THE UNIVERSITY  OF HONG KONG

Institute of Mathematical Research
Department of Mathematics

Workshop in Mathematical Biology and Statistical Genetic

July 19, 2006
Room 517, Meng Wah Complex, HKU

9:00 – 10:00	**Wing Kam Fung**, Department of Statistics and Actuarial Science, HKU *Imprinting and Linkage Analyses Based on Case-Parents Trios*
10:00 – 11:00	**James Cai,** Genome Research Center, HKU *Probabilistic Models of DNA Sequence Evolution with Context-dependent Mutation*
Tea Break	
11:15 – 12:15	**Benny C.Y. Zee**, Centre for Clinical Trials, CUHK *Application of Block Wavelet Shrinkage Principle Components Analysis on DNA Microarry Data*
Lunch Break	
14:00 – 15:00	**Hongzhe Li**, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine *Pathway-Based Regression Modeling of Genomic Data*
15:00 – 16:00	**Tsz Lung Chan**, Department of Mathematics, HKU *Community Structure of Networks and its Applications*
Tea Break	
16:15 – 17:15	**Tze Leung Lai**, Department of Statistics, Stanford University and Department of Mathematics, HKU *Stochastic Segmentation Models for Array-based Comparative Genomic Hybridization (array-CGH) Data Analysis*

Abstracts

James Cai, Genome Research Center, HKU

Probabilistic Models of DNA Sequence Evolution with Context-dependent Mutation

Context-dependent mutation processes manifests specific patterns of dinucleotide frequencies in the genomes of most organisms. The CpG-methylationdeamination process, for example, is one of the prominent mutation processes in vertebrates (CpG effect). Context-dependent mutation processes need to be incorporated in order to make more realistic models of DNA substitutions. Based on a general framework of nucleotide substitutions, we developed a method to identify the most relevant dinucleotide substitutions and to estimate their relative frequencies. Our method starts from a model of context-independent nucleotide substitution, then context-dependent substitution parameters are successively added to increase the likelihood of the model in describing given data. We applied the method on the dataset from major eukaryotic lineages and revealed contrasting patterns in dinucleotide substitution.

Tsz Lung Chan, Department of Mathematics, HKU

Community Structure of Networks and its Applications

Networks are everywhere. Some examples are the World Wide Web, the Internet, gene networks and metabolic pathways. One important research area is in the identification of community structure in networks. Communities are groups of vertices within which connections are dense and among which connections are sparse. In this talk, I will review some of the methods developed including spectral graph partitioning, edge removal based on betweenness measures and optimization of the modularity function. Also, recent developments on finding overlapping community structure will be discussed.

Wing Kam Fung, Department of Statistics and Actuarial Science, HKU

Imprinting and Linkage Analyses Based on Case-Parents Trios

The recombination rates in meioses of females and males are often different. Some genes that affect development and behavior in mammals are known to be imprinted, and more than 1% of all mammalian genes are believed to be imprinted. When the gene is imprinted and the recombination fractions are sex-specific, the conventional transmission disequilibrium test (TDT) is shown to be still valid for testing for linkage. The power function of the TDT is derived, and the effect of the degree of imprinting on the power of the TDT is investigated. A simple statistic for testing for imprinting effects is also developed. The proposed parent-of-origin effects test statistic (POET) is shown to be normally distributed and can be employed to test for imprinting in situations where the marker locus need not be a disease susceptibility locus and where the female and male recombination fractions are

sex-specific. Based on the POET, a novel statistic ITDT for testing for linkage is proposed, which is shown to be more powerful than the TDT in the presence of imprinting. The validity of the POET, TDT and ITDT are assessed by simulation. The power approximation formulae for the POET, TDT and ITDT are derived and the simulation results show that they are accurate. The simulation study on power comparison shows that the ITDT outperforms the TDT for imprinted genes. The improvement can be substantial in the case of complete paternal/maternal imprinting. (joint work with Yue-Qing Hu and Ji-Yuan Zhou)

Tze Leung Lai, Department of Statistics, Stanford University and Department of Mathematics, HKU

Stochastic Segmentation Models for Array-based Comparative Genomic Hybridization (array-CGH) Data Analysis

Array-based comparative genomic hybridization (array-CGH) is a high-throughput, high-resolution technique for studying the genetics of cancer. Analysis of array-CGH data typically involves estimation of the underlying chromosome copy numbers from the log flourescence ratios and segmenting the chromosome into regions with the same copy number at each location. We propose for the analysis of array-CGH data a new stochastic segmentation model and an associated segmentation algorithm that has attractive statistical and computational properties. An important benefit of this Bayesian segmentation model is that we can use the posterior distributions of the number and locations of the change-points to provide confidence assessments of the segmentation. Applications to real array-CGH data sets and simulation studies illustrate the advantages of the proposed approach.

Hongzhe Li, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine

Pathway-Based Regression Modeling of Genomic Data

High-throughput genomic data provide an opportunity for identifying pathways and genes that are related to various clinical phenotypes, including censored survival phenotypes. Besides these genomic data, another valuable source of data is the biological knowledge about genes and pathways that might be related to the phenotypes of many complex diseases. Databases of such knowledge are often called the metadata. In microarray data analysis, such metadata are currently explored in post hoc ways by gene set enrichment analysis but have hardly been utilized in the modeling step. In this talk, we present two pathway-based regression models, including the pathway-based generalized linear and Cox regression models and the nonparametric pathways-based regression (NPR) models to efficiently integrate genomic data and metadata. Such pathway-based models consider multiple pathways simultaneously and can allow complex interactions among genes within the pathways. We present a pathway-based gradient descent boosting procedure for identifying the pathways that are related to the phenotypes. Our simulation studies indicate that the proposed boosting procedure can indeed identify relevant pathways. Application to a gene expression data set on breast cancer distant metas-

tasis identified that Wnt, apoptosis and cell cycle regulated pathways are more likely related to the risk of distant metastasis among lymph-node-negative breast cancer patients. Results from analysis of other two breast cancer gene expression data sets indicate that the pathways of Metalloendopeptidases (MMPs) and MMP inhibitors, as well as cell proliferation, cell growth and maintenance are important to breast cancer relapse and survival. We also observed that by incorporating the pathway information, we achieved better prediction for cancer recurrence. Extensions to incorporate information contained in the pathway structures in the framework of spectral graph theory and to model the pathway-pathway interactions will also be discussed.

Benny C.Y. Zee, Centre for Clinical Trials, CUHK

Application of Block Wavelet Shrinkage Principle Components Analysis on DNA Microarry Data