

CONTRIBUTED SESSIONS

Thursday, 24th May 2007

2:00 p.m. – 3:40 p.m.

C-1 Analysis of Microarray Experiments

2:00 p.m. *An Iterative Nonparametric Quasi-Likelihood Method for Modeling Gene Expression Measurements*, **Qing Li, Nan Lin**, Washington University in St. Louis

It is essential to have a suitable error model for gene expression measurements in the statistical analysis of microarray data. The current models often use distributional assumptions, while the true probabilistic distribution is generally unknown. Strimmer (2003) proposed a semi-parametric model based on the extended quasi-likelihood method, which only requires the specification of mean and variance functions. Nevertheless, the specification of the variance function in the extended quasi-likelihood is still not an easy job. In this paper, we propose an iterative nonparametric quasi-likelihood approach that models the variance function nonparametrically. Our model only requires smoothness of the variance function and can easily incorporate the often-assumed monotonicity constraint of the gene expression variance function. The effectiveness of our model is illustrated by simulation studies.

2:20 p.m. *The High-level Similarity of Some Disparate Affymetrix GeneChip Expression Measures*, **Nandini Raghavan, Dhammika Amaratunga**, J&JPRD

A number of probe-set summarization techniques (PSSTs) have been proposed for summarizing Affymetrix GeneChips, including MAS4, MAS5, dChip, RMA, GC-RMA, FARMS, DFWFCB. The Affycomp website (<http://affycomp.biostat.jhsph.edu/>) lists several dozen more. Of serious concern to scientists and researchers is that the various PSSTs produce different results, at least as far as individual genes are concerned. An absolute comparison is impossible due to the lack of a gold standard. Instead, indirect means using, for example, spike-in experiments, in which known concentrations of mRNA have been added to the hybridization cocktail, have been used to assess them. However, these PSST comparisons, based as they are on individual gene results, miss the point that biological processes are driven by the actions of series of interacting genes. It is pertinent to inquire whether the different summarization schemes might in fact produce similar high-level results. In this talk, we present the results of such an evaluation.

2:40 p.m. *Graphical Modeling for Discrete Random Variables with Application to Tissue Microarray (TMA) Experiments*, **Corinne Dahinden, Peter Bühlmann**, ETH Zurich, Switzerland

Tissue microarrays (TMA) are composed of hundreds of tissue sections from different patients arrayed on a single glass slide. With the use of immunohistochemical staining, they provide a high-throughput method of analyzing potential biomarkers on large patient samples. The assessment of the expression level of a biomarker is usually performed by the pathologist on a categorical scale.

The analysis of the interaction of these biomarkers and in particular the estimation of the graph structure associated with the underlying discrete random variables, are of biological importance. Questions such as how the interaction pattern changes with progressing tumor grade or with survival time are of direct biological interest. However, the estimation of the interaction structure requires sophisticated techniques. Our approach is to fit an ℓ_1 -regularized log-linear model assuming a multinomial sampling scheme in order to obtain the graphical model. The regularization becomes necessary as after cross-tabulation of the samples in contingency tables, many cell entries remain zero, leading to so-called sparse contingency tables, where standard procedures fail to work.

We compare our approach with other methods for graphical modeling. Moreover, biological validation of the estimated interaction structure is done by mapping to known biochemical pathways.

3:00 p.m. *Consistent Probe Performance Yields Solutions Capable of Classification Across Independent Experiments*, **Kevin Thompson**, George Mason University

Microarrays are high throughput data collection technologies; those measuring gene expression allow investigators to simultaneously estimate the level of thousands of cellular transcripts present in a sample at the time of collection. Our lab has demonstrated that, by factoring in a combination of probe parameters and known sources of sample heterogeneity, it is possible to greatly improve the consistency of the expression profiles across experiments. We have shown, using experiments performed using Affymetrix GeneChip arrays, that by taking these factors into account the improved profile consistency enhances statistical testing and produces more robust predictive results with respect to recovering common gene lists from common disease states from independent experiments.

3:20 p.m. *Analysis of Comparative Genomic Hybridization Array Data Using Statistical Process Control*, **Xia Li, Adam Allred, Matthew Walter, Rhonda Ries, Timothy Ley**, and **William Shannon**, Washington University in St. Louis School of Medicine

We present statistical process control methods for analyzing high density comparative genomic hybridization array (aCGH) data. aCGH is an efficient two-colored chip-based array methodology used to scan an entire genome for DNA copy number differences. Sample DNA and reference DNA are labeled by green and red fluorescent dyes and

simultaneously hybridized to a chip. Using the commercially available NimbleGen technology we are able to query 385,000 oligo probes across a genome per chip. Probe spots with a $\log(\text{red}:\text{green dye ratio})$ near 0 indicate the same DNA copy number for that probe, while spots with $\log(\text{red}:\text{green dye ratio})$ significantly higher or lower than 0 indicate DNA copy number differences (copy number gains or losses) in the test sample relative to the reference sample. The goal is to identify DNA copy number changes that might be associated with disease and outcome.

Classical cytogenetic techniques already exist to identify large regions of DNA copy number changes. The statistical challenge with aCGH data is to use the fluorescent ratio information to identify 'short' regions of the genome with DNA copy gain or loss in the presence of random variation in the $\log(\text{red}:\text{green dye ratio})$ values. We have implemented an efficient statistical process control approach (i.e., control charts) to identify regions along the chromosomes where the log ratio differs from 0, perhaps associated with DNA copy number differences, and developed a scoring method based on mean squared error to rank order these segments. Drill-down graphics are used to display these regions efficiently. We developed this data analysis approach in response to a leukemia study using high density comparative genomic hybridization. In the leukemia experiments 32 small regions of DNA copy number changes were identified that were present in leukemia samples and absent in normal tissue.

3:40 p.m. – 4:00 p.m.

BREAK

4:00 p.m. – 5:40 p.m.

C-2

Visualization of Biological Data

4:00 p.m. *Integrating Network Views and Inference with Explorase*, Xiaoyong Sun, Michael Lawrence, Dianne Cook, Iowa State University

ExploRase is a tool for visually exploring microarray data and other bioinformatics experimental data. It uses R for statistical analysis and data processing, and GGobi for multivariate data visualization. The GUI is developed in RGtk2. This talk describes work that we are doing to incorporate network views, and network inference. The motivation for integrating networks is to examine experimental data in the context of biological expert knowledge.

4:20 p.m. *Network Models for time course microarray data*, **Inoue, L.Y.T., Telesca, D., Neira, M., Nelson, C., Gleave, M., Etzioni, R.**, University of Washington

Network models are the focus of a growing number of researchers concerned with discovering novel gene interactions and regulatory relationships between genes from expression data. In this talk we will discuss two approaches for inferring networks from time-course expression data. In the first half of the talk we will present a model-based approach that unifies the processes of inferring networks and clustering genes. Specifically, we provide a probabilistic framework for inferring clusters from gene expression profiles. Genes within the same cluster are expected to share a similar expression profile. We build a network over clusters using state-space models. In the second half of the talk, we will discuss an approach for inferring networks from time course microarray data which relies on modeling gene expression profiles as random functional transformations of a reference curve. Using measures of functional similarity and time order based on estimated warping functions we discuss time-varying networks.

We will illustrate the methods with simulation studies and a case study using time course microarray data arising from animal models on prostate cancer progression.

4:40 p.m. *"Heated" 3D Scatterplots for the Simultaneous Display of cgh Array and Gene Expression Data*, **Juergen Symanzik**, Utah State University, and **William Shannon**, Washington University School of Medicine

Plotting cgh array or gene expression data in simple scatterplots with the x-axis representing the location on the genome and the y-axis representing the recorded cgh or expression value (or its log) and the red-green heat maps are common ways of graphically representing cgh array and gene expression data. However, there lacks any good graphical representation that simultaneously displays cgh array and gene expression data to allow for quick associations of the form "high cgh" and "high expression" values that often are of particular interest in diseases such as breast cancer.

In this talk, we will show how 3-dimensional scatterplots can be used for this purpose, with the x-axis representing the cgh probe location on the chromosome, the y-axis representing the cgh value (or its log), and the z-axis representing the gene expression value of the gene expression location closest to the cgh probe location. We will discuss which effects different methods for selecting the values for the z-axis have, e.g., matching the xy-pair with the minimum or maximum of the gene expression values in a neighborhood of the cgh probe location. Such a minimum/maximum selection better allows for measurement errors and the effect that related events may take place at different locations, e.g., the deletion of some genes resulting in the amplification of cancer supporting gene expressions in a local neighborhood of the deletions (but not at exactly the same locations) as is the case with leukemia. We will further show how to use interaction on these graphics to find additional features in the data. All computations and graphics are done in R and its numerous packages.

5:00 p.m. *RJaCGH, a method for the analysis of CGH arrays with a Hidden Markov Model fitted via Reversible Jump MCMC*, **Oscar M. Rueda, Ramon Diaz-Uriarte**, Spanish National Cancer Center, Madrid, Spain

Genomic DNA copy number alterations (CNAs) are associated with complex diseases, including cancer: CNAs are indeed related to tumoral grade, metastasis, and patient survival. CNAs discovered from array-based Comparative Genomic Hybridization (aCGH) data have been instrumental for identifying disease-related genes and potential therapeutic targets. To be immediately useful in both clinical and basic research scenarios, aCGH data analysis requires accurate methods that do not impose unrealistic biological assumptions and that provide direct answers to the key question "What is the probability that this gene/region has CNAs?". Current approaches fail, however, to meet these requirements.

Here, we introduce a new method for identifying CNAs from aCGH; we use a non-homogeneous Hidden Markov Model fitted via Reversible Jump Markov Chain Monte Carlo, and we incorporate model uncertainty through Bayesian Model Averaging. We use a non-homogeneous Hidden Markov Model to incorporate explicitly the distance between genes/probes. By using Reversible Jump, we do not need to fix in advance the number of hidden states, nor do we need to use AIC or BIC for model selection (two criteria that are not well suited for Hidden Markov Models). We can therefore investigate the likely number of hidden states in the data and, more importantly, provide posterior probabilities that a gene or a set of genes is in a given state. To summarize results, we employ Bayesian Model Averaging, averaging over models with different states, thus incorporating model uncertainty and not conditioning our inferences to the selection of a particular model. Our method can be used to analyze data from each chromosome independently or all chromosomes together, offering both flexibility in the biological phenomena studied and increased statistical precision. Thus, our method provides a rigorous statistical foundation for locating genes and chromosomal regions with altered copy number and potentially related to cancer and other complex diseases.

The problem of finding minimal common regions is handled in a natural way, giving joint probabilities of alterations in specific locations in the genome across several samples, taking into account the variability in every array.

We show, using simulated and real data sets, that RJaCGH outperforms alternative methods (using as criteria classification error rate, false discovery rate, sensitivity, and specificity), and the performance difference is even larger with noisy data and highly-variable inter-probe distance, both commonly found features in aCGH data. The excellent performance of RJaCGH emphasizes the need to explicitly incorporate the inter-probe distance and use appropriate approaches (such as Bayesian Model Averaging) for accounting for model uncertainty.

5:20 p.m. *Visualization and Analysis of High Resolution Cell Image Data*, **Roy Welsch**, Massachusetts Institute of Technology

The mammalian cytoskeleton is a multi-component system responsible for maintaining the cell's structural integrity and transducing mechanical and chemical signals. Like

many in biology, this system involves hundreds of different proteins that transiently interact in specific protein complexes at discrete locations within the cell. Imaging fluorescently-labeled proteins in cells enables us to measure their location over time relative to structural markers. Undertaken quantitatively and on a large-scale, this approach enables us to build dynamic spatially-resolved models of cellular systems. These models provide a mechanism for sharing information, testing hypotheses, and also act as an active repository into which information from ongoing experiments can be added.

Imaging is an extremely rich information source. Acquisition of image data by optical microscopy is well-developed and as a result, several software options exist for image processing and image analysis. The combination of imaging and image analysis has led to the generation of large quantities of image-derived data but not necessarily information or understanding about the system studied. The absence of well-developed methods for multivariate data analysis, data visualization and spatial modeling of dynamic systems represent three major obstacles that currently limit use of image-derived data for systems modeling. In this paper we present some recent research in these areas, emphasizing data visualization. This is joint work with J. Evans, A. Samarov, R. Menjoge, J. Rajapaske, and A. Kumar.

Friday, 25th May 2007

8:30 a.m. – 10:10 a.m.

C-3

Protein-Protein Interaction and Microarrays

8:30 a.m. *Integration Of Relational And Hierarchical Network Information: Prediction Of Protein Function*, Xiaoyu Jiang, Eric Kolaczyk, Simon Kasifm, Boston University

In the current climate of high-throughput computational biology, the inference of a protein's function from related covariates, such as protein-protein interaction (PPI) relations, has become a canonical problem. Most existing technologies pursue this task as a Gene Ontology (GO) term-based classification problem. However, ontology structures are essentially hierarchies, with certain top to bottom annotation rules. We propose a probabilistic framework to integrate relational data, in the form of a PPI network, and the GO hierarchy, and offer two classifiers classifying GO terms marginally and jointly. Efficient dynamic programming algorithms for their computation are derived. We apply and evaluate our model in the yeast *Saccharomyces cerevisiae* for a whole-genome protein function prediction. It is found that substantial improvements may be obtained over non-hierarchical methods.

8:50 a.m. *Could Logistic Model Predict Yeast Protein-Protein Interaction from Subcellular Localization Pattern?*, Junfeng Liu, West Virginia University, Hongyu Zhao, Yale University, Jun Tan, & E. James Harner, West Virginia University

Compared to universal machine learning methods, parametric modeling approaches to classification is highly desirable due to its efficiency and interpretability. Classification by fitted probability from logistic regression have been intuitively applied in practice. This paper discusses some guidelines for the application of this method with a focus on misclassification error rates under certain feasible scenarios. A case study is done for the association between yeast protein subcellular localization pattern and protein-protein interaction.

9:10 a.m. *A Computational Model and Simulation of Demyelination Phenomenon in Neuroimmunological Disorders*, A. Zakharyan, University of Colorado, K. Kafadar, University of Colorado, D. Skundric, Wayne State University School of Medicine, and V. Zakarian, University of Colorado

To understand the dynamics of chemokine-cytokine-cell interactions and spatiotemporal changes in expression patterns in infiltration and demyelination phenomena during EAE,

we developed a computational model using VIXDUM software package. The computational model involves dozens of carefully modeled interactions such as chemokinesis, binding and release of various proteins. Pivotal processes that seem to play critical role in the model are the following. Cytokines produced by activated T Cells attach to and activate endothelial cells. Chemokines produced by endothelial cells penetrate interstitium and activate macroglia and astrocytes. Activated macroglia and astrocytes produce chemokines that chemoattract activated T Cells. Activated T Cells penetrate and attack the myelin. The model involves all these interactions as well as numerous parallel interactions and is defined in three spaces: vascular, endothelial and interstitial. The simulation results show the comparative role of individual components in the processes leading to demyelination. A number of parallel processes of activation, chemokine and cytokine production and chemoattraction influence the dynamics of the processes; the dynamics can be affected by changes in the parameters in some of the parallel processes. The developed model can be used to simulate various conditions that lead to cyclic infiltration and demyelination (remission-relapse) and/or to a progressive disease.

9:30 a.m. Using 'Flex' Motifs in Predictive Modeling of Antisense Oligonucleotides to Radically Increase Success Rate, Tamara B. Sipes, SciberQuest, Inc.

Antisense oligonucleotide technology allows the targeted reduction of mRNA expression through the *in vivo* application of short (~20 nt) DNA molecules. In theory, a base sequence complementary to a region of the transcript would hybridize to its mRNA target. Nevertheless, in practice some complementary antisense oligonucleotides are more active and more potent than others in suppressing specific gene expression. Previous research revealed a correlation between the short sequence motifs (tetramotifs or shorter) as well as certain thermodynamic restrictions and antisense oligo activity. Giddings *et al* (NAR 2002) presented an artificial neural network model that takes the absence or presence of forty tetramotifs of an oligo as input, and outputs a predicted activity of the oligo. The model correctly predicted 53% of the test instances, evaluated using cross-validation.

In our previous work we presented an approach that included the thermodynamic scores, structural features and motifs, in addition to several other descriptors that helped build a more efficient predictive model of oligo activity. Moreover, we used methods that produce human-readable output in the form of a hierarchical tree or a set of rules, suggesting the important factors in oligo activity classification. Building on previous work, we present a model built using our flex motif approach that use ambiguity codes in certain locations of the motif to reduce the attribute space and increase its potency. For example, the TYYC flex motif would allow C or T in the second and third location, a T in the first, and a C in the fourth. In order to preserve the predictive ability of fixed motifs, we searched for a minimal outer cover of the motifs. The flex motifs were able to increase our model performance and produce a model that correctly predicts 70% of the test instances, evaluated using 10-fold cross-validation. Compared to state-of-the-art model in the literature, this is an increase of model performance of more than 32%. Compared to the less than 10% standard success rate estimated in the literature, our

approach suggest a possible 7-fold reduction in the *in vivo* screening needed to discover an active oligo.

9:50 a.m. *Use of Dynamic Warping for the Alignment of DGGE Gels*, **David Baird**, AgResearch, Lincoln, New Zealand

Denaturing Gradient Gel Electrophoresis (DGGE) is a method for the separation of nucleic acids like DNA or RNA, and the analysis of proteins. The samples are placed on the edge of the gel and drawn through it by an electric field. The DGGE gel has a gradient of denaturing compound that at higher concentrations causes the DNA to melt, changing its rate of progress through the gel. Small changes in the DNA (as little as a single base substitution) cause the DNA to melt at different concentrations of denaturant, allowing separation of very similar DNA fragments. The resulting distribution of the DNA fragments is read using radioactive isotopes, fluorescent dyes in the samples, or staining with silver. This gives the familiar banded image, with each sample forming a lane in the gel. Issues which must be addressed in any statistical analysis are the alignment of the lanes to give consistent band positions, as often the rate that the lanes progress varies across the gel, and the lanes can diverge from parallel tracks. The use dynamic warping (the Viterbi algorithm) can be used to align the lanes so that the peak positions occur together. This talk will illustrate this method along with an approaches to identify the peaks in the intensity along the lanes (a peak = a band on the image), and the subsequent analysis of the peak data.

10:10 a.m. – 10:30 a.m.

BREAK

10:30 a.m. – 12:10 p.m.

C-4

DNA, the Lasso, and Variable Selection

10:30 a.m. *Drug Target Prediction Using Simultaneous Equation Models of Gene Regulatory Networks*, **Yingchun Zhou**, **Elissa Burk**, **Tim Gardner**, **Eric Kolaczyk**, Boston University

A major challenge in the development of new therapeutic drugs is the identification of the molecular targets of drug compound candidates. We consider the problem of identifying such candidates from DNA microarray data. Our approach is to cast the problem as one

of identifying outliers in a large, sparse system of simultaneous equations, describing the influence of both the underlying gene regulatory structure and the external effects of the potential drug candidate. Inference is conducted in two stages: (i) a LASSO-type regression for extracting the influence of the gene regulatory network, and (ii) a residual analysis for outlier detection. We present a variety of empirical results demonstrating the capabilities of our method.

10:55 a.m. *GLARS: S-PLUS and R package for Generalized Least Angle Regression*, **Tim Hesterberg**, Insightful Corp.

Least Angle Regression is a promising new technique for variable selection applications, offering a nice alternative to stepwise regression. It provides an explanation for the similar behavior of Lasso (L1-penalized regression) and forward stagewise regression, and provides a fast implementation of both. We'll demonstrate a prototype open-source S-PLUS/R package "glars" for generalized least angle regression, extending the work by outside collaborators. See www.insightful.com/Hesterberg/glars

11:20 a.m. *Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm*, **Markus Kalisch, Peter Bühlmann**, ETH, Zürich, Switzerland.

We consider the PC-algorithm for estimating the skeleton and equivalence class of a very high-dimensional directed acyclic graph (DAG) with corresponding Gaussian distribution, and we use it for inferring directed associations among biomolecular variables. The PC-algorithm is computationally feasible and often very fast for sparse problems with many nodes, i.e. variables, and it has the attractive property to automatically achieve high computational efficiency as a function of sparseness of the true underlying DAG. We prove uniform consistency of the algorithm for very high-dimensional, sparse DAGs where the number of nodes is allowed to quickly grow with sample size n , as fast as $O(n^a)$ for any $0 < a < \infty$. The sparseness assumption is rather minimal requiring only that the neighborhoods in the DAG are of lower order than sample size n . This setting is particularly attractive for biological applications, where the number of variables is in the thousands but the number of samples might be much less. We demonstrate the algorithm on a biological data set from *Bacillus Subtilis* with thousands of gene expressions and an additional continuous response variable.

11:45 a.m. *The Group and Smoothed Lasso for High-Dimensional Data*, **Lukas Meier, Peter Bühlmann**, ETH Zürich, Switzerland

Variable selection and the analysis of associations between variables is an important subject in many areas of bioinformatics. We propose two extensions of the popular Lasso and demonstrate their usefulness for DNA splice site detection and motif finding.

When dealing with categorical variables with higher-order interaction terms, models can become very high-dimensional. We propose to use the Group Lasso penalty with GLM models and present an efficient algorithm which is particularly suited for high-dimensional problems. The method is shown to be statistically consistent for sparse but high-dimensional problems, i.e. we allow the number of predictor variables to be much

larger than sample size. Moreover, we present a two-stage modification which allows for hierarchical model fitting. The methods are illustrated on the problem of DNA splice site detection where the predictors are (short) DNA sequences. When we are faced with different data sets, we can often get better estimators through a (weighted) combination of the data. This approach is particularly well-suited for time-course data. We propose the Smoothed (Adaptive) Lasso and show improved convergence rate and asymptotic consistency for variable selection. The success of the method is illustrated on a problem of motif finding.

12:10 noon – 2:00 p.m.

LUNCH

2:00 p.m. – 3:40 p.m.

C-5

Mutations, Metatarsals, Metabolomics, Metabolics, and Maps

2:00 p.m. *Maximum Likelihood Estimation of Mutation Rates Under the Haldane Model*, **Qi Zheng**, Texas A&M University System School of Rural Public Health

Estimation of mutation rates is an integral part of many scientific disciplines. The fluctuation test protocol proposed by Luria and Delbruck in 1943 remains the basic approach to experimentally measuring mutation rates. Several mathematical models have been proposed for estimating mutation rates under the fluctuation test protocol; this presentation focuses on a well-known model suggested by J.B.S. Haldane. Due to perceived mathematical difficulties, statistical inference about mutation rates under the Haldane model has been primarily based on statistical moments. The perceived difficulty was caused by the fact that the likelihood function possesses no closed form expressions. This talk outlines a novel computational approach to overcoming this difficulty. The proposed approach enables one to compute both point and interval estimates of mutation rates based on the likelihood function.

2:20 p.m. *Using a Triangle to Diagnose and Evaluate a Skeletal Deformity*, **Philip H. Demp**, Temple University

Hallux Abducto Valgus is a skeletal deformity characterized by a lateral deviation of the great toe with respect to the first metatarsal bone and by an abnormal position of the first metatarsal head (bunion deformity). A rectangular coordinate system was selected and (X,Y) coordinates for the three positions of the 1st, 2nd and 5th metatarsal heads were obtained. The three positions were considered as vertices of a triangle. A similarity (linear) transformation was performed in the complex plane. The 2nd and 5th vertices were mapped into the fixed positions (0,0) and (1,0) and the 1st vertex was mapped into a position (X,Y) such that the shape of the original triangle is preserved. In this way, all triangles are standardized with their differences in shape corresponding to unique points in the position of the 1st vertex. By applying the statistical method of cluster sampling in this pilot study, a strong discrimination of diagnostic values among points of non-deformity, mild to moderate deformity and severe deformity were found. All coordinates were obtained from dorsoplantar, weightbearing radiographs of the foot. The shape of a standardized triangle, given by a uniquely derived (X,Y) point position, can be used by the surgeon for a preoperative evaluation.

2:40 p.m. *Detecting Peaks in Gas and Liquid Chromatography Mass Spectrometric Metabolomics Data*, Michael Lawrence, Heike Hofmann, Suh-yeon Choi, Dianne Cook, Olga Nikolova, Eve Wurtele, Iowa State University

Gas or liquid chromatography mass spectrometry (GC-MS or LC-MS) is a common method for the high-throughput measurement of a metabolome. A metabolome is the collection of metabolites in a biological sample. The presence of a metabolite is represented by a sharp peak in mass spectral intensity along the chromatographic time domain. The area under the peak is an indicator of the quantity of the metabolite. Random and background noise and peak convolution interfere with the detection of metabolite peaks. Data preprocessing is a necessary and difficult step in the detection and quantification of metabolite peaks. We present methods for robustly denoising the time series, detecting peaks, and quantifying metabolite levels in LC-MS and GC-MS data. The noise estimation relies on loess smoothing, and the peak detection is based on a filtering approach that fits distribution functions to the peaks.

3:00 p.m. *Indexing and Searching Databases of Metabolic Pathways*, Robert Grossman, University of Illinois at Chicago and Open Data Group, **Xiangjun Liu**, University of Illinois at Chicago, and **Greeshma Neglur**, University of Illinois at Chicago

In prior work, we introduced an algorithm for assigning unique keys to chemical compounds. We also showed that by using these keys, data quality problems in various online databases of chemical compounds could be identified and corrected. For example, the National Cancer Institute (NCI) online database of chemical compounds contains approximately 235,000 structures, but approximately 14% of them have duplicate entries. Finally, we also showed that commonly used identifiers for chemical compounds, such as UNIQUE SMILES, are not in fact unique.

In this talk, we extend this work to an algorithm that computes keys for metabolic pathways. We also describe how this algorithm can be used as a basis for a system that searches across online databases of metabolic pathways, such as KEGG and MetaCyc.

3:20 p.m. *Strategies for Genome Scans in Complex Pedigrees Using Dense Marker Maps*, Natascha Vukasinovic, Monsanto Company, Animal Genomics and Breeding, St. Louis, MO

Advances in genome sequencing and high-throughput genotyping have stimulated use of dense marker maps in genome scans and fine mapping of genes that influence quantitative traits (quantitative trait loci or QTL). So far, two general approaches have been used to map quantitative trait loci (QTL) in pedigrees: linkage analysis (LA) and linkage disequilibrium (LD). The LA approach utilizes information on recombination events between genetic markers in a pedigree. The LD approach utilizes information on historical recombinations. Recently, a method that combines linkage analysis and linkage disequilibrium information, the LA/LD method, has become common for genome analysis. This method is based on predicting the Identity-By-Descent (IBD) probability between all haplotypes of the individuals in a pedigree at the putative QTL position. The calculation of IBD probabilities is based on similarity of markers in a haplotype. The matrix of IBD probabilities is used as a correlation matrix between haplotypes and the variance associated with the putative QTL can be estimated by maximum likelihood methods. However, with complex pedigrees and several hundreds or even more markers on a chromosome, genome scans can become computationally very tedious. In this work, we explore strategies for optimal genome screening and QTL fine mapping in complex pedigrees using a very dense marker map. We apply the LA/LD method to a simulated complex multigenerational pedigree and investigate accuracy and computational aspects of this method in scenarios involving different sizes of QTL and different haplotype lengths.

3:40 p.m. – 4:00 p.m.

BREAK

4:00 p.m. – 5:40 p.m.

C-6

Gene Expression Data Analysis

4:00 p.m. *Genome-Wide Co-Expression Based Prediction of Differential Expressions*,
Yinglei Lai, George Washington University

Microarrays have been widely used for various disease studies to detect novel disease related genes. They enable us to study differential gene expressions at a genomic level. They also provide us with informative genome-wide co-expressions. Although numerous methods have been proposed for identifying differentially expressed genes, genome-wide co-expressions have not been well considered for this issue. Incorporating genome-wide co-expression information in the differential expression analysis may improve the detection of disease-related genes.

In this study, we propose a statistical method for predicting differential expressions through the local linear regression between differential expression measures and expression correlation measures. The smoother span parameter is determined by optimizing the correlation between the observed and predicted differential expression measures. A mixture normal quantile based method is used to transform data. We use the gene-specific permutation procedure to evaluate the significance of a prediction. Two published microarray data sets were analyzed for applications. For the data set collected for a prostate cancer study, the proposed method identified numerous genes with weakly differential expressions. These genes have been shown in literature to be associated with the disease. For the other data set collected for a type 2 diabetes study, no significant genes could be identified by the traditional methods. However, the proposed method identified many genes with significantly low false discovery rates.

4:15 p.m. *A Framework for Learning Classification Models Using Whole Genome DNA Copy Number Data*, Jie Cheng, Joel Greshock, Tal Zaks, Kwan Lee,
GlaxoSmithKline

A simple and robust framework that integrates data preprocessing, feature selection and model building is proposed from analyzing high dimensional DNA copy number data. There are four steps in finding a classification model: (1) Use data cut-off threshold (d) to convert small values to zero; (2) Filter out features with low mean difference (m) between the two phenotypic groups; (3) Use 10,000 permutations to filter out features with insignificant p values (p) of mean difference. (4) Use the remaining features to construct a simple weighted voting model (no parameters).

To find optimal models, grid search method is used to search the parameter space of (d , m , p). For each combination of parameters, leave one out cross validation (including the feature selection steps) is performed and area under the ROC curve is calculated. The confidence interval of the AUROC of the optimal model is estimated using 500 bootstrapping. Statistical significance of the optimal model is measured by random class label permutation 500 times. Experimental results on a drug efficacy study in oncology

show that the proposed approach can effectively discover important biomarkers targeted by this drug.

4:30 p.m. *Gene Expression Data Analysis by Chemometric Methods*, David X. Zhu, Richard J. Goeke, David L. Baker, James H. Hamburg, David E. Booth, Stephane E. Booth, Kent State University

Since the beginning of the Human Genome Project, it has been known that the knowledge of the base pair sequence of the human genome would be helpful for both the diagnosis and development of treatment modalities for human cancers. The present review considers two questions. First, how can chemometric methods for the analysis of gene expression (microarray) experimental data be used with the fact that genes cause cancer (Varmus, 1993) to presumptively identify biochemical pathways of disease as well as help in diagnosis. Secondly, how can such information be used to develop new treatment modalities? We will give an example of an analysis of the first type and discuss recent approaches on the literature for the second.

4:45 p.m. *Learning gene association networks from high dimensional data*, Jie Cheng, Kwan Lee, Discovery Analytics, GlaxoSmithKline

Inferring (reverse-engineering) large-scale gene networks from limited continuous data is a challenging problem in bioinformatics. This problem is also closely related to covariance matrix estimation in Statistics. We propose a novel approach to learn network structures based on our earlier work on Bayesian network learning from multinomial data. The basic idea is to learn the network structure in three phases: drafting, thickening and thinning. In drafting phase, marginal correlations are used to generate an initial guess of the structure. In thickening and thinning phases, low order partial correlations are used to constantly modify the network in a parsimony manner (in terms of both the number of tests and the order of the tests), based on the analysis of the network structure at that moment. Simulation data and public functional genomics data are used to evaluate the performance of the proposed method. The comparisons with other competing approaches (such as shrinkage based method GeneNet) are also given.

5:00 p.m. *Tests for the Hypothesis of Exchangeability*, Reza Modarres, George Washington University

We discuss the hypothesis of bivariate exchangeability and consider test statistics based on the ordering of the Euclidean interpoint distances. The runs test of exchangeability is an adaptation of the well-known multivariate runs test. The nearest neighbor test of exchangeability is based on the number of nearest neighbor type coincidences after the observations are folded. The rank test of exchangeability compares the within and between ranks of the interpoint distances. We also consider the sign test of exchangeability, which uses the sign of the observations in specific regions, and a bootstrap test of exchangeability based on the maximum distance between the mirror images. We compare the power of these methods in a Monte Carlo study which shows different power orderings of the methods, depending on the alternative hypothesis.

5:15 p.m. *Learning of Regulatory Modules and Predictive Models of Global Transcriptional Dynamics: Application to the Extremophile Halobacterium NRC-1*, Richard Bonneau, Nitin Baliga, Marc Faciotti, Vesteinn Thorsson, David Reiss, Courant Institute, New York University

Our system for network inference and modeling consists of three major components: cMonkey (a method for learning co-regulated biclusters and pathways), the Inferelator (regulatory network inference). We describe our application of these methods to *Halobacterium* and several other model organisms. This effort represents one of the first coordinated functional genomics efforts in archaea and in particular, under hypersaline conditions.

We have described an algorithm, the *Inferelator*, which infers regulatory influences for genes and/or gene clusters from mRNA and/or protein expression levels. The procedure can simultaneously model equilibrium and time-course expression levels, such that both kinetic and equilibrium expression levels may be predicted by the resulting models. Through the explicit inclusion of time, and gene-knockout information, the method is capable of learning causal relationships. It also includes a novel solution to the problem of encoding interactions between predictors. We discuss the results from an initial application of this method to the halophilic archaeon, *Halobacterium NRC-1*. We have found the network to be predictive of 150 newly collected microarray datasets and have also validated parts of the network using ChIP-chip. This network offers a means of deciphering how this intensely tough organism maintains homeostasis and responds to wide varieties of metabolic, genetic and environmental states.

Saturday, 26th May 2007

8:30 a.m. – 10:10 a.m.

C-7

Clustering and Mixture Models

8:30 a.m. *Models for the Analysis of Antarctic Soil Samples*, Lidia Rejto, Alfred Renyi Mathematical Institute of the Hungarian Academy of Sciences, Budapest, Hungary, and University of Delaware, **Gabor Tusnady**, Alfred Renyi Mathematical Institute of the Hungarian Academy of Sciences, Budapest, Hungary, **Craig Cary**, University of Delaware, **Julie Smith**, University of Delaware

In this paper, we show methods to analyze data from Antarctic Dry Valley soil samples. Soil samples were divided for analysis of both the microbial community and physiochemical soil qualities such as soil moisture, pH, conductivity, ammonia, nitrate, and carbon. Genomic DNA was extracted and amplified to generate bacterial community profiles by automated ribosomal intergenic spacer analysis (ARISA). Modern molecular techniques such as ARISA allow us to estimate community diversity of the mostly unculturable bacteria found in natural environments. ARISA is a DNA fingerprinting method based on length heterogeneity in the internal transcribed spacer (ITS) region between ribosomal genes. A presence/absence table was generated for all of the ARISA fragment sizes found in the 16 transects. Each soil transect was measured at 5 locations close to each other. 725 organisms with different sizes were found in the 16 transects. In this paper, we build up parametric stochastic models for the ARISA data and then estimate the parameters with the help of the soil chemistry variables.

There are papers dealing with soil chemistry and ARISA data. The data set we are coming from different conditions and using traditional factor analysis and clustering methods. The unique feature of this analysis is that it provides a novel method to describe bacteria presence/absence in the different sites with the help of the soil chemistry data.

8:45 a.m. *Detection of Syntenic Regions in Bacterial Genomes Through Statistical Clustering*, Siew-Ann Cheong, Cornell University, **Paul Stodghill**, **Dave Schneider**, USDA Agricultural Research Service, **Chris Myers**, Cornell University

Syntenic regions are groups of successive genes found in the same order within different bacterial strains and species. These extended genomic elements, as well as the lineage-specific regions between them, offer important insights into the gene regulatory networks and evolutionary histories of closely related bacteria. Presently, syntenic regions are identified by determining the best bipartite match between BLAST or FASTA homologs of genes in the various sequences, or from whole-genome alignments obtained using

MUMmer, MAUVE, or WABA. In this talk, I will describe an alternative method for synteny detection that do not require sequence alignments. Instead, I start from statistical segmentations of the sequences to be compared, and hierarchically cluster the segments across different sequences.

9:00 a.m. *Bayesian Modeling and Inference in Single Cell Dynamic Networks*, **Jarad Niemi**, Duke University

Biologists are beginning to measure protein levels within individual cells using time-lapse fluorescent microscopy. The fluctuations in these protein levels are described by complex dynamic networks typically represented by differential equations. We present some aspects of our work on creating discrete-time stochastic models based on these differential equations. These models include components accounting for intrinsic biological randomness, noise, and measurement error. We then discuss Bayesian methods for parameter estimation in these models, with examples from synthetic gene circuit studies.

9:15 a.m. *Multi-Level Mixture Models - Flexible Clustering of High-Dimensional Biological Data*, **Rebecka Jornsten**, Rutgers University

Mixture models have become immensely popular tools for analyzing high-dimensional and heterogeneous biological data sets, e.g. for clustering of gene expression data. The outcome of clustering analysis is often interpreted in a subjective manner, such as clusters appearing to represent a particular shape or profile (constant or increasing).

In this talk, we present a multi-level sparse parametrization of the mixture model components. This representation allows for an objective and direct comparisons between components or clusters in multi-factor experiments, and can in addition cluster the data using several distance metrics simultaneously (e.g. euclidean, 1-|correlation|). The levels in the multi-level parametrization can correspond to factor levels in a multi-factor experiment, or a data transformation (such as standardization, which links euclidean distance to a correlation-based distance metric).

We motivate the use of a sparse multi-level parametrization from an efficiency, or budgeting standpoint. If cluster profiles substantially overlap for a subset of experimental factors/variables, modeling these clusters separately is a waste of parameters. If we pool clusters for the overlapping portion of variables (e.g. shape, or one factor level), we both save parameters and obtain better estimates, since a larger number of genes contribute to the common cluster profile estimates. Using this more efficient data description, we can re-allocate the parameters (or complexity budget) to where they are better needed - for example in detecting other clusters with more distinct patterns.

We apply the multi-level mixture modeling framework to analyze several gene expression studies on differentiating stem cell-lines, conducted at the Keck center for collaborative neuroscience at Rutgers University. The multi-level mixture models provide interpretable clustering results that have in part been experimentally validated.

9:30 a.m. *The k-means decomposition method for identifying biological functions of unknown transcripts in the yeast transcriptome*, **Sungchul Ji**, Rutgers University, **Wonsuk Yoo**, Wayne State University School of Medicine, and **Andrei Zinovyev**, The Curie Institute, Paris

When a frozen solid-phase solution of, say, sodium chloride is heated, the liquid phase containing both water and sodium chloride ions will first form, leaving behind pure solid-phase clusters of water molecules. This happens because the intermolecular forces among water molecules are stronger than those among water and sodium chloride ions. Similarly, when a set of objects are analyzed by a k-means clustering method and the composition of the resulting clusters are examined as a function of k, those objects that are closely related (either structurally or functionally) are expected to remain as clusters longer than those not closely related. Thus, in this analogy, the independent variable k is analogous to temperature T in a phase diagram. For convenience, we may refer to the method of identifying functionally related objects by increasing k in this manner as the ‘k-means decomposition’ method. Equivalent results should be obtainable by changing the k value in the opposite direction, namely, from a large initial value equal to or less than the number of the objects in the set being analyzed toward 1, the minimum possible k value. Previously we referred to this method as the ‘freezing-out’.

We recently applied the ‘k-means decomposition method’ (implemented by the ViDaExpert software developed by one of us (AZ) to the transcriptome of budding yeast measured with DNA arrays under the condition of glucose-galactose shift at 6 time points (0, 5, 120, 36, 450 & 850 minutes). In analogy to the order parameter used in physics that undergoes an abrupt change at critical points, we defined a novel ‘order parameter’ unique to cell metabolism called ‘transcript density (d_T)’ as the ratio of two numbers – i) the fraction of the total number (i.e., n'/n in Table 1) of transcripts with a given function that remain clustered as k is increased, and ii) the fraction of the total number of the clusters to which the functionally related transcripts remain localized as k is increased. When a double logarithmic plot is made between d_T and k, straight lines were obtained with good p-values, indicating that there exists a power law relation between d_T and k. That is, $d_T = ak^w$ holds, where a is a constant and w is the ‘critical exponent’ whose numerical values were found to depend on the biological functions of the clustered transcripts, varying approximately from $1/4$ to $4/4$.

9:45 a.m. *On Some Computational Issues in Marginal Latent Mixture Analysis*, **Yan Yang**, Arizona State University, **Douglas Simpson**, University of Illinois

Data with bound-inflated responses are common in many areas of application. Often the data are bounded below by a real number (e.g., zero) with excess observations at the boundary value. We consider a general class of latent mixture models for inflated discrete and semi-continuous data that combines a degenerate distribution at the bound and a discrete or censored distribution that can also produce observations equal to the bound. The latency resulting from not being able to identify which distribution has generated a boundary value leads to a pseudo-likelihood for correlated bound-inflated data that

cannot be factorized. We implement both the EM and Quasi-Newton algorithms to estimate the class of mixture models and compare the two methods. The asymptotic covariance matrix is adjusted by the sandwich estimator using the theory of generalized estimating equations. The methods are illustrated with an ultrasound safety study in laboratory animals.