



UNIVERSITY OF WASHINGTON

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DECEMBER 1, 2004

Biocomplexity Faculty Search Committee
c/o Prof. Rob de Ruyter van Steveninck
Department of Physics, Swain Hall West 117
Indiana University
Bloomington, IN 47405-7105

Dear Prof. van Steveninck:

It is honestly a delight to write this letter in support of Saurabh Sinha's application for a position in your department. I was Saurabh's Ph.D. advisor. We worked together and continue to collaborate in the area of Computational Molecular Biology. Since earning his Ph.D. in 2002, Saurabh has spent the past two years as a postdoc with Eric Siggia at Rockefeller University, in order to further deepen his understanding of biology and computation. This is an uncommon move for most computer scientists, who can find faculty positions right out of graduate school, but very common for biologists. It is a move I encourage in all my Ph.D. students, since I cannot provide the depth of biology training I believe they should have.

In a nutshell, Saurabh is a delight to work with: extremely energetic, self-motivated, adept at the relevant mathematics, skilled at programming and problem-solving, and creative. Add to that the depth of understanding of biology and computer science he has acquired and you have a wonderful candidate for your computational biology research position.

The earliest problem Saurabh and I tackled successfully was that of predicting protein binding sites in DNA, eventually resulting in a tool called YMF that is freely available on the internet. Given a set of genes believed to be regulated by the same protein, the problem is to determine the most likely DNA patterns ("motifs") that serve as the protein's binding sites. The approach we used was to identify the most statistically overrepresented motifs in the input DNA sequences. Saurabh's research on this project consisted of (1) working out the details of the statistical overrepresentation theory, which were quite complex, (2) implementing the resulting algorithm, (3) running experiments on well studied families of co-regulated genes in yeast, to confirm that the algorithm could find their known binding sites, and (4) running experiments on families of functionally related genes in yeast, to see if we could discover novel binding sites. This was a challenging project, and Saurabh performed it superbly. The results of the experiments were quite exciting, in the sense that, for most of the gene families on which Saurabh experimented, the known binding site ranked at or near the top in statistical significance, and he discovered many excellent candidates for novel binding sites. Saurabh presented this work at the highly competitive 2000 *International Conference on Intelligent Systems for Molecular Biology* (ISMB), and we later wrote two papers directed at biologists that appeared in the prestigious journal *Nucleic Acids Research*. As a result, YMF has been downloaded by over 500 users and used in its web-based form countless times.

Saurabh went on to work on two related projects. The first was joint work with fellow graduate student Mathieu Blanchette, and dealt with postprocessing the output of motif discovery algorithms such as

YMF. Often the motifs reported by a computational method as being statistically significant include hundreds of small variations of just a few true motifs, the overrepresentation of these variants being explained solely by these true motifs. Saurabh and Mathieu devised a statistically sound method for finding the few true motifs among a multitude of close variants, and demonstrated that it recovers significant motifs with a very low false positive rate. This paper was presented by Saurabh at the highly competitive 2001 *International Conference on Intelligent Systems for Molecular Biology* (ISMB).

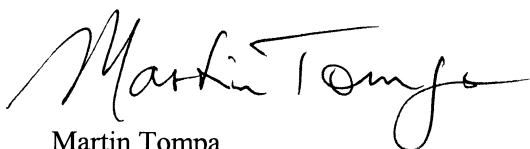
The second project was an innovative approach to the motif discovery problem, formulating it as a task of feature selection for classification between sequences that have the motif and those that do not. This approach enabled Saurabh to build on existing techniques from machine learning, resulting in a very general framework for finding a large class of motifs. This work was solely authored by Saurabh while still a graduate student, and presented at the very competitive 2002 *International Conference on Computational Biology* (RECOMB).

During Saurabh's recent postdoc years he has broadened his expertise in gene regulation and evolutionary models. Eric Siggia's lab was the perfect place to do this, and Saurabh continues to be very productive. He developed Stubb, an algorithm for genome-wide detection of modules of binding sites, which earned the SGI Best Paper Award at the 2003 *International Conference on Intelligent Systems for Molecular Biology* (ISMB). He went on to use the evolutionary and hidden Markov models developed in Stubb to solve a problem with which we had struggled for a few years, namely how to integrate both intra- and inter-species sequence data to better predict regulatory elements. Saurabh graciously included Mathieu Blanchette and me as coauthors, but the creativity and work of developing the software called PhyME were all his. Indeed, I would have never been able to come up with the evolutionary model and the differential equations necessary for the method. This paper has just been accepted for publication in *BMC Bioinformatics* and the software is again freely available on the internet.

Saurabh is remarkably versatile, with substantial research accomplishments in fields quite diverse from computational biology. In the summer and autumn of 1999, Saurabh did a six-month internship with Ramarathnam Venkatesan at Microsoft Research on cryptographic algorithms. This resulted in a paper at a DIMACS workshop on cryptography, and an invitation to collaborate in another internship in the summer of 2000. The second internship resulted in a paper on software watermarking in the 2001 *International Workshop on Information Hiding* and one on verification of code execution in the 2002 *International Workshop on Information Hiding*.

As you undoubtedly know, the recent revolutions in molecular biology have created incredible demand for qualified computational biologists today, both from universities and industry. For the coming years, until fledgling graduate programs in computational biology start turning out such qualified individuals, the supply will be dwarfed by the demand. It won't be easy to attract someone with Saurabh's qualifications under these conditions, and I wish you luck.

Sincerely,



Martin Tompa
Professor of Computer Science and Engineering
Adjunct Professor of Genome Sciences