

STATISTICS COLLOQUIUM

FRIDAY, APRIL 27, 2007

TALK: 4:00 PM — SCIENCE CENTER RM. 705

RECEPTION: 5:15 PM — SCIENCE CENTER, 7TH FLOOR

“Decoding Mammalian Gene Regulatory Programs through Efficient Microarray, ChIP-chip and Sequence Analysis”

Hongkai Ji
Department of Statistics
Harvard University

ABSTRACT

Knowledge on how gene regulatory programs are encoded in the genome and executed in living cells is a key to understand human diseases. The goal of my thesis is to explore efficient statistical strategies to dissect mammalian gene regulation.

An empirical hierarchical Bayes approach was first proposed to analyze gene expression data collected from microarray experiments. Through a closed-form variance shrinkage estimator, information from multiple genes is pooled to increase the statistical power of multiple hypothesis testing. The approach also allows various types of subject matter knowledge to be incorporated conveniently. Caveats in controlling false discovery rate (FDR) will be discussed. When variance shrinkage estimator is employed, inappropriate use of permutations may result in underestimation of FDR.

Based on the hierarchical empirical Bayes approach, a TileMap method was developed for the tiling array data analysis. The method combines the hierarchical model with a Moving Average (MA) method or a Hidden Markov Model (HMM). Unbalanced Mixture Subtraction (UMS) was proposed to provide approximate estimates of false discovery rate for MA and model parameters for HMM. Applying TileMap to ChIP-chip allows one to detect transcription factor binding regions at a 500-2000 base pair resolution level. Systematic evaluations showed that the method significantly increased the performance of protein-DNA binding region detection compared to previously existing methods.

Finally, a comparative analysis involving six human and mouse transcription factors was performed to explore common characteristics of mammalian chromatin immunoprecipitation data and potential issues in their analysis. The cross-study comparisons, to the best of our knowledge the first to analyze multiple datasets, revealed the importance of matched genomic controls in the de novo identification of 6-30 base pair long transcription factor binding motifs, raised issues about the interpretation of ubiquitously occurring sequence motifs, and demonstrated the clustering tendency of protein binding regions for certain transcription factors.

The methods developed here were applied to dissect gene regulatory programs in mouse Sonic Hedgehog (SHH) signaling pathway. Through a combined analysis of gene expression, ChIP-chip and sequence motifs, new enhancers that are targeted by transcription factor GLI1 were discovered.