

Cornell New Life Sciences Initiative/  
Department of Computer Science/  
Department of Biological Statistics & Computational Biology

# COLLOQUIUM

Tuesday, February 15, 2005

4:30pm, B17 Upson Hall

*Refreshments at 4:00pm in the Upson 4<sup>th</sup> Floor Atrium*

**Saurabh Sinha**

The Rockefeller University

*Hidden Markov Models and Mutant Fruitflies:  
A Case Study in Computational Biology*

High throughput technologies in molecular biology are generating vast amounts of data, allowing the groundwork for biological discovery to be done computationally. However, the data is (i) noisy, (ii) heterogeneous, (iii) incomplete, and (iv) voluminous. This makes it crucial to devise smart algorithms that analyze this data and extract biological knowledge from it.

In this talk, we will explore probabilistic models and algorithms that learn, from the genomes of multiple species, an important kind of biological information -- that of gene "regulation". Probabilistic constructs like the Hidden Markov Model are combined with simple mathematical models of evolution, to realistically capture the biological features of the data. The combined model integrates heterogeneous sources of data in a principled manner. A maximum-likelihood technique is then employed, via expectation-maximization algorithms, to identify the functional parts of the genome. Such computational identification, when followed up with genetic experiments, leads to an understanding of how a single cell develops into an adult organism, how organisms function, and how new species evolve -- three central problems in science today. Our algorithms were used to predict evolutionary differences between two species of fruitfly, that were then confirmed by wet-lab experiments.

The biological concepts used will be explained during the talk.