*Gene Regulation: Session Introduction*

S. Batzoglou and L. Pachter

Pacific Symposium on Biocomputing 8:3-4(2003)

# GENE REGULATION

SERAFIM BATZOGLOU
*Department of Computer Science*
*Stanford University*
*Palo Alto, CA 94305*
*serafim@cs.stanford.edu*

LIOR PACHTER
*Department of Mathematics*
*UC Berkeley*
*Berkeley, CA 94720*
*lpachter@math.berkeley.edu*

The year 2003 marks the completion of the sequencing phase of the human genome, and will coincide with the 50th anniversary of Watson and Crick's description of the structure of DNA. Despite such remarkable and rapid progress in elucidating the structure and composition of our genome, the problem of understanding the complexity of its function is largely unsolved. The unraveling of the regulatory code is certainly one of the next great frontiers in molecular biology.

Progress in the understanding of gene regulation will have to be driven by experimental data. Unfortunately, unlike genomic sequence data, gene expression experiments are not as "clean." EST data is notoriously messy, and gene chip experimental data is non-trivial to analyze. Furthermore, there are very few experimentally confirmed transcription factor binding sites, especially in the human genome, although there is a bit of data in other organisms. The situation is similar for interaction pathways, where there are few well studied and characterized examples.

The computational challenge is formidable and at the same time progress is essential, both in order to facilitate the understanding of experimental data, and also to drive the experimental efforts themselves. There are numerous examples of such efforts in the papers represented in this track, consisting of creative and original approaches both for the analysis of data and for the generation of hypotheses for experiment. It is interesting to note that the mathematical and computational techniques used include such diverse methods as differential equations, probabilistic methods, combinatorics and statistics

The variety of organisms studied is large. Some researchers are clearly concentrating on first understanding simpler regulation systems, thus following the advice of the mathematician George Polya who suggested: "if there is a problem you can't solve, then there is an easier problem you can't solve: find it." Thus, Eskin, Gelfand and Pevzner have wisely concentrated on analyzing bacterial promoter regions, and De Hoon et al. have examined time-ordered gene expression data from *Bacillus Subtilis*, both efforts resulting in successful

analysis for which there is a combination of existing data for verification and the realistic possibility of further experiment.

A large amount of research is currently being devoted to gene expression analysis, which is not surprising since there is a lot of chip array data to analyze now, and it is provides a first-order glimpse into the structure of regulation.
The paper by Berrar et al. suggests a new probabilistic approach to clustering expression data while Murali and Kasif explore new directions for detecting conservation of expression patterns. Jaeger et al. and Ganesh et al., in two separate papers demonstrate that pre-filtering can improve clustering performance on expression data. These analyses are important and should be of immediate interest to biologists who are generating expression data.

Expression analysis is just a preliminary step towards understanding entire regulatory networks, and there is already interesting computational work in this direction. The paper by Segal, Battle and Koller presents probabilistic model of cellular processes and show how to learn the processes from gene expression data. The paper by Li and Yuan on the relationship between gene expression profiles and survival data is a refreshing application of gene expression data not commonly examined by computational biologists.