

STANFORD UNIVERSITY
DEPT OF STATISTICS
DEPARTMENTAL SEMINAR

4:15 p.m., Tuesday, October 9, 2001
Sequoia Hall Rm. 200
(Cookies at 3:45 in 1st Floor Lounge)

Xiaole Liu
Stanford Medical Informatics
Stanford University
Stanford, CA 94305

Computational statistical algorithms for finding transcription factor binding sites

The rapid development of sequencing technology enabled human genome and many other genomes to be sequenced and publicly available. The microarray technology has also become considerably more robust and sensitive. The combination of the two allows biologists to study gene expression and transcription regulation at a genome level. Given a set of upstream DNA sequences whose downstream genes are clustered together based on similarity in gene expression profile, or a set of DNA sequences enriched in chromatin immunoprecipitation followed by microarray experiments (ChIP-array), it is desirable to conduct computational analysis to find common sequence motifs that are the potential transcription factor binding sites regulating transcription. I will review the established algorithms for discovering common DNA motifs in a set of sequences, and propose two computational statistics approaches, BioProspector and MDscan. BioProspector searches for common sequence motifs from any general cluster of DNA sequences, especially potential transcription factor binding sites from upstream sequences of genes clustered by expression profile similarity. BioProspector adopts a Gibbs sampling motif discovery strategy, but provides many improvements. Motifs can have one-block, two-block, or two-block palindromic patterns. BioProspector allows variable copies of a motif per sequence, and uses background model with Markov dependency to improve the specificity of motifs. The statistical significance of a discovered motif can be calculated by Monte Carlo simulation. Preliminary results for testing each BioProspector feature have been very encouraging. A BioProspector web site is setup for biologists to load their sequences on the server for motif discovery. Another program, MatrixScan, is developed to search the genome for more potential sites using a discovered motif matrix. MDscan is a fast and novel algorithm

that looks for motifs from a set of sequences when one has confidence that a subgroup of the sequences contains the motif more abundantly. It can be used to find protein-DNA interaction sites from sequences selected by ChIP-array experiments because the sequences highly enriched by ChIP-array are very likely to contain the real protein-DNA interaction sites, and with multiple copies per sequence. The comparison of MDscan with several other motif-finding programs shows the advantage of MDscan in both speed and accuracy. It also succeeds in identifying the correct motifs from all published ChIP-array experiments. My research is designed to use computational statistics to find transcription factor binding sites and aid in studying the underlying mechanism of gene expression and transcription regulation. A better knowledge of transcription regulation can provide insights into understanding biological systems, and ultimately promote human health.