*Joint Learning from Multiple Types of Genomic Data: Session Introduction*

A. Hartemink and E. Segal

Pacific Symposium on Biocomputing 9:262-263(2004)

# JOINT LEARNING FROM MULTIPLE TYPES OF GENOMIC DATA

A.J. HARTEMINK
*Department of Computer Science and*
*Center for Bioinformatics and Computational Biology*
*Duke University, Box 90129*
*Durham, NC 27708-0129*
*amink@cs.duke.edu*

E. SEGAL
*Department of Computer Science*
*Stanford University*
*Stanford, CA 94305*
*eran@cs.stanford.edu*

Recent technological advances enable us to collect many different types of data at a genome-wide scale, including DNA sequences, gene and protein expression measurements, protein-protein interactions, protein structural information, and protein-DNA binding data. These data provide us with a means to begin elucidating the large-scale modular organization of the cell. Indeed, much recent work has been devoted to the analysis of these data for this purpose. However, most of this work has been devoted to the analysis of a single type of data at a time, using other types of data only for validation.

In contrast, results jointly learned from more than one type of data are likely to lead to new insights that might not be as readily available from analyzing one type of data in isolation. For instance, experimental genomic datasets often contain errors arising from imperfections in the applied technology. Thus, some of the findings of methods that analyze a single type of data may be erroneous. If we assume that technological errors across different genomic datasets are largely independent, then the probability of error in results that are supported by two different types of data is dramatically reduced.

The Joint Learning from Multiple Types of Genomic Data session at PSB 2004 was created to provide a forum for novel methods that use more than one type of data in their analysis and do so jointly. Our goal in organizing this new session at PSB is two-fold: first, we hope to encourage the computational biology community to develop methods that are capable of integrating the large number of different types of data that are becoming increasingly available; second, we

hope to stimulate the discovery of new biological insights that would be difficult or impossible to identify in the analysis of only single types of data.

Based on the number of excellent papers submitted, the session has clearly tapped into a growing interest in such joint methods. Because of this large number of quality submissions, we were able to accept nine papers for publication. Interestingly, almost every one is different from the others in terms of the types of data used and the goal of the study. Some examples include: combining sequences from multiple organisms, or combining phylogenetic trees with sequence, for the task of detecting *cis*-regulatory motifs; combining gene expression and sequence for detecting operon structure; combining protein sequences with tertiary structural information for classifying proteins; combining protein-protein interaction data with gene expression for learning regulatory networks; and combining text from the literature with protein sequences for discovering functional domains in proteins. The methods employed for the joint learning were also very diverse, and included probabilistic methods, support vector machines, and methods from combinatorial optimization.

Taken together, these papers represent a fairly thorough cross-section of the most promising directions in this field. As more types of data become widely available, it is our belief that these kinds of unified approaches are likely to produce great insights into the complex biological systems that we are trying to better understand.

The session co-chairs are grateful to those who submitted papers to the session for their contributions in advancing the field of joint learning, and especially grateful to those who reviewed submissions for their contributions in selecting the most outstanding papers to present this year, which was a challenging task given the large number of excellent submissions.