SAURABH SINHA — RESEARCH STATEMENT

INTRODUCTION: My research focuses on the development of algorithms for solving problems in molecular biology, with an emphasis on gene regulation and its evolution. The expression of genes into proteins is regulated by transcription factors that bind to their cognate sites, thereby activating or repressing the gene. In metazoans, binding sites are organized into *cis*-regulatory modules, where input from multiple transcription factors is integrated into a common output – the gene's expression pattern. Transcription factors may regulate other transcription factors, forming a network of regulatory interactions that determines cell fate and function. For example, segmentation of the fly embryo is almost exclusively done by a transcriptional network. By unraveling these networks, we can solve the puzzle of how a particular gene is expressed exactly at the right times and places in the organism. Moreover, it has recently emerged that evolution explores novelties in the body plans of animals by tinkering with the regulatory networks, rather than creating new genes. We therefore need to shift our focus from coding sequences to the larger, less understood, non-coding regime. The biological objective of my research is to infer regulatory networks from non-coding sequences and explain how their evolution gives rise to the staggering diversity of organisms.

A vast amount of genomic data is being generated today, creating opportunities for biological discovery on an unprecedented scale. However, inherent redundancies in biological systems, measurement errors, and the magnitude of the data necessitate sensitive and efficient computational tools to extract biological information from the data. My research aims at deducing regulatory interactions by means of algorithms that integrate heterogeneous, incomplete and noisy sources of data in a statistically sound framework. To study how evolution affects gene regulation to create species diversity, we need to compare the deduced regulatory networks across species. The time is ripe for such a full-blown multi-species comparison, with a rapidly growing number of sequenced genomes. My research addresses such evolutionary comparisons with computational tools that are based on realistic models of evolution. The study of regulatory evolution needs close collaboration with experimental biologists, because the cross-species comparison must also happen at the phenotypic level if we are to understand the link "from DNA to diversity".

Current Research

MOTIF FINDING: The simplest unit of the transcriptional network in the sequence is the binding site. The short length and high variability of binding sites obscures their detection in the much longer promoter or enhancer sequences. In my doctoral thesis, I addressed the problem of *ab initio* discovery of binding sites, with algorithms that are guaranteed to find the most statistically overrepresented motifs. We applied one of our algorithms, called "YMF", on functional classes of genes in the yeast *S. cerevisiae*, and reported several novel motifs [7]. A binding site involved in the hypoxia-induced pathway in the bacterium *M. tuberculosis* was discovered by YMF, and verified experimentally by another group. YMF has been downloaded by over 500 users, and its web server has handled 20-30 processing requests every month for the past two years.

MODULE DETECTION: We get a higher resolution picture of regulatory interactions when we discover how binding sites are organized in modules, and this became clear during my post-doctoral experience with the fruitfly genome. The knowledge of canonical binding sites, e.g., from motif-finding tools, makes it possible to computationally search for modules genome-wide. While previous tools searched with simple rules on counts of binding site consensus elements, the *Ahab* algorithm from our group took a probabilistic approach reflecting the energetics of protein-DNA interaction. The algorithm was sensitive to modules with multiple weak sites, placed appropriate non-uniform weights on different transcription factors, and required no parameters to be trained by hand. I developed the *Stubb* algorithm to extend this approach to handle multi-species data [5]. It takes as input a set of *weight matrices* characterizing binding sites of different proteins, and scores each module-length sequence in the genome for clustering of binding sites. The highest scoring sequences are predicted as modules. Stubb combines orthologous module-length sequences into one score, allowing binding sites to occur in conserved (orthologous) and non-conserved regions, in a framework that is consistent with evolution. The binding sites in conserved regions are scored under an assumption of common descent. We evaluated the algorithm on the segmentation pathway in fly [2], and demonstrated that comparison with a second fly genome using Stubb boosts the recovery of modules by 20-50%. Another significant advance in the Stubb algorithm was to model positional correlations between sites, so as to reflect their cooperativity, and this led to improved detection of targets for dimers of Hox proteins and Exd, in the fly genome.

MODELS OF EVOLUTION: During cross-species comparison of biological signals, it is crucial to model their (evolutionary) relationship accurately. We therefore proposed a stochastic model that captures how a binding site can evolve when constrained to bind the same protein. This model was the key to integrating evolutionary comparison

with Stubb's probabilistic approach. Based on the same model, we built a new algorithm (called "PhyME") for finding binding sites in orthologous sequences, with a design principle of accurately reflecting the observed properties of evolution [1]. The algorithm was successful on test data from yeast, fly, and mammals, a versatility of application domains not seen in other similar programs.

EVOLUTION OF REGULATION: In order to clearly demonstrate the evolutionary dynamics of gene regulation, we zoomed in on $\sim$100 modules involved in segmentation of the fly embryo, and computationally screened them for changes in regulatory content between *D. melanogaster* and *D. pseudoobscura*. We cataloged distinct classes of change – severe mutations in specific binding sites, large indels carrying several sites, sites disappearing and reappearing elsewhere in the module, and even entire modules being lost or gained. We recorded the blastoderm expression pattern driven by each of $\sim$16 screened modules (from either species), and correlated the inter-species differences in expression pattern with sequence-level differences highlighted computationally [17]. An interesting finding was that "duplicate" modules with overlapping functions tend to "exchange" some of their functionality, via turnover in their binding site compositions. The two compared species diverged about 30 Myrs ago, their proteins are largely conserved, and the gene expression patterns are also mostly similar. In contrast, we found considerable change in how the regulatory information is encoded at the sequence level. This project involved intensive interplay between bioinformatics and experimental genetics. I had the opportunity to participate in the analysis of embryos, from looking at them under the microscope, to creating tools that automatically extract expression patterns from photographs of embryos and calibrate them. In a separate project, we compared regulatory modules from three species of fly, and carefully tallied the insertions and deletions using maximum parsimony criteria [16]. We found more indels than substitutions (in terms of base-pairs), a fact that needs greater attention in the "conservation filters" used in bioinformatics. We also found insertions to be more common than deletions, and the indels to be enriched in short local duplications (tandem repeats). It is vital that the observed indel patterns inform the tools of evolutionary comparison, and that tandem repeats get incorporated into bioinformatic models of regulatory regions.

Future Directions

My future research will infer how changes in the regulatory network map to changes in biological function and body plans of organisms. I will devise algorithms to chart the history of modules over many species, in terms of binding site composition. The algorithms will adopt a fresh approach to the sequence alignment problem, with scores motivated by evolution rather than by information theory. The most difficult challenge will be to infer which of the computationally detected sequence changes are phenocritical. For this, we will need a better understanding of the mapping from regulatory sequence to gene expression, and extensive experimental work to train the models. Completely new challenges will arise when we study evolution over greater time scales, e.g., between fly and mosquito. A collaboration with another post-doc in our group has shown changes in the DNA-binding domains of a few embryonic factors, that should reflect in their binding sites. Moreover, the non-coding sequences are not easily alignable between the species. We thus have the problem of remote-homology detection for regulatory modules, which we must solve to understand how another long germ-band insect parses its gene expression patterns into modules.

I will investigate regulation in other paradigms in the fly, e.g., cell fating in the nervous system, signal transduction in the eye, wing and other imaginal discs, etc. A typical data set in these paradigms includes microarray data that reveal scores of participating genes, whose spatial patterns are then derived. In these less charted regimes, there are likely to be unidentified transcription factors, that will have to be deduced from the data. Motif-finding algorithms will play a crucial role here, by locating new binding sites that feed into the module detection step.

The vast amount of genomic data being generated today includes complete genomes of multiple species, microarray data, spatial patterns of gene expression, libraries of enhancer traps, mutant analysis data, etc. Integrating these different axes of information into a unified probabilistic framework will be a persistent theme in my research, with emphasis on handling incompleteness of data. For instance, I will extend the multi-species sequence-based model of Stubb to incorporate gene expression information. We are already working on a model that maps the spatial expression patterns of embryonic transcription factors in fly into the spatial patterns of target genes, using the binding site information in the sequence.

My research plan, as outlined above, brings together ideas from computer science and statistics, to solve real world biological problems. In this scientific pursuit, I envision establishing close collaborations with biologists, enabling a fruitful and absolutely crucial exchange between computation and experimental verification.

SAURABH SINHA — TEACHING STATEMENT

I grew up harboring a deep respect for the teaching profession; a respect that was instilled in me since childhood, and that now feeds my aspiration to embrace this most noble of professions. One of the main attractions of a faculty position, to me, is the opportunity to interact with students, and motivate them to learn new ideas. The flow of thoughts between teacher and student is not a one-way street, and I believe that my teaching experience will reward me with fresh ideas and perspectives from my students.

As a teaching assistant in the Computer Science department at the University of Washington, I had the opportunity to teach quiz sections for three different classes - Computer Programming I (CSE 142), Computer Programming II (CSE143), and Machine Organization & Assembly Language (CSE 378). The goal of the quiz sections was to teach in a more interactive fashion than possible in lectures. I gained regular exposure to classroom scenarios through these courses, and because of the small size of the classes ($\sim$ 20 - 30 students), I had many opportunities to understand the students' concerns and problems. I also assisted in courses on Compiler Construction, Theory of Computation, and graduate courses on Automata, Algorithms, and Applied Algorithms, for all of which I held regular office hours to help students. Occasional fill-in duties for course instructors exposed me to larger classroom situations. During all these experiences, I learned to approach the subject from the students' perspective, with an appreciation of their background. I also understood the importance of direct interaction with students. A visit to their laboratory before assignments, or extra office hours before examinations – any form of one-on-one interaction led to a much higher comfort level for the students, which would then reflect in their class performance. In their feedback at the end of the courses, I found them appreciating this individual attention more than anything else.

Teaching the subject I am most familiar with, Computational Biology, presents unique challenges and excitement. Typical classes comprise students from both biological and quantitative backgrounds, and making the material accessible to everyone's aptitude, without compromising on the content, is challenging. I have presented my research in short courses at Rutgers University and Cornell Medical School, in seminars at various universities, as well as in several conferences with large audiences. From this experience, I have learned to draw the right balance between technical rigor and comprehensibility of the material.

I look forward to teaching undergraduate and graduate level courses on Computational Biology, tailored for students with purely biological or purely quantitative backgrounds, as well as a mix of the two. I would also like to design a graduate-level course on "Probabilistic Methods in Computational Biology", that will explore ideas in probability theory and statistics, with applications to bioinformatics. I am also excited about the possibility of co-teaching courses with biologists, wherein both quantitative and biological aspects of the same problems are presented in depth. My teaching experience also qualifies me to teach graduate level courses on Algorithms, Data Structures, Complexity theory, and undergraduate courses on several core topics in Computer Science, including Theory, Machine Learning, Compilers, and Machine Organization.

I firmly believe that a student, particularly at the graduate level, has the strength of an open mind, free from the biases that knowledge and experience bring. Therefore, I am very excited about having the opportunity to mentor students, and guide their research. I have had the privilege of working with wonderful mentors who gave me the chance to decide upon the problems or techniques that I believed in and wanted to work on. I will adopt the same approach toward the students I mentor, rather than treat them as tools in my research. My research advisees will principally work on problems in bioinformatics, but I will also love the opportunity to co-advise students with an experimental focus.

In ending, I include some of the comments that my students had on my teaching, as part of their anonymous, end-of-term evaluations of the quiz sections:

"Saurabh's help (contributed most to my learning); he did a fantastic job. I struggled in the beginning and he gave up a ton of his free time in providing me with help. Without his input I would have dropped this class weeks ago." ... "Saurabh's effectiveness in teaching difficult material (contributed most to my learning). Also extra help one on one whenever needed." ... "I think Saurabh did a great job! His help is available whenever in need! Thanks for dropping by in the lab in the middle of the night to help." ... "Talking to TA (was the aspect of the class that contributed most to my learning). Clearly, effort was put into making the class useful for us."