



Workshop on Computational Biology and Statistics

(For celebrating the 10th anniversary of
The Department of Statistics at University of California, Los Angeles)

Oct 19-20, California, USA

Time: October 19-20, 2008

Venue: IPAM (Institute of Pure and Applied Mathematics), UCLA, Los Angeles, California, USA

Organizers: Ker-Chau Li and Qing Zhou, Department of Statistics, UCLA

In connection with the celebration of the 10th anniversary of the Statistics Department, ULCA, a workshop highlighting the application of Statistics in Computational Biology will take place in the Institute of Pure and Applied Mathematics (IPAM), UCLA from October 19 to October 20.

The rapid growth in the public repertoire of biological data and knowledge resource, including the completion of genome sequencing for human and many species, the stride in the SNP detection and international HapMap project and the accumulation of full genome microarray gene expression data under a number of conditions for numerous organisms and tissues provide a rich research environment for sophisticate quantitative methods to overcome the seemingly insurmountable difficulties due to the enormous complexity of data structure and exceedingly high dimensionality. This workshop offers a golden opportunity to appreciate the role of Statistics in modern biology.

Program

Oct 19, Sunday

8:30-9:00 Registration

9:00-9:15 Opening Remark

9:15-10:45 Oral presentation Section I (Chair: Ker-Chau Li)

9:15-10:00 Christopher Lee, UCLA: Deciphering Genome Evolution and Function Using Multiple Measures of Selection Pressure

10:00-10:45 Fengzhu Sun, USC: The Application of EM Algorithms in the Study of Molecular Networks

10:45-11:00 Coffee break

11:00-12:30 Oral presentation Section II (Chair: Ker-Chau Li)

11:00-11:45 Matteo Pellegrini, UCLA: New methods for processing high-throughput sequencing data: bisulfite sequencing and transcriptional profiling

11:45-12:30 Tianwei Yu, Emory University: Adaptive Processing of High-Resolution LC/MS Data for Large-Scale Metabolomics Studies – New Algorithms and an R Package

12:30-2:00 Lunch (for speakers with statistics faculty)

2:00-3:30 Oral presentation Section III (Chair: Wei Sun)

2:00-2:45 Steve Horvath, UCLA: Weighted gene coexpression network analysis and causality testing for finding complex disease genes

2:45-3:30 Ching-Ti Liu, Boston University: A forest-based approach in genome-wide association study

3:30-3:45 Coffee break

3:45-5:15 Oral presentation Section IV (Chair: Qing Zhou)

3:45-4:30 Wei Sun, UNC: Genome-wide Multiple Loci Mapping in Experimental Crosses Using Dense Genetic Markers

4:30-5:15 Hsuan-Yu Chen, Academia Sinica: Molecular Signatures, Bioinformatics, and Clinical data in Cancer- Toward Personalized Medicine

Oct 20

9:00-10:30 Oral presentation Section V (Chair: Qing Zhou)

9:00-9:45 Jun Liu, Harvard University: Inference of Patterns and Associations Using Dictionary Models

9:45-10:30 Chiara Sabatti, UCLA: Genome wide association studies in a founder population

10:30-10:45 Coffee break

10:45-12:15 Oral presentation Section VI (Chair: Wei Sun)

10:45-11:30 Xianghong Zhou, USC: Integrated Approaches to Mapping Genome to Phenome

11:30-12:15 Richard Llewellyn, UCLA: Annotating Proteins with Generalized Functional Linkages

12:15-1:30 Lunch (for speakers with statistics faculty)

1:30-3:45 Oral presentation Section VII (Chair: Chiara Sabatti)

1:30-2:15 Mary Sehl, UCLA: A Race to the Death among Stem Cells: Implications for Cancer Therapy

2:15-3:00 Shinsheng Yuan, Academia Sinica: A network analysis on human microRNA expression

3:00-3:45 Chun Houh Chen, Academia Sinica: Matrix Visualization for High Dimensional Biomedical Data with Categorical Nature

3:45-4:00 Closing Remark

Abstracts

The Application of EM Algorithms in the Study of Molecular Networks

Fengzhu Sun, Molecular and Computational Biology Program, Department of Biological Sciences, USC <http://www-rcf.usc.edu/~fsun>

Expectation Maximization (EM) algorithms are widely used in many different fields. In this talk I will give two examples of using EM algorithms to the study of molecular networks. The first is the prediction of protein domain interactions based on protein interactions and the second is the identification of network motifs in stochastic molecular networks.

Rui Jiang, Zhidong Tu, Ting Chen, Fengzhu Sun (2006) [Network motif identification in stochastic networks](#). *Proc Natl Acad Sci USA* 103:9404-9409

Hyunju Lee, Minghua Deng, Fengzhu Sun, Ting Chen (2006) [An integrated approach to the prediction of domain-domain interactions](#). *BMC Bioinformatics* 7:269

Deng MH, Mehta S, Sun FZ, Chen T (2002) [Inferring domain-domain interactions from protein-protein interactions \(supplement\)](#). *RECOMB2002*:117-126. Also on *Genome Research* 12:1540-1548.

New methods for processing high-throughput sequencing data: bisulfite sequencing and transcriptional profiling

Shawn Cokus, Matteo Pellegrini
MCD Biology, UCLA

High-throughput sequencing has undergone remarkable increases in efficiency over the past few years. Some of the most efficient machines are those produced by Illumina, which sequence approximately 1-3 billion bases every three days as millions of ~36 base reads. We have developed novel techniques to process and interpret such data that extend the standard data analysis pipeline. There are two main areas we address with our software: a probabilistic base calling module and a mapping tool that aligns short reads to reference sequences while accounting for base call quality.

The first aspect of our software involves the estimation of the sequence of each read. Our base calling tool starts with Solexa base calls and attempts to correct systematic errors in these using multidimensional Gaussian mixture models. Starting with Solexa base calls we estimate the distribution of each base using Gaussian hyper-ellipsoids. These are then used to infer the probability that each base in a read is an A, C, G or T.

Our second tool aligns these probabilistic reads to sequenced genomes uses suffix trees. Full use is made of information available at every base of every read as to the probability of A, C, G, and T, and scoring of reads is statistically grounded via algorithmic, non-heuristic consideration of

whole-genome likelihoods. Our tool also supports sodium bisulfite-converted reads suitable for study of DNA cytosine methylation at the single base level.

We apply these tools to a variety of data sets. The first involves the estimation of DNA methylation in Arabidopsis using bisulfite sequencing. The second involves the measurement of the transcriptional landscape in Chlamydomonas. We use this data to estimate transcript levels of each gene, differential expression in differing conditions and to correct existing gene models.

Adaptive Processing of High-Resolution LC/MS Data for Large-Scale Metabolomics Studies – New Algorithms and an R Package

Tianwei Yu, Emory Univ.

Liquid chromatography-mass spectrometry (LC/MS) profiling is a promising approach for the quantification of metabolites from complex biological samples. Significant challenges exist for the analysis of LC/MS data, including noise reduction, peak identification and quantification, and peak alignment. Vast amounts of data require the algorithms to be highly efficient. Here we present a set of algorithms for the processing of high-resolution LC/MS data. The major technical improvements include the adaptive tolerance level searching rather than hard cutoff or binning, the use of non-parametric methods to fine-tune intensity grouping, and the model-based estimation of peak intensities when m/z sharing occurs. The algorithms are implemented in an R package apLCMS, which can efficiently process large LC/MS datasets.

Weighted gene coexpression network analysis and causality testing for finding complex disease genes

Speaker: Steve Horvath, Departments of Human Genetics and Biostatistics, UCLA

Weighted gene co-expression network analysis (WGCNA) facilitates a systems genetic view of gene expression data and SNP marker data. The network framework makes it straightforward to integrate gene expression data, clinical traits, and genetic marker data. This talk covers several theoretical topics including network construction, module definition, network based gene screening, differential network analysis, and causality testing with the software NEO. The methods are illustrated using several applications including gene expression and genotype data from an F2 mouse intercross. Related articles and material can be found at the following webpage <http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/>

A forest-based approach in genome-wide association study

Ching-Ti Liu, Boston University

Multiple genes and gene-by-gene interactions are believed to underlie most complex diseases. However, such interactions are difficult to identify. While there have been recent successes in identifying genetic variants for complex diseases, it still remains difficult to utilize haplotype information to identify gene-gene interactions. To overcome this difficulty, we propose a forest-based approach and a concept of variable importance. The proposed approach is demonstrated by simulation study for its validity and illustrated by a real data analysis for its use. Analyses of both real data and simulated data based on published genetic models show the effectiveness of our approach.

Genome-wide Multiple Loci Mapping in Experimental Crosses Using Dense Genetic Markers

Wei Sun, UNC

Multiple loci mapping has been a long-time challenge for genetic studies of complex traits. Many existing methods have been designed and evaluated for genetic markers that are not densely distributed. With recent advances in genotyping techniques, most current genetic studies employ genotype arrays with dense genetic markers across the genome. In this study, we propose two new multiple loci mapping methods, namely Bayesian Adaptive Lasso and Iterative Adaptive Lasso, and we evaluate the proposed methods as well as several existing methods using both sparse and dense marker maps. We show that the proposed methods have improved variable selection ability and/or efficiency compared to most existing methods, especially when the genetic markers are densely distributed.

Molecular Signatures, Bioinformatics, and Clinical data in Cancer- Toward Personalized Medicine

Yi-Chuing Hsu and Hsuan-Yu Chen, Institute of Statistical Science, Academia Sinica

Cancer is a complex disease. There are three major levels to describe the underlying biological processes: genomic DNA, RNA, and protein. Changes of elements in above three levels can be measured by RT-PCR-based assays or high throughput technologies like microarray. Yet how to combine the genomic data with clinical outcome data for shedding light on cancer biology is a major challenge.

In this talk, I will describe the recent progress in utilizing data from gene-expression in mRNA and microRNA to predict survival of lung cancer. At the mRNA expression level, we found a five-gene signature which could predict patients' overall and relapse-free survival in two sample cohorts from Taiwanese population and in Western population. At the microRNAs expression level, we also found a set of five microRNAs that could predict patients' outcome in the testing sample and is validated by samples from another hospital which is located far from the original hospital collecting the training samples. Moreover, by functional assay, the 5 microRNAs were shown to have biological functions in tumor invasiveness.

Our bioinformatics approach is powerful in extracting useful information from large scale genomic and clinical data. The obtained molecular signatures could improve the traditional prognosis by cancer stage. They can be used to guide the proper clinical treatment assignment according to the patients' molecular risk scores. The combined use of bioinformatics, epidemiology and biostatistics could aid the progress in personalized medicine.

Inference of Patterns and Associations Using Dictionary Models

Jun Liu, Department of Statistics, Harvard University

Pattern discovery is a ubiquitous problem in many disciplines. It is especially prominent in recent years due to our greatly improved data-generation capabilities in science and technologies. The method I present here is motivated by the "motif-finding" and "module-finding" problems in biology, i.e., to find sequence patterns (i.e., "words") that seem to appear more frequent than

usual in a given set of text sequences (i.e., sentences) and to find which of these "words" tend to co-occur in a sentence. A challenge in the motif-finding problem is that there are no spacings and punctuations between the words and the dictionary of "words" is unknown to us. Existing methods are mostly "bottom-up" approaches, i.e., to build up the dictionary starting with single-letter words and then concatenate some existing words that appear to occur next to each other in sentences more frequently than chance. Our new approach is a top-down strategy, which uses a tree structure to represent the relationship among all possible existing words and uses the EM algorithm to estimate the usage frequency of each word. It automatically trims down most of the incorrect "words" by letting their usage frequencies converge to zero.

The module-finding problem is closely related to the well-known "market basket" problem, in which one attempts to mine association rules among the items in a supermarket based on customers' transaction records. It is also related to the two-way clustering problem. In this problem, we assume that the words are given, and our goal is to find subsets of words that tend to co-occur in a sentence. We call the set of co-occurring words (not necessarily orderly) a "theme" or a "module". We can generalize the dictionary model to the "theme"-model and use a similar EM-strategy to infer these themes. I will demonstrate its applications in a few examples including an analysis of chinese medicine prescriptions and an analysis of a chinese novel.

This is based on a joint work with Ke Deng and Zhi Geng.

Genome wide association studies in a founder population

Chiara Sabatti, UCLA

Genomewide association studies (GWAS) of longitudinal birth cohorts enable joint investigation of environmental and genetic influences on complex traits. We report GWAS results for nine quantitative metabolic traits (triglycerides, high density lipoprotein, low density lipoprotein, glucose, insulin, C-reactive protein, body mass index, and systolic and diastolic blood pressure) in the Northern Finland Birth Cohort 1966 (NFBC1966), drawn from the most genetically isolated Finnish regions. We replicate most previously reported associations for these traits and identify nine novel associations, several of which highlight genes with metabolic functions. The currently identified loci, together with quantified environmental exposures, explain little of the trait variation in NFBC1966.

Integrated Approaches to Mapping Genome to Phenome

Xianghong Zhou, USC

In this talk, we will report our recent effort in utilizing the rapidly accumulating body of genomics data, especially the enormous amount of public microarray data, together with the associated phenotypic and environmental context information to reconstruct the biological basis of phenotypes. Traditional association studies have been relatively successful at relating genetic polymorphisms to phenotypes. However, they have met difficulties in elucidating the gene-gene interactions that contribute to complex phenotypes. Here, we develop novel methods aimed at deriving genome-wide molecular networks of genotype-phenotype associations. Furthermore, we develop methods to perform phenotype prediction and computational diagnosis utilizing public genomics databases, particularly the large public microarray repositories, to create an automated disease diagnosis database.

Annotating Proteins with Generalized Functional Linkages

Richard Llewellyn, UCLA

Functional linkages describe pairwise relationships between proteins that work together to perform a biological task. The function of one protein, when known, suggests the function of any unannotated partners that are targets for protein function prediction. Function prediction may improve by combining many pairwise relationships, but noise may also increase. Proteins with conflicting annotations of varying quality are often linked to the target with unequal levels of strength. The challenge of combining these pairwise relationships into a unified description of protein function requires balancing the contribution of the linked proteins with each other and other sources of evidence. We have developed a Bayesian framework, known as Generalized Functional Linkages (GFL), to combine the information in functional linkages with existing knowledge about a protein, such as negative experimental results, its cellular location, or its individual molecular function. The result of GFL is a distribution over Gene Ontology biological process annotations that retains conflicting inferences from multiple lines of evidence, but can provide single best estimates of target protein function. GFL handles proteins with multiple annotations and addresses uncertainty in these annotations as well as incompleteness in the Gene Ontology itself. We have demonstrated the performance of GFL with links defined by zorch, the result of an algorithm used to quantify connectivity in protein-protein interaction networks. We show that functional linkages quantified by zorch are good predictors of the biological processes of proteins in *S. cerevisiae*, even when using only indirect or high-throughput protein-protein interactions. GFL increases the accuracy and coverage of protein function prediction by combining multiple lines of lower quality evidence that may be insufficient alone.

A Race to the Death among Stem Cells: Implications for Cancer Therapy

Mary Sehl, UCLA

Cancer stem cells have been shown to play a role in both originating and driving the progression of several different malignancies. Recent interest has arisen in developing therapies that target cancer stem cells. Because normal stem cells are essential for the maintenance and repair of normal tissues, there has been a recent push to identify therapies that selectively kill cancer stem cells, while sparing normal stem cells. We here address the question of what is the relative difference in death rates of cancer stem cells and healthy stem cells required to assure that healthy stem cells are present once all cancer stem cells have been eliminated? Modeling stem cell population dynamics as Kendall's birth-death process without immigration, we examine the probability distributions of the extinction times for cancer and healthy stem cells. By employing a theorem from the asymptotic theory of extreme value statistics, we identify the asymptotic limiting distribution of these extinction times. This allows us to compare their distributions for varying values of the two death rates. The most meaningful comparison involves the mean and variance of the normal stem cells at the extinction time of the cancer stem cells. This calculation can be accomplished by conditioning on the asymptotic time to extinction of the cancer stem cells.

A network analysis on human microRNA expression

Shinsheng Yuan, Institute of Statistical Science, Academia Sinica, Taiwan

We investigated the global pattern of microRNA expression in human. Three microRNA expression datasets of different quantification protocols and different experimental contexts were obtained. Despite the overall substantial differences in the pairwise coexpression, we are able to find two common features among the three: (1) the correlation in expression of microRNAs may be induced by the microRNA sequence similarity; (2) when two microRNA families have a pair closely located microRNAs, their correlations are generally enhanced. We further found that in general there are no correlations between microRNA and their known target gene expression. Lastly, we employ the novel statistical method of liquid association (LA) to analyze the relationship between microRNA and their target genes. The expression pattern of regulation of Hsa-mir-125a on ERBB2 network is discussed in detail. Our short list of high LA score genes includes EPS8, epidermal growth factor receptor pathway substrate 8.

Matrix Visualization for High Dimensional Biomedical Data with Categorical Nature

Chun-houh Chen, Institute of Statistical Science, Academia Sinica, Taiwan

Exploratory data analysis (EDA, Tukey, 1977) has been introduced and extensively used for more than 30 years yet boxplot and scatterplot are still the major EDA tools for visualizing continuous data in the 21st century. On the other hand, multiple correspondence analysis (MCA) type of methods (HOMALS: Gifi, 1990; MCA: Benzecri et al., 1973; Dual Scaling: Nishisato, 1984) and mosaic plots (Hartigan and Kleiner, 1981; Friendly, 1994) are most popular in practice for visualizing multivariate categorical data. But all these methods lose their efficiency when data dimensionality gets really high (hundreds/thousands), particularly when data is of nominal nature.

The categorical generalized association plots (cGAP) is an extension of the generalized association plots (Chen, 2002; Tien et al., 2008; Wu et al., 2008), which was developed as a matrix visualization environment for high-dimensional categorical data. Integrating matrix visualization with HOMALS's reduced joint space for samples and variables of categorical nature, cGAP can effectively present complex patterns for thousands of categorical variables for thousands of subjects in one matrix visualization.

Data generated and collected from biomedical experiments and studies are quite often of categorical nature. In this talk cGAP will be applied to analyze several such high dimensional categorical data sets. We believe GAP and cGAP related matrix visualization techniques have great potential to become major data/information visualization tools for next generation EDA. Related information can be obtained at:

<http://gap.stat.sinica.edu.tw/Software/index.htm>