Statistics Seminar Series

Statistical Models for Discovery and Clustering of DNA Binding Motifs

Shane T. Jensen
Department of Statistics
Harvard University

Tuesday, February 3, 2004, 4:00-5:00pm
Mathematics and Computer Science (MCS) Building, Room 135
111 Cummington Street, Boston

Tea and Cookies at 3:30pm in MCS 137

Abstract: In order to elucidate the detailed mechanisms of cell regulation, biologists are interested in discovering the DNA binding motif of regulatory proteins as well as modeling the common characteristics between different motifs. Bayesian models for the discovery of binding motifs are reviewed, and we introduce an approach based on optimization of a scoring function that allows us to improve upon the motif predictions found by previous stochastic algorithms. In addition, this scoring function formulation allows us to relax several standard model assumptions such as known motif size and motif abundance. Our approach has shown improved performance in both simulation studies as well as real-data examples. After a set of motifs has been discovered, the common structure between them is investigated by a Bayesian non-parametric hierarchical model. This non-parametric model enables similar motifs to be partitioned together into clusters. This approach has substantial benefits over traditional clustering methods. The clustering inference gained can be built into more sophisticated search algorithms in order to find new binding motifs in large genomes. Our strategy is also being used to predict co-regulated genes on the basis of the clustering of motifs found in the regulatory region of genes that are conserved between different but closely-related species.

This is joint work with Professor Jun S. Liu, Department of Statistics, Harvard University and Professor Richard M. Losick, Department of Molecular and Cell Biology, Harvard University.

For directions and maps, please see http://math.bu.edu/research/statistics/statseminar.html.