Statistical and Computational Challenges in Omics Data Integration (SCC-ODI) Workshop

February 16-17, 2015

SPEAKER TITLES/ABSTRACTS

Keith Baggerly
MD Anderson

"Integrating Data, Assays, and Pathways: the 20,000 Foot View"

The "big data" revolution impacting biology for the past 15 years has been driven by the incredibly rapid development of assays capable of simultaneously measuring some type of "expression levels" for thousands of "genes" at once. Since these assays were initially expensive to run, these assays in turn sparked broad interest on the analytical side in "large p, small n" analyses, which were inverted from problems facing statistics for most of the 20th century. The hope (and hype) was that examination of the patterns revealed would quickly lead to more accurate diagnosis and treatment of disease.

Today¹s situation is rather different. Many assays have now reached the point in terms of stability and cost that datasets of hundreds of samples or more are not uncommon, so n is no longer all that "small". Further, in many cases the same samples have now been examined with multiple assays. Since dramatic improvements in patient care have not quickly emerged from single analyte studies, the hope is now partially shifting to ensembles of assays or more sophisticated (and larger-scale) use of individual ones (e.g., perturbation and heterogeneity studies).

In this talk, we will survey some of the types of assay data now available, where to get the data, and some constraints on their use (TCGA and CCLE will be used as examplars here). We will give some examples of basic data integration involving pairs of assays, and then touch on larger aggregates. We will also describe some initial attempts to exploit existing biological knowledge of coherent subsets of genes (pathways) to ask more focused questions.

Alberto Cassese
Rice University

"A Bayesian Model for the Identification of Differentially Expressed Genes in Daphnia Magna Exposed to Munition Pollutants"

Biological effects of water purification is a novel interesting topic arising in the biological community. In this talk I will focus on a Bayesian hierarchical model developed for the identification of differentially expressed genes in organisms exposed to chemical compounds in water. We have data on Daphnia Magna organisms exposed to a serial purification system that

comprises four consecutive purification stages of progressively more contaminated water. In particular, we model the expected expression of a gene in a pond as the sum of the mean of the same gene in the previous pond plus a gene-pond specific difference. Differential expression is identified employing a variable selection mechanism with prior probability that accounts for available information on the concentration of chemical compounds present in the water. We perform posterior inference via MCMC stochastic search techniques. In the application, we reduce the complexity of the data by grouping genes according to their functional characteristics, based on the KEGG pathway database. This also increases the biological interpretability of the results. Our model successfully identifies a number of pathways that show differential expression between consecutive purification stages. We also find that changes in the transcriptional response are more strongly associated to the presence of certain compounds, with the remaining contributing to a lesser extent. We discuss the sensitivity of these results to the model parameters that measure the influence of the prior information on the posterior inference.

Junxiao Hu
University of Colorado Denver

"Exploratory Study of Gene Expression in COPD Using Network Analysis and Kernel Machine Methods"

To identify molecular pathways for Chronic Obstructive Pulmonary Disease (COPD) related traits and to characterize new subtypes, several different omics data sets on a subset of subjects were collected from the COPDGene genetic epidemiology study. Our group has performed analyses on each of the individual data types (gene expression, biomarker, metabolites) to identify COPD related features, along with analyses of pairwise interactions (e.g., between genes and metabolites, genotypes and biomarkers). As an extension of our previous research, we are interested in applying more complex models to the gene expression data to describe the associations between identified pathways and the COPD phenotypes. We hypothesize that a more complex model, which allows for the possibility of nonlinear gene expression effects and more complicated inter-module interactions within proposed gene pathways, will improve the identification of genomic features associated with COPD phenotype signatures. We will present preliminary results of weighted correlation network analysis (WGCNA) for building functional modular networks for COPD related phenotypes. Modules and highly connected hub genes were selected as candidate gene pathways. Then, kernel machine methods were applied to test for nonlinear associations with the COPD phenotypes. Our goal is to extend this framework for multiple dimensional phenotypes and different types of omics data to define pathways.

Yufeng Liu
University of North Carolina

"Sparse Regression Incorporating Graphical Structure Among Predictors"

With the abundance of high dimensional data in various disciplines, sparse regularized techniques are very popular these days. In this talk, we use the structure information among predictors to improve sparse regression models. Typically, such structure information can be modeled by the connectivity of an undirected graph. Most existing methods use this graph edge-by-edge to encourage the regression coefficients of corresponding connected predictors to be similar. However, such methods may require

expensive computation when the predictor graph has many edges. Furthermore, they do not directly utilize the neighborhood information. In this work, we incorporate the graph information node-by-node instead of edge-by-edge. Our proposed method is quite general and it includes adaptive Lasso, group Lasso and ridge regression as special cases. Both theoretical study and numerical study demonstrate the effectiveness of the proposed method for simultaneous estimation, prediction and model selection.

Qiongshi Lu
Yale University

"Post-GWAS Prioritization through Integrated Analysis of Functional Annotation"

We develop a SNP-prioritizing method that integrates p-values from GWAS studies and functional scores predicted by GenoCanyon, a genomic function prediction tool developed in our lab. We apply our statistical framework to the COPDGene GWAS results. Globally, genomic loci associated with COPD or lung function have improved rankings using our prediction score than using p-values. Locally, SNPs with consistently strong signals across different GWAS studies are favored under our framework even if the p-values do not suggest so.

Sandra Safo,
Emory University

"Sparse Analysis of High Dimensional Data with Application to Data Integration"

A core idea of most multivariate data analysis methods is to project higher dimen- sional data vectors on to a lower dimensional subspace spanned by a few meaningful directions. Many multivariate methods, such as canonical correlation analysis (CCA), and principal component analysis (PCA), solve a generalized eigenvalue problem. We propose a general framework, called substitution method, with which one can easily obtain a sparse estimate for a solution vector of a generalized eigenvalue problem. We employ the idea of direct estimation in high dimensional data analysis and suggest a flexible framework for sparse estimation in all statistical methods that use generalized eigenvectors to find interesting low-dimensional projections in high dimensional space. We illustrate the framework with sparse CCA for joint analysis of two sets of variables to study the idea that changes in one set of variable may be associated with the second set of variable, and vice-versa. Our method makes no assumptions of the underlying covariance structures, it can be used for very high dimsional problems, and conver- gence to a stationary point is guaranteed. We compare our method to existing sparse CCA approaches via simulation studies and real data analysis using gene expression and copy number variation data from a breast cancer study.

Ali Shojaie
University of Washington

"Network-Based Pathway Enrichment Analysis of Omics Data"

Biological networks provide valuable insight into behavior of biological systems. They also offer new clues into mechanisms of initiation and progression of complex diseases. I will describe a flexible framework for network-based pathway enrichment analysis of omics data in diverse applications. I will then discuss some of the challenges in analysis of networks of multiple types of omics data, as well as new directions for network-based analysis of diverse omics data types.

Francesco Stingo
MD Anderson

"A Bayesian Approach to Biomarker Selection through Mirna Regulatory Networks"

The availability of cross-platform, large-scale genomic data has enabled the investigation of complex biological relationships for many cancers. Identification of reliable cancer-related biomarkers requires the characterization of multiple interactions across complex genetic networks. MicroRNAs are small non-coding RNAs that regulate gene expression; however, the direct relationship between a microRNA and its target gene is difficult to measure. We propose a novel Bayesian model to identify microRNAs and their target genes that are associated with survival time by incorporating the microRNA regulatory network through prior distributions. We assume that biomarkers involved in regulatory networks are likely associated with survival time. We employ non-local prior distributions and a stochastic search method for the selection of biomarkers associated with the survival outcome. Using simulation studies, we assess the performance of our method, and apply it to experimental data of kidney renal cell carcinoma (KIRC) obtained from The Cancer Genome Atlas. Our novel method validates previously identified cancer biomarkers and identifies biomarkers specific to KIRC progression that were not previously discovered.

Jiehuan Sun
Yale University

"Discovery of Novel Loci Associated with COPD by Pulling Information from Case-Control Status, Related Clinical Feature, and Functional Annotation"

In the COPDgene project, standard GWAS using case-control status has been conducted to identify genetic variants associated with COPD, but only a small number of loci have met statistical significance. The COPD patients were followed longitudinally with their clinical information recorded. Most clinical information can be, but has not been, employed to infer loci associated with COPD. Here, we will use the number of exacerbations for each COPD patient over time as a longitudinal trait to find loci that might be associated with COPD, together with case-control status and functional annotation.

Xiting Yan
Yale University

"Identifying Disease Heterogeneity from Gene Expression Data by Integrating Pathway Information in Asthma Patients"

It is increasingly evident that the pathobiologic alterations in asthma are heterogeneous and that analysis of the airway transcriptome can define patterns of gene expression that reveal are clinically meaningful transcriptional endotypes of asthma (TEA clusters). To determine if this is possible, we conducted a novel unsupervised Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway based clustering analysis of gene expression in the induced sputum of adult asthma patients. The identified TEA clusters were correlated to demographical, physiological and inflammatory phenotype of the disease. To validate the associated phenotypes, logistic regression analysis was applied to the expression profiles in matched blood samples, which defined an expression profile in the circulation to determine the TEA cluster assignment in a

cohort of children with asthma. In both cohorts, two of the identified TEA clusters are associated with phenotypes of severe diseases: a history of a near fatal asthma, a history of hospitalization for asthma, and weakly overlap with SARP clusters and TH2 high/low defined disease. This suggests that the TEA clusters are unique and driven by biologic phenomena that are “upstream” or parallel to Th2 inflammation.

Bin Zhu
National Cancer Institute

“Integrating Clinical and Molecular Data for Survival Prediction in TCGA”

Patient survival outcome prediction is an important clinical question. Clinical variables such as tumor stage, age and race are traditionally used to classify each individual into different risk groups. However, without considering the heterogeneity among subjects in the same risk group, the prediction of survival time is inevitably imprecise with large uncertainty. Comprehensive molecular profiling of tumor samples reveals inter-individual heterogeneity in the genomic background and tumor characteristics with unprecedented detail. By integrating clinical and molecular data at the genomic, epigenomic, transcriptomic, and proteomic level, we aim to achieve a more precise survival prediction. We consider a kernel-based approach which facilitates the incorporation of prior knowledge, such as pathway information, and allows the integration of multiple molecular data types with possible multiple views. In this talk, we will outline our research proposal, including imputation of missing molecular data and multiple kernel construction, learning and selection. Some preliminary survival prediction results will be demonstrated using the TCGA lung adenocarcinoma data set.