

Joint Learning from Multiple Types of Genomic Data: Session Introduction

A.J. Hartemink and E. Segal

Pacific Symposium on Biocomputing 10:445-446(2005)

JOINT LEARNING FROM MULTIPLE TYPES OF GENOMIC DATA

ALEXANDER 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*

ERAN 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, information regarding protein structure and localization, 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 first appeared at PSB 2004 to provide a forum for novel methods that use more than one type of data in their analysis, and do so jointly; this year's session represents a continuation of that conversation. Our goal in organizing these sessions 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. We received 28 submissions of very high quality; we were able to accept 8 papers for publication, 6 of which were accepted for oral presentation. The papers represent a wide range of goals and approaches. Some examples include: combining DNA sequence data and gene expression data for improved detection of transcription factor binding sites, and at a higher level of detection, combining sequence data from multiple organisms for the task of identifying genomic regions that are bound by multiple transcription factors; predicting the function of uncharacterized genes by combining protein-protein interaction data, protein complex data and gene expression data; using gene expression data to better predict gene structure; and combining protein-DNA binding data, *cis*-regulatory binding data and gene expression for predicting regulatory networks. The methods employed for the joint learning were also very diverse, and included support vector machines, Bayesian networks and other Bayesian methods, non-negative matrix factorization, random forests, techniques from the data-mining community, 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.