have developed my own sequence of courses in Bioinformatics Methods, and published an introductory bioinformatics textbook, *Developing Bioinformatics Computer Skills* (O’Reilly & Assoc. 2001), which has been widely adopted and translated into several languages in its first edition.

The Bioinformatics Methods course series is designed to train life science students to properly use common bioinformatics tools. When I first arrived at Virginia Tech, Bioinformatics was a newly-defined discipline and graduate programs in this area were just beginning to emerge nationwide. There was little consensus on what should be taught to beginning students in Bioinformatics. I set out to develop a practical course for upper level undergraduates and beginning graduate students in the Biology Department. The course has both a lecture and a laboratory component, and the laboratory exercises are designed to provide the students with an experience of using bioinformatics tools in scientific inquiry.

A course of this kind, which introduces standard bioinformatics methods and their application and relevance in the life sciences, is a useful introduction both for life science students and for quantitative scientists, who often focus on the theoretical side of bioinformatics method development without much connection to the application side. I could easily teach or co-teach such a course in a different institutional setting. My training is in Biophysics and Computational Biology, and I could also develop or co-teach courses in Molecular Biophysics, Molecular Biology, Protein and Nucleic Acid Structure, or related topics, at undergraduate or graduate level. I look forward to developing a more specialized graduate course in the future. My recent research has focused on genome comparison and genome evolution, and I have given some thought to developing a course in molecular and genomic evolution and phylogenetic analysis methods. An advanced topics course in design and analysis of DNA microarray experiments is another possibility.

What I most enjoy about teaching, and about training graduate students is working with them one-on-one to develop problem-solving strategies. In past semesters, I have graded students using both standard multiple choice and short-answer questions to test their understanding of basic concepts. However, I also base a substantial part of the students’ grade on laboratory exercises or projects, evaluating not only their ability to find the right answer but their ability to develop and remember a process of getting to an answer using bioinformatics tools.

Graduate Program Development

At Virginia Tech, I have had extensive involvement in Bioinformatics teaching and graduate program development. I’ve served on an ad-hoc committee charged with development of a new graduate program curriculum (the Genetics, Bioinformatics, and Computational Biology Program. This program received approval from the State Council of Higher Education in 2003) and I have subsequently served on the GBCB Program steering committee. As this program has evolved, I have spent quite a bit of time researching the required elements of graduate training in Bioinformatics and Computational Biology and working with faculty from diverse backgrounds to create an appropriate program.