THE ROCKEFELLER UNIVERSITY
1230 YORK AVENUE · NEW YORK, NEW YORK · 10021-6399

November 23, 2004

Biocomplexity Faculty Search Committee
Department of Physics
Indiana University
Swain Hall West 117
Bloomington, IN 47405-7105

Dear Faculty Search Committee:

Saurabh Sinha has been a member of my group for 2.5 years, first supported by a competitively awarded W.M. Keck Foundation fellowship from The Rockefeller University and in his final year by an NIH grant awarded to U. Gaul and me. He came with very strong bioinformatics skills and there was little for me to teach him on that score. The environment in which he chose to postdoc is on the quantitative side physics oriented, and otherwise, entirely biological. He started working with U. Gaul his first year and now does so in a completely independent manner, as often with me as without me. Much to his credit he has evolved during this period from providing tools to problems posed by biologists (e.g. how to treat expression array data), to asking questions about a biological system, understanding what computations are needed, proposing specific cases to test experimentally, and then returning to the calculations when the experiments only partially bear out prior expectations. I would expect that he could be very productive in a liberal genetics/developmental biology department.

His first project was to generalize a code we developed for finding regulatory modules to multiple species. This was done in a few months with very little input from me, and extended the original product to allow for either the fitting or imposition of positional correlations between the binding factors. (The correlation option was used to search for Hox-Exd factor binding sites using data from R. Mann, which has not been followed up experimentally.) It is not a trivial matter to optimally use multispecies data, since we have shown that there is only a modest correlation between proven factor binding sites and interspecies conserved sequence. However, binding sites that occur in conserved sequence blocks convey less information to discriminate the module from surrounding sequence than if the same two sites reside between two aligned blocks. (To see this consider the limit of two very similar species, they are identical by descent, except if there are indels that carry a binding site, which is then new information about the function of that piece of regulatory sequence.) The resulting 'Stubb' code summarizes in one score based on a surrogate for the binding free energy of factors to DNA, the contributions of aligned and unalignable sequence. The next 6 months were then spent processing the unassembled, unannotated *D.pseudoobscura* sequence (the first official annotation still has not appeared), and then comparing over a hundred known and predicted segmentation modules for the 50 or so key genes that define the anterior-posterior patterning of the embryo. The problem was to make sense of the changes and decide how much of the numerics was to be trusted. It should be emphasized

that the prior literature on regulatory evolution in fly was largely concentrated on one stripe module of the even skipped gene, which we now feel is not terribly representative. After all this is biology and not physics; there are no 'hydrogen atoms'. A considerable amount of software had to be generated, but less easy was looking at the output, and assimilating the literature from the 80's onward about this pathway.

A ranked list of interesting changes (and similarities) was presented to Ulrike in August 2003, and of order 24 constructs were made, injected (into *D.melanogaster*), and a dozen insitu hybridizations done in both species for the segmentation genes whose regulatory modules we were testing, (or for their input factors). A second round of experiments was started around May of 2004 to follow up on predictions that were wrong (e.g. there appears to be an unconventional activator influencing eve stripe37 in *D.pseudoobscura* binding to a short sequence insertion not present in *D.melanogaster*); plus cases were predictions worked and we wanted to verify the mechanism (e.g. we find a great reduction in sites for factor X to module Y, express Y in a background of reduced X). Writing will begin by November 2004. We feel a much better paper has emerged as a result of this iteration, and integration of experiment and calculation, but it takes time.

While the experiments were progressing, we turned our attention to the evolution of regulatory sequence, using first three way comparisons between *D.mel*, *D.psu*, and limited regions of *D.virilis*, and more recently we included the completed *D.yakuba* genome. Among our new findings are the prevalence of approximate tandem repeats for the generation of new sequence (indels account for more base pairs of change than do point mutations), plus the finding of a general expansion of regulatory sequence with the corollary that most of the euchromatic sequence in the fly is functional (others have shown that nonfunctional sequence is rapidly lost).

By early 2005, there will be more fully sequenced fly genomes than yeast. There are now many high profile papers that isolate by interspecies comparisons the ~5% of the human genome sequence that appears to be under selection, only a third of this codes for proteins, and the rest is suspected to be regulatory, but there is very little direct evidence for this, or how to parse it. A few conserved modules have been checked, but the ones that do not work out do not get reported, and no one has come close to showing that the conserved blocks recapitulate the endogenous expression of the gene. Fly is very useful intermediary between yeast and mammals, since most of the signaling pathways common to higher metazoans occur in fly and a great deal is known about gene regulation in fly beyond embryogenesis. We do not believe simple sequence conservation will suffice to pick out what is functional, but there is much more information than percent identity in these genomes. Hence, the future requires a fusion of molecular evolution with the specifics of wing and eye patterning, development of specific cell types etc. Saurabh will be well launched into this area by the time he leaves. He will also be abreast of the contributions miRNA's make to development by Ulrike, who has new experiments underway following her work with C. Sander's group at MSKCC. He is also playing with more

subtle ways to infer AP patterning in the mosquito from what is known in fly in collaboration with another postdoc in the lab who is trained in protein structure prediction.

Saurabh is certainly the strongest person in my group, and knowns more biology than two earlier alumni who have settled very productively into faculty jobs at Columbia and UCSF in biology departments. He knows less fly biology than N. Rajewsky who recently went to NYU, but is sharper and much easier to work with. He is the first professional computer scientist to pass through my group, and I would happily take more like him. He would do equally well in a computational biology department, as in biology itself.

Sincerely;

Professor Eric D. Siggia

EDS:mjl