Colloquium

UNIVERSITY OF GEORGIA
DEPARTMENT OF STATISTICS

Hao Wu
Department of Biostatistics and Bioinformatics
Rollins School of Public Health
Emory University

"Genomic 'bump finding'"

Exploring genomic landscapes of different biological endpoints is an important approach for understanding biological processes and disease etiologies. Examples of these endpoints are sequence composition, DNA methylation, histone modifications, and binding sites for different transcription factors. With the completion of human genome project and advances of high-throughput technologies, tightly spaced measurements have been collected from linear chromosomes to create unbiased maps at the whole-genome scale. Detecting regions of interests from these data can be categorized as a general "bump finding" problem, where a bump is defined as a genomic location for which data behaves differently from the majority of the genome.

In this talk I will present several examples with the general theme of bump finding. In the first example we propose using Hidden Markov Models to search for CpG islands (CGI) from DNA sequence. The main advantage of our approach over others is that it summarizes the evidence for CGI status as probability scores, which provides flexibility in the definition of a CGI and facilitates the creation of CGI lists for many species. In the second example we construct a hierarchical model to detect transcription factor binding sites (TFBS) by jointly analyzing multiple related ChIP-chip datasets. This model captures the locational correlation among datasets, which provides basis for sharing information across experiments. Simulation and real data tests illustrate the advantage of the joint model over strategies that analyzes the individual dataset separately. Time permitting, I will present a unified frame work of TFBS detection and clustering from ChIP-seq data, which is an extension of the second example. I will show that jointly analyzing multiple ChIP-seq datasets can not only improve peak detection, but also cluster the detected peaks according to their binding probabilities in different samples.

Thursday March 3rd, 2011
ROOM 306
Statistics Building
University of Georgia
Athens, GA 30602
3:30 P.M. – Room 306, Statistics Building
Refreshments following talk at 4:30 P.M. in room 230 (The Cohen Room)