

# **39<sup>th</sup> Symposium on the Interface: Computing Science and Statistics**

**Theme:  
Systems Biology**

May 23–26, 2007

DoubleTree Hotel  
Philadelphia, Pennsylvania

## **PROGRAM CHAIR**

**Alan J. Izenman  
Temple University**

## **PROGRAM CO-CHAIR**

**Zoran Obradovic  
Temple University**

The Interface Symposium is hosted by the **Center for Statistical & Information Science (Alan J. Izenman, Director)**, the **Department of Statistics**, and the **Center for Information Science & Technology (Zoran Obradovic, Director)**, Temple University.

# INVITED SESSIONS

Thursday, 24<sup>th</sup> May 2007

---

10:30 a.m. – 12:10 p.m.

---

## I-1

### Genetic Association Studies

Organizer: Charles Kooperberg, University of Washington

**10:30 a.m.** *Quantifying and Extracting Information in Whole-Genome Association Studies*, Dan L. Nicolae, University of Chicago

Genome-wide association studies aim to detect the DNA variation modifying disease risk in humans. The challenges in analyzing the data generated by these studies come from the complexity of both the phenotype and of the genetic data. The statistical analyses and the interpretation of results from whole-genome studies can be greatly improved by detailed information on how the SNPs included in the genotyped set interrogate all known variation across the genome. I will describe methods for efficiently extracting the information available in the genotype data and in reference databases.

**11:00 a.m.** *An Adaptive Clustering Algorithm to Model Haplotype-Disease Association in Case-Control Association Studies*, James Y. Dai, University of Washington

A haplotype is the combination of closely linked markers on a chromosome. It has been shown that incorporating information from adjacent markers is more effective than modeling single-locus association. However haplotype analyses tend to require a fairly large number of parameters when many markers are involved, thereby losing power to detect disease association. Recently, several clustering algorithms have been proposed to reduce the dimensionality using haplotype genealogy. One limitation of existing algorithms is that the similarity measure between haplotypes does not account for the haplotype-trait association. We propose an adaptive hierarchical clustering algorithm which combines the haplotype similarity and trait association in defining clusters. Simulations and data application are performed to evaluate the proposed algorithm.

**11:30 a.m.** *Patterns of Linkage Disequilibrium Reveal Errors and Deletions in Population Genetic Data*, Paul Scheet, University of Michigan

While average genotyping accuracy for large-scale studies of population genetic variation is now very accurate, a small number of individual single nucleotide polymorphisms (SNPs) may exhibit unusually high error rates, perhaps due to pathological patterns in

the intensities on which genotype calls are based. The segregation of a genomic duplication or deletion (copy number polymorphism; CNP) in the population may present suspicious genotypes at consecutive SNPs in the same individuals. Simple filters, particularly a test for Hardy-Weinberg equilibrium (HWE), are usually applied to attempt to identify such errors and anomalies. These single-marker filters ignore the correlation among genotypes at nearby SNPs (linkage disequilibrium; LD).

In this talk, I will introduce new filters based on identifying SNPs whose genotypes produce unusual patterns of LD. By using a flexible model for variation in a sample of unphased multilocus genotypes, I scan the observed genotype data, evaluating the potential for a marker position to have at least one error by constructing a likelihood ratio (LR) statistic for each SNP in the sample. Additionally, I assess directly the evidence for the presence of deletions. I apply the methods to data from The International HapMap Project as well as to data from a large-scale association study. These methods offer considerable improvement over traditional tests, allowing for the detection of errors and CNPs which were missed by HWE alone, and they will be available in the software package fastPHASE.

**12:00 noon. *Floor Discussion***

## **I-2**

### **Statistical Methods in Proteomics**

*Organizer: Francoise Seillier-Moiseiwitsch, Georgetown University*

**10:30 a.m. *Statistical Preprocessing of Mass Spectrometry Data*, Kevin Coombes, MD**  
Anderson Cancer Center, Houston

Mass spectrometry (MS; especially MALDI or SELDI) is being applied to discover disease-related proteomic patterns in complex mixtures of proteins. A typical MS data set arising in a clinical application contains tens or hundreds of spectra, where each spectrum contains many thousands of intensity measurements. In order to make sense of this volume of data, analysts must perform extensive pre-processing, including calibration, denoising, baseline subtraction, normalization, peak detection, peak quantification, and peak matching. In this talk, I will discuss current approaches for pre-processing MS data in a coherent way.

**11:00 a.m. *Modeling and Spot Detection of Two-Dimensional Gels*, Kimberley Sellers,** Georgetown University

Two-dimensional gel electrophoresis is a powerful tool in the field of proteomics because of its ability to quantify and compare protein abundance for numerous proteins at once. In particular, the Difference Gel Electrophoresis (DIGE) method allow for, say, normal and disease protein samples to be compared on the same gel in order to circumvent much of the systematic variation that exists in traditional two-dimensional gel data. I will discuss results and ongoing work regarding removal of systematic variation existing in

the data, and spot detection and quantification of protein spots. The goal here is to provide an automated procedure that accounts for and removes artifacts introduced in the experimental and imaging process, and to better detect and quantify protein spots.

**11:30 a.m. *Analysis of 2D-Gels: A Global Approach*, Valeriy Korostyshevskiy, Georgetown University, & Françoise Seillier-Moiseiwitsch, Georgetown University**

Two-dimensional polyacrylamide gel electrophoresis is currently one of the techniques of choice to separate and display all the proteins expressed in a tissue. In the resulting protein maps for groups of patients, we seek to identify proteins that are differentially expressed. I will describe a comprehensive analytical approach that deals with preprocessing, alignment and differential analysis. Preprocessing removes the bulk of the background noise. It involves smoothing, selecting regions containing spots and gradient thresholding. Images are corrected to account for current leakage and subsequently aligned using cubic-spline transformations. The alignment is formulated as a quadratic programming problem that is optimized using an interior-point method. Wavelets are then utilized to summarize the aligned images, and statistical tests performed on the wavelet coefficients. These novel statistical tests were developed with the experimental design and the low sample size in mind.

**12:00 noon. *Floor Discussion***

## **I-3**

### **Problems and Prospects for Integrative Biology and the Grid**

*Organizers: J. Robert Beck, Fox Chase Cancer Center, Philadelphia*

**10:30 a.m. *Good Cells, Bad Cells, and the Grid*, J. Robert Beck, Fox Chase Cancer Center**

Systems biology deals with the computational modeling of networks of interacting components. Although much of the work to date in this discipline has focused on molecular processes, cellular organization is particularly amenable to analyses utilizing systems techniques. Because of its inherent complexity, cellular systems research is also appropriate for study using Grids: distributed computational, data, and community resources that can be brought to bear concurrently on biomedical problems. In this presentation we will illustrate examples of all three types of biomedical Grids applied to topics in cancer research.

**11:00 a.m. *Trustable Architectures for a Biomedical Research Grid*, Frank J. Manion, Fox Chase Cancer Center, Philadelphia**

One of the goals of the Grid community is to enable sharing of resources at a large scale in self assembling communities termed "Virtual Organizations". Sharing of biomedical research data in a grid environment, particularly data derived from research involving

human subjects, presents specific challenges in diverse areas. These range from supporting systems needed by the regulatory community, as well specifics of grid services and applications. In this talk, we will describe recent work with the regulatory community in deriving their perceived requirements, and discuss possible approaches for future grid security systems.

**11:30 a.m. *Parallel Implementation of Nonnegative Matrix Factorization (NMF) Algorithms using High-Performance Computing (HPC) Cluster*, Karthik Devarajan and Guoli Wang, Fox Chase Cancer Center, Philadelphia**

NMF is a powerful method for decomposing a high-dimensional non-negative matrix  $V$  into two matrices with non-negative entries,  $V \sim WH$ . In the context of a  $p \times n$  gene expression matrix  $V$  from  $n$  samples on  $p$  genes, each column of  $W$  defines a metagene and each column of  $H$  represents the metagene expression pattern of the corresponding sample. In contrast to existing methods, NMF provides a unique parts-based, local representation of the original data. The factorization is based on the Poisson likelihood of generating  $V$  from  $WH$  using the generalized Renyi's divergence of order  $\alpha$  ( $\alpha \neq 1$ ). While the NMF update rules for  $W$  and  $H$  guarantee convergence to a local minimum based on random initial values, the solution may not be unique across runs due to the stochastic nature of initial conditions. We exploited this feature to evaluate the consistency of its performance based on various measures of cluster consensus. In order to assess whether a given rank  $k$  provides a biologically meaningful decomposition of the data, we developed a model selection procedure that utilizes consensus clustering and provides a quantitative evaluation of the robustness of the factorization.

The implementation of the steps in the model selection procedure for any real high-dimensional dataset is computationally very intensive. In order to efficiently apply this method for large-scale biological data from high-throughput studies, we parallelized this algorithm on the HPC cluster using the Message-Passing Interface (MPI)/C++ platform. We also created an integrated package that consists of data input, initialization and model parameterization, model selection, graphical display and output of results via a graphical user interface that communicates between a Windows desktop and the HPC cluster. The interface and connection between the desktop and HPC cluster were built using C#, and the rest of the package implemented in C++ on top of a MPI.

**12:00 noon. *Floor Discussion***

## **I-4**

### **Modelling of Ethanol Abuse and HIV/AIDS**

*Organizer: Yasmin H. Said, George Mason University*

**10:30 a.m. *Developing Integrative Models of the Interaction of Alcohol Use and HIV/AIDS Progression Using Behavioral and Biological "Bedside and Bench" Research*, Kendall J. Bryant & Ron Braithwaite, National Institute on Alcohol Abuse & Alcoholism/NIH**

Multiple generalized conceptual models for the effect of past and present alcohol use on a variety of intermediate biological parameters with AIDS outcomes are being developed and replicated in large samples. In general alcohol use is associated with earlier loss of life. These models include rates of viral mutation and resistance accumulation in addition to co-occurring disorders (HCV, tobacco use, etc) to predict survival endpoints. Additional animal laboratory and human observational work suggested that the concept of viral "set-points" early in the initiation of infection is important in predicting rates of progression. Higher viral loads and less active immunological response predicts more rapid progression and earlier time of death. These findings are consistent with animal models when using laboratory feed and exposed primates - with alcohol consuming primates more rapidly progressing and dying in about half the time of sucrose feed animals. With the increased emphasis on integrative models (and the ability to move between the "bench and the bedside" research paradigms) it was suggested that initiation of HIV infection processes while under the influence of alcohol or for individuals with past heavy drinking could change the characteristics of models which predict disease outcomes. Can "new" models which take into account initial set points related to alcohol use, nutritional deficits, and other features of frailty or immunological susceptibility be developed to accurately identify highest risk individuals for rapid progression and death? This presentation will be given in two parts: first, a review of conceptual models being put forward for the interaction of current and lifetime drinking for HIV+ individuals (Kendall Bryant), and then a mathematical, mechanistic model of the specific impact of non-adherence (taking into account viral set points) as an example of translational integrative work (R. Scott Braithwaite).

**11:00 a.m.** *Assessing Interventions Related to Negative Effects of Ethanol on HIV/AIDS Spread*, **Yasmin H. Said & Edward J. Wegman**, George Mason University

Alcohol abuse leads to serious individual and societal outcomes. Among these, we identify irresponsible behavioral outcomes notably promiscuous sexual contacts as well as other violence related outcomes including DWI crashes with fatalities, assault, suicide, murder, sexual assault, domestic violence, child abuse and other. The risky sexual contacts in turn can lead to infections with AIDS, HIV, and other STDs. Social, structural, and environmental factors are major influences on HIV-related behaviors yet the dearth of evaluation tools impedes progress in intervention development. Our research is intended to provide a policy tool for evaluation of interventions using a dynamic agent-based simulation. Alcohol abusers are embedded in a social network that includes the user, family and friends, producers and distributors of alcohol products, law enforcement, remediation and intervention facilities, and treatment facilities, which are coupled to insurance programs. This complex network is reminiscent of more traditional biologic ecology systems. The basic idea is to formulate a model of this network with the goal of exploring short- and long-term interventions that reduce the overall probability of negative outcomes. Intervention approaches will target networks and behavioral settings and provide alcohol users with socially meaningful and rewarding behavioral options that are consistent with valued pro-social identities. Historically, modeling attempts have

focused on specific negative outcomes. The unique feature of this work is that we are attempting to explore the simultaneous reduction of all negative outcomes.

**11:30 a.m. *Using Research Data to Parameterize Risky Sexual Behavioral and Alcohol Use*, W.F. Wieczorek, K.S. Marczyński, T.H. Nochajski & S. Tetewsky,**  
Center for Health and Social Research, Buffalo State, Buffalo, NY

The effort to develop an agent-based model of alcohol-related behavior and acute outcomes is a daunting challenge. The agent-based approach is designed to model the alcohol ecological system, including individuals, population subgroups, activities, locations, alcohol availability, and acute outcomes (both acute problems and benign). The model utilizes a directed graph approach to estimate acute outcomes for every member of the population (i.e., agents) over a single day. The model is run repeatedly to develop estimated outcomes for an entire year; a stochastic component ensures individual-level estimates so that all members of a population subgroup do not have the same outcome. A major challenge in developing this type of model is the large number of parameters that need to be estimated for the various subgroups. Epidemiological data on drinking and acute outcomes is not usually collected to inform this type of model. Therefore, it is a challenge to extract reasonable parameter estimates from the epidemiological data currently available. One topic of interest in this agent-based model is alcohol-related risky sexual behavior. Data on risky sexual behavior among DWI offenders, a group known for high quantity episodes of alcohol use, will be presented. The sample is 822 male and female DWI offenders recruited from courts who provided in-depth interviews including details of risky sexual behaviors and sexually transmitted diseases. The offenders also provided blood samples that were tested for a history of hepatitis B infection (HBV). Hepatitis B is usually considered an STD. The combination of the frequency of risky sexual behavior will be linked with the self-reports of STDs and the HBV information to estimate marginal probabilities to parameterize the model.

**12:00 noon *Floor Discussion***

---

**12:10 noon – 2:00 p.m.**

**LUNCH**

---

---

2:00 p.m. – 3:40 p.m.

---

## I-5

### Statistical Methods for Genetic Association Studies

Organizer: Gang Zheng, National Heart, Lung, and Blood Institute, Bethesda, MD

**2:00 p.m.** *Aspects of the Statistical Multiple Comparisons Problems in Genetic Association Studies of Sporadic Disease*, **Philip Rosenberg**, National Cancer Institute, Bethesda, MD

In case-control studies of unrelated subjects, gene-based hypothesis tests consider whether any tested feature in a candidate gene - single nucleotide polymorphisms (SNPs), haplotypes, or both - are associated with disease. Standard statistical tests are available that control the false-positive rate at the nominal level over all polymorphisms considered. However, more powerful tests can be constructed that use permutation resampling to account for correlations between polymorphisms and test statistics. A key question is whether the gain in power is large enough to justify the computational burden. We compared the computationally simple Simes Global Test to the min P test, which considers the permutation distribution of the minimum  $p$ -value from marginal tests of each SNP. In simulation studies incorporating empirical haplotype structures in 15 genes, the min P test controlled the type I error, and was modestly more powerful than the Simes test, by 2.1 percentage points on average. When disease susceptibility was conferred by a haplotype, the min P test sometimes, but not always, under-performed haplotype analysis. A resampling-based omnibus test combining the min P and haplotype frequency test controlled the type I error, and closely tracked the more powerful of the two component tests. This test achieved consistent gains in power (5.7 percentage points on average), compared to a simple Bonferroni test of Simes and haplotype analysis.

**2:30 p.m.** *Conditional Likelihood Methods for Haplotype-Based Association Analysis Using Matched Case-Control Data*, **Jinbo Chen**, University of Pennsylvania

I will present two conditional likelihood approaches for estimating haplotype-related odds ratio parameters using SNP genotype data from matched case-control studies, where controls are individually matched to cases on some selected factors. These approaches address the issue that haplotypes cannot be inferred with certainty from SNP genotype data (phase ambiguity). One approach is based on the likelihood of disease status conditioned on the total number of cases, genotypes, and other covariates within each matching stratum, and the other is based on the joint likelihood of disease status and genotypes conditioned only on the total number of cases and other covariates. The joint-likelihood approach is generally more efficient, particularly for assessing haplotype-environment interactions. Estimates from the joint-likelihood approach are obtained by a "one-step approximation" to the standard Newton-Raphson algorithm, with estimates from the first approach used as the starting values.

**3:00 p.m.** *Two-Stage Analyses of Case-Control Genetic Associations*, **Gang Zheng**, National Heart, Lung, and Blood Institute, Bethesda, MD

Three two-stage analyses for population-based case-control association studies will be discussed. All three approaches are based on comparing Hardy-Weinberg disequilibrium coefficients between cases and controls and the genotype frequencies between cases and controls. The first approach is to find two significance levels for the two comparisons between cases and controls, which maximizes the overall power. The second approach is to find the first significance level based on a specific conditional power for the first comparison between cases and controls. Then adaptively determine the second significance level. The third approach is to use the first comparison to determine an underlying genetic model such as recessive, additive or dominant. Then incorporate this model information into the second comparison of genotype frequencies between cases and controls. These three approaches are compared through simulation studies.

**3:30 p.m.** *Floor Discussion*

## **I-6**

### **Proteomics Data Analysis**

*Organizer: Slobodan Vucetic, Temple University*

**2:00 p.m.** *From Diverse Genomics Data to Protein-Protein Interaction Networks*, **Olga Troyanskaya**, Princeton University

Understanding and modeling biological networks is a key challenge in modern systems biology. Broad availability of diverse functional genomic data should enable fast and accurate generation of network models through computational prediction and experimental validation. I will discuss our recent work on developing systems that predict biological networks and pathways from heterogeneous experimental results. Using these systems, we have modeled multiple known processes in *Saccharomyces cerevisiae*, characterized unknown components in these processes through computational predictions and experimental validation, and identified novel cross-talk relationships between biological processes.

**2:30 p.m.** *Enhancing the Analysis of MS Proteomic Profiles Using Prior Knowledge and Past Data Repositories*, **Milos Hauskrecht**, University of Pittsburgh

Data analysis approaches for MS whole-sample proteomics study signals in the MS spectra and use them to build classification models. A typical MS proteomic result reported in the literature includes classification statistics (e.g. accuracy of the model, its Area under the ROC curve) and a set of peak locations that carry the signal responsible for the differences in the sample groups. The identification of protein species responsible for these and other MS signals is either not attempted at all or it is limited to one (or a very few) peak locations, typically, because of the high cost of existing protein id methods. However, the identity of these species is critical for independent validation of

the signals and their clinical utilization as disease biomarkers. High cost of lab-based protein-id solutions prompts the development of computational tools for screening potential differential peak candidates and for eliminating peaks that are unlikely to reveal any new information. In this talk we describe two approaches to achieve these goals: (1) in-silico protein-id methods that let us assign protein labels to peaks in whole-sample MS profiles directly, (2) data-analysis approaches that let us filter out unspecific biomarker candidates using data stored in MS data-repositories.

**3:00 p.m. *Methods for Protein Identification and Quantification From Tandem Mass Spectrometry Data*, Predrag Radivojac, Indiana University**

Shotgun proteomics refers to the use of bottom-up proteomics techniques in which the protein content in a biological sample mixture is digested prior to separation and mass spectrometry analysis. In this talk, I will address our approaches to two of the major challenges in shotgun proteomics, protein identification and label-free protein quantification.

We proposed a new concept of peptide detectability and showed that it could be an important factor in explaining the relationship between a protein's quantity and the peptides identified from it in a high-throughput proteomics experiment. We define peptide detectability as the probability of observing a peptide in a standard sample analyzed by a standard proteomics routine and argue that it is an intrinsic property of the peptide sequence and neighboring regions in the parent protein. To test this hypothesis we first used publicly available data and data from our own synthetic samples in which quantities of model proteins were controlled. We then applied machine learning approaches to demonstrate that peptide detectability can be predicted from its sequence and the neighboring regions in the parent protein with satisfactory accuracy. The utility of this approach for protein quantification is demonstrated by peptides with higher detectability generally being identified at lower concentrations over those with lower detectability in the synthetic protein mixtures. These results establish a direct link between protein concentration and peptide detectability.

In our second approach, we used peptide detectability to address a major challenge in shotgun proteomics, that of the assignment of identified peptides to the proteins from which they originate, referred to as the protein inference problem. Redundant and homologous protein sequences present a challenge in being correctly identified, as a set of peptides may in many cases represent multiple proteins. One simple solution to this problem is the assignment of the smallest number of proteins that explains the identified peptides. However, it is not certain that a natural system should be accurately represented using this minimalist approach. We propose a reformulation of the protein inference problem by utilizing peptide detectability. We also propose a heuristic algorithm to solve this problem and evaluate its performance on synthetic and real proteomics data. In comparison to a greedy implementation of the minimum protein set algorithm, our solution that incorporates peptide detectability performs favorably.

**3:30 p.m. *Floor Discussion***

## I-7

### Genomic Analysis Across Platforms

Organizer: Giovanni Parmigiani, Johns Hopkins University

**2:00 p.m.** *Multiple Laboratory Comparisons of Microarray Platforms*, **Rafael Irizarry**, Johns Hopkins University

Microarray technology is a powerful tool able to measure RNA expression for thousands of genes at once. Various studies have been published comparing competing platforms with mixed results: some find agreement, others do not. As the number of researchers starting to use microarrays and the number of cross platform meta-analysis studies rapidly increase, appropriate platform assessments become more important. Here we present results from a comparison study that offers important improvements over those previously described in the literature. In particular, we notice that none of the previously published papers consider differences between labs. In this talk we present the results of study performed by a consortium of ten labs from the DC and Baltimore area formed to compare three heavily used platforms using identical RNA samples: Appropriate statistical analysis demonstrates that relatively large differences exist between labs using the same platform, but that the results from the best performing labs agree rather well.

**2:30 p.m.** *A Hierarchical Model and R Software for Finding Differential Gene Expression in Multiple High-Throughput Platforms*, **Robert Scharf**, Johns Hopkins University

An increasing number of high throughput gene expression studies are publicly available, many of which sample a common study population with comparable clinical covariates and overlapping scientific aims. By increasing statistical power and broadening the sampled population, combining high throughput gene expression data across studies can be helpful. Our approach to estimating differential expression in multiple studies extends ideas that have worked well in single studies. In particular, statistical models that borrow strength across genes when estimating gene-specific parameters are generally preferable to those that do not because (1) genes are not independent and (2) high throughput technologies typically affect all genes in a similar way. Explicitly acknowledging the within and across study dependence, we develop a hierarchical model for the common gene set in the simple setting of a binary covariate. We test our model by simulation using lung cancer data from three different high throughput technologies. Because concordant and discordant findings may be biologically relevant and plausible, approaches for mining both types of patterns will be discussed.

**3:00 p.m.** *Finding significant large-average sample-variable blocks in high-dimensional data*, **Andrew Nobel**, University of North Carolina, Chapel Hill

Exploratory analysis of gene expression and other high dimensional data often begins with row and column clustering, which yields a partition of the data matrix (or heat map) into disjoint sample-variable blocks. Of particular interest in practice are "monochromatic" blocks whose entries are large (red) or small (green) on average. In

conjunction with clinical and functional annotation, large average blocks are often the starting point for subsequent biological analysis, including the identification of pathways and new disease subtypes.

We describe an efficient algorithm, belonging to the general category of biclustering methods, for identifying large average blocks in high dimensional data. Like other biclustering methods, the algorithm improves on independent sample variable clustering in several respects: the blocks it identifies can overlap and they need not cover the entire data matrix (features that better reflect underlying biology); the inclusion of samples and variables in a block does not depend on their expression values outside the block. The algorithm relies on a simple measure of statistical significance that provides an objective basis for comparing and selecting among submatrices of different sizes and average intensities. We will discuss an application of the algorithm in a multi-platform setting, and will compare its performance to several related biclustering methods.

**3:30 p.m. *Floor Discussion***

## **I-8**

### **Systems Approaches for Alcohol Use Modelling**

*Organizer: Yasmin H. Said, George Mason University*

**2:00 p.m. *Systems Perspectives on Prevention of Alcohol-Related Problems: Research Opportunities and Resources*, Gregory Bloss, NIH/National Institute on Alcohol Abuse and Alcoholism**

Alcohol research affords a wide range of opportunities for applied studies addressing intellectually challenging problems with great social significance. Because alcohol-related behaviors and consequences embody rich interactions among biological, behavioral, socio-cultural, and economic factors, trans-disciplinary and systems perspectives are needed to address prevention of alcohol problems in appropriate multidimensional terms. This presentation identifies the magnitude and significance of alcohol-related problems using several alternative measures that suggest specific areas where further study is needed. Research opportunities examining the complex range of influences on several aspects of behavior leading to a broad set of outcomes are discussed, including epidemiologic studies, simulation models of individual and aggregate behavior and outcomes, characterization of the interactions of genetic and environmental factors on behavior and outcomes, and studies of the interactions of public policy and economic behavior of consumers and alcohol producers and sellers. The role of public policy is highlighted, with a close look at the Alcohol Policy Information System, a recently-developed tool to facilitate policy-related research.

**2:25 p.m. *Systems Ecology and Community Systems: Understanding Social Problems in Social Terms*, Paul J. Gruenewald, PIRE**

This presentation will outline current approaches to understanding alcohol problems from the perspectives of genetics, neurobiology, cognition, and social developmental theory. Methodological individualism dominates these approaches and is complemented by reductionist classical economic and micro-sociological frameworks that attempt to explain alcohol problems from an individual choice perspective. These approaches do not address the social dynamics that underlie the formation of macro-sociological systems (social, economic and political) and that determine the plausible range of individual choice or the consequences of such choices in social contexts. An alternative community systems perspective is introduced that begins to address these macro-sociological issues from a population perspective. Although enabling study of the community systems, this approach has no connecting theory relating these systems to individual choice. An alternative ecological approach is outlined that provides statements of the primary individual-social ecological dynamics that support alcohol problems in community settings. This social ecological approach provides a set of predictions about the likely course of development of alcohol problems in in developed and developing societies.

**2:45 p.m.** *Applications of Small World Network Models in Alcohol Epidemiology*,  
**Robert A. Wilson**, University of Delaware

This presentation focuses on Small World Models of alcohol abuse in structured populations, such as residential communities and college campuses. Small World Models have the capacity to incorporate a variety of forms of connectivity besides personal acquaintance, such as geographical proximity and common organizational membership. The models can also incorporate a resilience dimension that indicates the susceptibility of each individual in a network to alcohol abuse. The models have the capacity to simulate the effect of rewiring alcohol abusers into networks of non-abusers, either as the result of treatment or membership in self-help organizations.

**3:05 p.m.** *Multilevel Examination of Neighborhood Influences on Drinking-and-Driving Behavior*, **W.F. Wiczorek; K.S. Marczyński; A. Delmerico; S.J. Tetewsky**, Buffalo State College, Buffalo, NY

Research over approximately the past decade has clearly identified that there are neighborhood influences on drinking and related behaviors such as alcohol-impaired driving. However, this research has been conducted in limited geocultural settings primarily in the U.S west and south. Also, this research has focused primarily on identifying the role of alcohol availability as the neighborhood factor in combination with individual attributes to better understand the etiology of drinking and related outcomes. While previous research provided some insights, it was not designed to provide a broader assessment of the potential role of various neighborhood-level systems on a specific alcohol-related behavior. The purpose of this paper is to examine the role of neighborhood systems (e.g., alcohol availability, neighborhood cohesion, other crime, educational factors, socioeconomic status, etc.) on drinking-and-driving behaviors of individuals that live in those areas. Data are from 3,701 random-digit dial telephone interviews conducted with persons aged 15-45 years in Erie County, New York. The

telephone-based assessment was conducted by trained interviewers using a structured interview to obtain a broad assessment of demographics, drinking practices and consequences. Neighborhood-level data for this project utilizes zip code areas to capitalize on the rich amount of data available for those areas. The neighborhood-level data include alcohol outlets (total availability, off-premise availability and on-premise availability), neighborhood cohesion (transiency), crime, education, substance abuse admission rate, poverty/socioeconomic status, and other potential neighborhood-level risks factors. The data analytic approach utilizes hierarchical linear modeling to incorporate multilevel data (individual and neighborhood) to predict drinking and driving. The analysis uses a basic individual-level model of demographic and drinking variables that remains constant to focus the examination on the neighborhood-level variables. The results are highly relevant to the development of a general ecological systems model of drinking and related outcomes by providing insights on how to potentially condition individual-level probabilities of drinking and driving based on neighborhood system factors.

**3:30 p.m. *Floor Discussion***

---

**3:40 p.m. – 4:00 p.m.**

**BREAK**

---

**4:00 p.m. – 5:40 p.m.**

---

**I-9**

## **Computational Analysis of Gene Regulation**

*Organizer: Vincent Carey, Harvard Medical School*

**4:00 p.m. *Learning Regulatory Programs That Accurately Predict Gene Expression,***  
**Christina Leslie**, Columbia University

Studying the behavior of gene regulatory networks by learning from high-throughput genomic data has become one of the central problems in computational systems biology.

Most work in this area has focused on learning structure from data -- e.g. finding clusters of potentially co-regulated genes, or building a graph of putative regulatory

"edges" between genes -- and has been successful at generating qualitative hypotheses about regulatory networks. Instead of adopting the structure learning viewpoint, our focus is to build predictive models of gene regulation, i.e., gene regulatory programs that allow us to make accurate quantitative predictions on new or held-out experiments (test data).

In our approach, called the MEDUSA algorithm, we learn a prediction function for the regulatory response of genes, using a boosting technique to avoid overfitting as we select features from a high-dimensional search space. In particular, we discover motifs representing putative regulatory elements whose presence in the promoter region of a gene, coupled with the expression state of a regulator in an experiment, is predictive of differential expression, and we combine this information into a global predictive model for gene regulation. We will describe recent results on the yeast hypoxic response, where we have used MEDUSA to propose the first global model of the oxygen sensing and regulatory network, including new putative context-specific regulators and motifs. Our analysis of hypoxia in yeast is joint work with our wet lab collaborator, Dr. Li Zhang at Columbia University.

**4:30 p.m. *SVMs and Probabilistic Approaches for Classifying Promoters*, Anirvan Sengupta**, Rutgers University

We discuss how likelihood based approaches to regulatory site detection, with minimal input from biophysics of protein-DNA interaction, lead naturally to low degree polynomial kernels for appropriate imbedding of sequences in  $R^n$ . We study the performance of these one-class SVMs on real and quasi-synthetic data. The method allows us to score sites as well as set a threshold to separate out functional sites from non-functional ones. We call this method QPMEME (Quadratic Programming Method for Energy Matrix Estimation).

Combining evidence from heterogeneous sources (motifs, gene expression, phylogenetic comparison,  $\hat{u}$  to verify of regulatory interactions is getting to be the way to compensate for limitations of inference based on individual data types. In many cases, each kind of data allows us to rank genes according to the likelihood of the gene being the target of regulation by a particular mechanism. Often, one has to choose cutoffs, separately, for each of these ranking and then use some meta-classifier to combine the results to decide whether or not a gene in question is appropriately regulated or not. We discuss a simple non-parametric method for combining ranked data for discovering correlation between high ranks. The threshold is drawn on the combined data in a principled manner. We show how well this method works for a particular yeast dataset, where we have experimentally tested the predictions from this method.

Finally, if time permits, we discuss an extension of QPMEME, to be used on aligned sequences, as well as multi-site classifiers for complex developmental promoters that integrate information combinatorially.

**5:00 p.m. *Data-Driven Biophysical Modeling of Gene Expression Regulation*, Harmen Bussemaker**, Columbia University

It is the dynamic balance between transcription from DNA to messenger RNA and subsequent mRNA degradation that determines the steady-state mRNA abundance for each gene in a genome. However, while regulation of transcription rate by DNA binding transcription factors has been intensively studied, both experimentally and computationally, regulation of the transcript turnover rate by RNA binding factors has received far less attention. We took advantage of the fact that information about the condition-specific activity and sequence-specific affinity of RNA binding regulatory factors is implicitly represented in the steady-state mRNA abundances measured using DNA microarrays. Thus, by fitting a model based on a physical description of molecular interactions, we were able to gain quantitative insight into the mechanisms that underlie genome-wide regulatory networks. We developed a novel algorithm, MatrixREDUCE, that predicted the sequence-specific binding affinity of several known and unknown RNA-binding factors and their condition-specific activity, using only genomic sequence data and steady-state mRNA expression data as input.

We identified and computationally characterized the binding sites for six mRNA stability regulators in the yeast *S. cerevisiae*, which include two known RNA-binding proteins, Puf3p and Puf4p. We provide computational and experimental evidence that regulation of mRNA stability by the discovered factors is dynamic and responds to a variety of environmental stimuli. For example, little was previously known about the functional role of Puf3p, but our computational results suggest that Puf3p functions to destabilize mitochondrion-related transcripts when metabolite repressing sugars are present and in response to the drug rapamycin. We were able to experimentally confirm these predictions by growing a transformed strain expressing a hybrid mRNA designed to contain a functional Puf3p binding site in different culture conditions and measuring its half-life after a transcriptional shut-off.

Our work suggests that regulation of mRNA stability is not a special case phenomenon, but rather a pervasive regulatory mechanism that rapidly adapts cellular processes to a changing environment.

**5:30 p.m. *Floor Discussion***

## **I-10**

### **Information Extraction from Biomedical and Clinical Text**

*Organizer: David Shera, Children's Hospital of Pennsylvania*

**4:00 p.m. *The Global Linguistic Challenge and Approaches*, Kevin Cohen, University of Colorado Health Science Center**

The history of natural language processing over the past three decades has focussed around two types of approaches: rule-based systems and statistical ones. This talk will review the two approaches, compare the strengths and weaknesses of each, and show how they can be synthesized to create robust systems for the analysis of free text. A

range of systems addressing tasks as varied as named entity recognition, document retrieval, question-answering, and automatic summarization will be reviewed. All have in common the application of a combination of statistical and rule-based techniques for text mining from biomedical literature.

**4:30 p.m. *Mining the Bibliome: Supervised Extraction of Biomedical Entities from Text for Biological Application*, Peter S. White**, Children's Hospital of Pennsylvania

Modern biomedical research requires the synthesis of experimentally-derived datasets with phenotypic and disease descriptions from the primary literature. Machine-learning approaches to text mining, which combine a variety of linguistic and content features, are now able to identify with accuracy and coverage sufficient for practical application many of the occurrences of specified types of biological entities in biomedical texts. Our overall objective is to create a process for extracting, relating, and delivering mentions of biological entities of interest to particular research efforts. We have established a text mining procedure for efficient biomedical knowledge extraction, by developing accurate, supervised discriminative learning methods for extracting biomedical entities. These methods have been incorporated in a flexible and scalable annotation system. Our extractors have been trained and tested on biomedical and clinical text, and we have accurately normalized extractor output for targeted entity classes by combining rules and machine-learned matching. A production system that extracts and normalizes mentions of genes and malignancies from all MEDLINE abstracts with high accuracy has been developed. As a prototype of this approach, we have created a public query and delivery interface for retrieval of literature records and normalized text relevant to particular aspects of genomics, with links to relevant curated data and visualization of results integrated with genomic position. The interface is available at <http://fable.chop.edu>. This work supports the transformation of text extraction methods into tools practical for biomedical researchers interested in linking molecular and clinical descriptions of disease processes.

**5:00 p.m. *Striving to Make Text-Mining Useful for Disease-Gene Identification*, Andrey Rzhetsky**, Columbia University

The information overload in molecular biology is a mere example of the status common to all fields of the current science and culture: an ever-strengthening avalanche of novel data and ideas overwhelms specialists and non-specialists alike. The help of relieving the information overload may come from the text-miners who can automatically extract and catalogue facts described in books and journals. My talk will touch the following questions: What can large-scale analyses of scientific literature tell us about both active and forgotten knowledge? What can such analyses tell us about the scientific community itself? How do mathematical models help us to differentiate true and false statements in literature? How will text-mining help us to find cures for human and non-human maladies?

**5:30 p.m. *Floor Discussion***

## I-11

### Computational and Statistical Methods for Genome-Wide Studies of Biological Systems

Organizer: Sandrine Dudoit, University of California, Berkeley

**4:00 p.m.** *Systems Biology Insights From Host-Pathogen Interaction Studies*, **Imola K. Fodor**, Lawrence Livermore National Laboratory

*Yersinia pestis*, the causative agent of plague, is one of the most virulent human pathogens. To characterize the changes that *Y. pestis* undergoes as it leaves the flea vector and enters the mammalian host, a Phenotype MicroArray (PM, Biolog, Inc) experiment was performed. Virulence of the bacteria was induced in vitro by shifting the temperature and calcium concentration from the physiological conditions mimicking the flea vector to that of the mammalian host. The respiration of *Y. pestis* cells grown in triplicate under the two conditions were monitored for three days by taking one measurement every fifteen minutes in each of close to 2000 wells preloaded with separate chemicals. The chemicals were representative of microbial organisms, and included compounds related to the catabolic pathways for carbon and nitrogen, as well as tested the sensitivity of the cells to antibiotics and other inhibitory agents. Preliminary results suggest the existence of biochemical pathways not indicated by the current genome annotation and known biochemistry of *Y. pestis*. In addition, strain- and condition-dependent growth was observed in a number of wells, reflecting underlying biological differences. The goal is to design new therapeutics by exploiting the differences.

**4:30 p.m.** *Simple Models for Integrating Gene Expression and Genetic Marker Data to Characterize Disease-Related Genes*, **Steve Horvath**, University of California, Los Angeles

Gene mapping approaches that are based on the genetics of gene expression have been fruitful in identifying genetic regulatory loci related to complex traits. Recently, several groups have proposed to use microarray data to construct gene expression networks and to identify network modules (sets of tightly correlated genes) and highly connected (hub-) genes. Here we describe a simple data mining approach for integrating microarray and genetic marker data. To illustrate the method we apply it to liver gene expression from an F2 mouse intercross and describe a novel systems biology approach for identifying body weight related genes. We examine the large-scale organization of gene co-expression networks in female mouse liver and annotate several modules in terms of 20 physiological traits and genetic marker data. We identify chromosomal loci (referred to as module quantitative trait loci, mQTL) that perturb the modules. We describe a novel approach for integrating intramodular connectivity with genetic marker information. Using the mQTLs and the intramodular connectivity of a body weight related module, we describe which factors determine the relationship between gene expression profiles and weight. Our approach results in the identification of genetic targets that influence genetic modules (pathways) that are related to the clinical phenotypes of interest. Our systems-

level integration of gene co-expression network analysis with the underlying genetic loci perturbing this network allows us to model the relationships between genes and complex traits. This is joint work with Anatole Ghazalpour, Sudheer Doss, Eric Schadt and Jake Lusis.

**5:00 p.m. *Enhancing Motif Finding Models Using Multiple Sources of Genome-Wide Data*, Sunduz Keles, University of Wisconsin**

Identifying binding locations of transcription factors within long segments of non-coding DNA is a challenging task. We develop a general framework for integrating multiple sources of genomic data into the task of motif finding based on sequence data. Some examples of such data include data from ChIP-chip experiments (transcription factor binding, nucleosome occupancy etc) and multiple species sequence data. At the core of our framework are conditional mixture models that model sequence data based on multiple sources of genomic data. We develop efficient estimation techniques for this model and study the statistical properties of our estimators. Finally, we illustrate the utility of this integrative approach using several yeast and human transcription factors.

**5:30 p.m. *Floor Discussion***

**Friday, 25<sup>th</sup> May 2007**

---

**8:30 a.m. – 10:10 a.m.**

---

**I-12**

**Data and Decision Fusion**

*Organizers: Frank Hsu, Fordham University, & Amy Braverman, JPL*

**8:30 a.m. *Uniform Random Guesses and Random Decision Forests*, Tin K. Ho**, Bell Laboratories, Alcatel-Lucent

Learning in everyday life is often accomplished by making many random guesses and synthesizing the feedback. Kleinberg's analysis of this process resulted in a new method for classifier design -- stochastic discrimination (SD). The method constructs an accurate classifier by combining a large number of very weak discriminators that are generated essentially at random.

SD is an ensemble learning method in an extreme form. Studies on other ensemble learning or decision fusion methods have long suffered from the difficulty of properly modeling the complementary strengths of the components. The SD theory addresses this rigorously via the mathematical concepts of enrichment, uniformity, and projectability.

I will explain these concepts via a very simple numerical example that captures the basic principles of the SD theory and method. The focus is on a fundamental symmetry in point set covering that is at the core of the theory. Better understanding of this will be useful for the analysis of other decision fusion methods. An example is my invention of random decision forests that was a result of these pursuits.

**9:00 a.m. *A Censored-Data Approach to Modelling Partial Rankings on the Ranking Lattice*, Guy Lebanon**, Purdue University

Statistical models on permutations or their partial ranking analogues, are often of limited practical use for large due to computational consideration. We explore the use of Mallows-based models for partial rankings for large and derive some efficient procedures for their use. The derivations are largely possible through the introduction of a combinatorial framework called the ranking lattice on the set of partial rankings. The ranking lattice also serves to present a new perspective to unifying ranking and classification and derive new frequentist and Bayesian models. The resulting models exhibit a certain coherence that most other ranking models lack.

**9:30 a.m. *Regulatory Module Discovery with Heterogeneous Data by Multipartite Coclustering*, Li-San Wang, University of Pennsylvania**

The problem of analyzing different kinds of functional and genomic data is a key challenge in molecular systems biology due to the sheer size and format heterogeneity of the data. Moreover, the analysis is often exploratory as the underlying biological system is largely unknown, and the objective is usually ambiguous. To address this ubiquitous problem, we proposed a graph-theoretic approach that is flexible enough to handle a wide range of experimental data, intuitive, and easy to implement. In our formulation, a multipartite graph is created where each vertex partition corresponds to one biological domain and edges between partitions represent relations based on available biological data. Significant properties of the graph such as dense subgraphs (co-clusters) may represent biologically important concepts such as portions of a functional pathway. To this end, we developed a generic stochastic-search method to find dense subgraphs on multipartite graphs.

We show the utility of our formulation by exploring tissue-specific transcriptional regulation in the human genome: in a tri-partite graph of transcription factors, their putative target genes and the tissues in which the target genes are differentially expressed, a dense subgraph may reveal knowledge about tissue-specific transcriptional modules. We validate our approach through the analysis of Cardiac, Skeletal, and Smooth muscle data.

This is collaboration with Logan Everett and Sridhar Hannenhalli of PCBI.

**10:00 a.m. *Floor Discussion***

## **I-13**

### **Computational Techniques for Structural Genomics**

*Organizer: Daisuke Kihara, Purdue University*

**8:30 a.m. *In Silico Prediction of Protein Interactions: Nothing That Nature Does Not Already Know*, Carlos J. Camacho, University of Pittsburgh**

The 3D sequence and structure of proteins encode all the information necessary for most protein interactions to succeed against all odds, from folding to reach out to their often unique substrate. In this presentation, we will review recent progress in our understanding of the biophysics of protein interactions and how this "know how" has helped develop the field of in silico predictions of protein interactions.

**9:00 a.m. *From Sequences and Structures to Interactions of Proteins: Some Recent Developments*, Yaoqi Zhou, IUPUI, Indianapolis**

Detecting hidden structural similarity and making an accurate alignment between two seemingly unrelated sequences is challenging. Solution to this problem will have a significant impact in building phylogenetic trees, locating conserved motifs and domains, and predicting secondary and tertiary structures of proteins. In this talk, we will show how evolution-derived sequence profiles (the probability of an amino-residue type at a given sequence position derived from multiple sequence alignment of homologs) can make a significant improvement in accuracy of multiple-sequence alignment and template-based structure prediction through their combination with secondary and/or tertiary structural information. Our fold recognition methods, called SPARKS and SP3, were ranked as one of the most accurate, fully-automatic, structure-prediction servers in a recent meeting that assessed various structure-prediction techniques (CASP 6, Dec., 2004). Multiple sequence alignment method, SPEM, is also found to be superior to current state-of-the-art techniques in various alignment benchmarks. SPARKS, SP3, and SPEM are freely available as servers and downloadable standalone tools at <http://theory.med.buffalo.edu>. In addition to sequence alignment and structure prediction, I also discuss a new energy function that provides an accurate prediction of protein-ligand, protein-protein and protein-DNA binding affinities.

**9:30 a.m. *Surface-Shape-Based Protein Structure Classification and Search*, Daisuke Kihara**, Purdue University

The tertiary structure of proteins provides indispensable information about function and evolution of proteins. The tertiary structure of proteins has been represented in various different ways. The primary source of information about a protein structure is the Cartesian coordinates of atoms in the structure determined by experimental methods. Using the coordinates, the most common way to compare and classify protein structures is to compute the root mean square deviation (RMSD) of corresponding atoms (usually only atoms on the main chain are used) of pairs of structures. Alternatively, a distance map, which stores distances of every pair of atoms or amino acid residues, or the spatial arrangement of secondary structures in a protein can be used to concisely represent a structure. Here we introduce methods to represent surface shape of protein structures. The surface shape representation of proteins is more biologically intuitive because it is the surface that is responsible for function of proteins, such as catalytic activity or interaction with other proteins or molecules. An practical advantage of the surface representation is that it allows a quick real-time search of protein global and local shape of proteins in a large database. It is also possible to map physicochemical properties of surface atoms naturally in the representation.

**10:00 a.m. *Floor Discussion***

## I-14

### Biological Networks

Organizer: *Natasa Przulj, University of California, Irvine*

**8:30 a.m.** *Predicting domain interactions from protein-protein interaction networks*  
**Teresa Przytycka, NIH/NLM/NCBI**

Comprehending the cell functionality requires knowledge about the functionality of individual proteins as well as the interactions among them. Proteins typically contain two or more domains, and a protein interaction usually involves binding between specific pairs of domains. Identifying such interacting domain pairs is an important step towards determining the protein-protein interaction network. We demonstrate that evolutionary parsimony principle combined with combinatorial optimization techniques leads to a very successful approach to detecting domain-domain interactions from the topology of the protein-protein interaction network.

**8:50 a.m.** *Learning Biology From Networks*, **Chris Wiggins, Columbia University**

One central problem in biology is learning networks from biology --- inferring biological networks from the copious but noisy data presented by modern, high-throughput biotechnologies. A related central problem is learning biology from these networks --- inferring which possible evolutionary design principles could have resulted in the observed network topologies, and relating these structure of these networks to their function. I will present some progress made on these problems using approaches from machine learning and information theory.

**9:15 a.m.** *Geometric Local Structure in Biological Networks*, **Natasa Przulj, University of California, Irvine**

The recent explosion in biological and other real-world network data has created the need for improved tools for large network analyses. Several new mathematical techniques for analyzing local structural properties of large networks have recently been developed. Our work introduces small induced subgraphs of large networks, called graphlets. We use graphlets to develop "network signatures" that quantify local structural properties of a network. Based on these network signatures, we design two novel "network agreement" measures. These measures lead us to new, well-fitting geometric graph models of biological networks.

**9:35 a.m.** *Algorithmic and Analytical Methods for Functional Characterization of Molecular Interaction Networks*, **Mehmet Koyuturk, Purdue University**

Species-specific interaction data makes it possible to study cellular organization through a systems perspective. On the other hand, standardized libraries of functional annotation provide comprehensive understanding of the function of individual molecules, which are part of these interaction networks. Coupled analysis of these two sources of data, i.e., extending molecular annotations to pathways and sub-networks, is a critical component

of functional characterization of cellular signaling at the systems level. In this talk, we introduce a framework for projecting molecular interaction networks onto the space of functional attributes using multigraph models. We first demonstrate that annotations of pairwise interactions do not generalize to indirect relationships between processes. Motivated by this result, we formalize the problem of identifying statistically over-represented pathways of functional attributes. We discuss the hardness of this problem in terms of the non-monotonicity of common statistical significance measures. Then we propose a statistical model that emphasizes the modularity of a pathway, evaluating its significance based on the coupling of its building blocks. Comprehensive results on the *E. coli* transcription network demonstrate that our approach is effective in identifying known, as well as novel biological pathway annotations.

**10:00 a.m. *Floor Discussion***

## **I-15**

### **JCGS Highlights**

*Organizer: Luke Tierney, University of Iowa & David van Dyk, University of California, Irvine*

**8:30 a.m. *Simulating Two-Dimensional Gaussian Random Fields: Fast and Exact Algorithms*, Hana Sevcikova, University of Washington**

The circulant embedding technique uses the fast Fourier transform to generate exact realizations of stationary and intrinsically stationary Gaussian random fields, as they are widely used in a variety of applications, ranging from simulation studies in spatial statistics to environmental risk assessment and computer graphics.

In two-dimensional applications, the technical requirement of a nonnegative definite periodic embedding for the covariance matrix is frequently violated, thereby hindering the use of the circulant embedding approach. Our work addresses this challenge through an extension of the standard embedding to cut-off and intrinsic embeddings, respectively. The resulting methods are fast and exact, in the sense that they are computationally feasible, and that the realizations have exactly the desired multivariate Gaussian distribution. The approach is based on a suggestion by Michael Stein, who proposed nonnegative definite periodic embeddings based on suitably modified, compactly supported covariance functions.

In the presentation, various techniques for simulating two-dimensional Gaussian random fields are described and compared, including the cut-off embedding and intrinsic embedding techniques. Numerical experiments illustrate current computational limits. A simulation study on geostatistical inference for random fields points at the consequences that inadequate choices of simulation algorithms might entail.

This is joint work with Tilmann Gneiting, Don Percival, Martin Schlather and Yindeng Jiang.

**9:00 a.m.** *Flexible, Optimal Matching for Comparative Studies: A Network Flows Algorithm and an R Package*, **Ben B. Hansen**, University of Michigan, & **Stephanie Olsen Klopfer**, Merck Research Laboratories

In the matched analysis of an observational study, confounding on covariates  $\mathbf{X}$  is addressed by comparing members of a distinguished group ( $Z=1$ ) to controls ( $Z=0$ ) only when they belong to the same matched set. The better matchings, therefore, are those whose matched sets exhibit both dispersion in  $Z$  and uniformity in  $\mathbf{X}$ . For dispersion in  $Z$ , pair matching is best, creating matched sets that are equally balanced between the groups; but actual data place limits, often severe limits, on matched pairs' uniformity in  $\mathbf{X}$ . At the other extreme is full matching, the matched sets of which are as uniform in  $\mathbf{X}$  as can be, while often so poorly dispersed in  $Z$  as to sacrifice efficiency.

We present an algorithm, used by the R package `textsc{optmatch}`, for exploring the intermediate territory. Given requirements on matched sets' uniformity in  $\mathbf{X}$  and dispersion in  $Z$ , the algorithm first decides the requirements' feasibility. In feasible cases, it furnishes a match that is optimal for  $\mathbf{X}$ -uniformity among matches with  $Z$ -dispersion as stipulated. We offer an illustration, and compare our method to a commonly used alternative, greedy matching, which is neither optimal nor as flexible but is algorithmically much simpler. The comparison finds meaningful advantages, in terms of both bias and efficiency, for our more studied approach.

**9:30 a.m.** *Mean-Mean Multiple Comparison Displays for Families of Linear Contrasts*, **Richard M. Heiberger**, Temple University

The Mean--Mean Multiple Comparison display is a succinct and compact display of the results of traditional procedures for multiple comparisons of population means or linear contrasts involving means. In one plot, the MMC display simultaneously provides the sample means themselves with correct relative distances, the point and interval estimates of the  $k(k-1)/2$  pairwise differences, the point and interval estimates for arbitrary contrasts of the level means, declarations of significance, and confidence interval widths that are correct for unequal sample sizes. The construction of the software to display the plots depends on two sets of underlying software, one for constructing the confidence intervals of the contrasts, and the second for designing the graphical displays. We discuss the effects of differences in the underlying software models and how they affect the construction of our display functions.

**10:00 a.m.** *Floor Discussion*

---

**10:10 a.m. – 10:30 a.m.**

**BREAK**

---

---

**10:30 a.m. – 12:10 p.m.**

---

## **I-16**

### **Indexing and Search in Biological Data**

*Organizer: Mohammed J. Zaki, Rensselaer Polytechnic Institute*

**10:30 a.m. *Indexed Biosequence Similarity Search in the Age of Comparative Genomics*, [Jeremy Buhler](#), Washington University in St. Louis**

High-throughput DNA sequencing has not only dramatically increased the size of modern biosequence databases but also fundamentally changed the character of their content. Genome projects for a biological clade, e.g. fruit flies or cereal grasses, now routinely sequence several representative species at once, allowing both quantitative measurements of conservation and reconstruction of sequence evolution in the clade. The data from these projects and others may be stored not as individual sequences but as multiple alignments of homologous sequences, or as probabilistic models (profile HMMs, SCFGs) constructed from such alignments.

Efficient indexed similarity search for biosequences is well-studied; the highly successful BLAST software is perhaps the best-known example of a tool for such searches. However, fast computational comparison of multiple alignments and probabilistic models has received much less attention. The extra information captured in these new data objects should improve the quality and informativeness of similarity search, but exploiting this information efficiently requires new algorithms to implement indexed search for these objects. In this talk, I will describe my group's recent work on developing such algorithms, with an emphasis on techniques for seed pattern design and alphabet selection.

**11:00 a.m. *Indexing for Success: Effective Index-Based Methods for Querying Biological Sequences*, [Jignesh Patel](#), University of Michigan**

The current ongoing revolution in life sciences is producing new and exciting discoveries at a remarkable pace. The driving factor behind these advances is the emergence of new high-throughput methods (such as gene sequencing) and the use of computational tools to analyze the data that is produced by these methods. Data analysis for life sciences applications often requires sophisticated querying on various complex data types, such as sequences, geometrical structures, and graphs. Many of these datasets are large and are rapidly growing in size. To query these datasets efficiently, effective indexing techniques and associated querying techniques are urgently needed. In this talk, I will describe the ongoing efforts in the Periscope group at the University of Michigan to address this issue, and demonstrate various practical index methods that address these problems.

**11:30 a.m. *Querying and Mining in Graph Databases*, [Ambuj Singh](#), University of California, Santa Barbara**

A number of scientific endeavors are generating data that can be modeled as graphs: high-throughput biological experiments on protein interactions, high throughput screening of chemical compounds, social networks, ecological networks and food-webs, database schema and ontologies. Access and analysis of the resulting annotated and probabilistic graphs is crucial for advancing the state of scientific research, accurate modeling and analysis of existing systems, and re-engineering of new systems. I will discuss a set of scalable querying and mining tools for graph databases. Specifically, I will discuss querying for similar graphs and subgraphs, discovery of significant subgraphs in a graph database, and the mining of well-connected clusters in large probabilistic graphs.

**12:00 noon. *Floor Discussion***

**I-17**

## **Integrative Systems Biology in Cancer Research**

*Organizer: Ramana V. Davuluri, Ohio State University*

**10:30 a.m. *Stochastic Modelling and Estimation in Dynamic Cellular Networks, Mike West***, Duke University

This talk will discuss aspects of our work on modelling high-resolution, dynamic data on mRNA and protein fluctuations in complex cellular networks. We will focus on specific examples of regulatory networks in the Rb/E2F pathway, and describe work on: (a) novel and fundamentally stochastic, discrete-time models at the levels of individual cells; (b) how these models relate to mechanistic, differential equations models at the aggregate cell population-level; (c) approaches to Bayesian model analysis and parameter estimation at the cell-specific level; and (c) the evaluation of multiple sources of intrinsic cell-specific biological randomness, noise and measurement error.

**11:00 a.m. *Integration of Annotations for Guiding Analysis of Signaling in Cancer Cells, Michael Ochs***, Fox Chase Cancer Center, Philadelphia

Over the past 50 years, molecular and cellular biology have elucidated details of cell signaling processes and linked errors in signaling to the development of cancer. Utilizing such knowledge to guide analysis within a Bayesian framework permits the linking of changes in gene expression to transcriptional regulation and to signaling processes. We explore this approach with examples from model organisms and from studies of targeted therapeutics.

**11:30 a.m. *Modern Data-Mining Tools in Predicting Cis-Regulatory Transcriptional Modules: A Case Study with a Random Forest Approach, Ramana Davuluri***, Ohio State University

SMAD transcription factors lie at the core of one of the most versatile cytokine signaling pathways, the transforming growth factor  $\beta$  (TGF- $\beta$ ) pathway, which has been

implicated in the regulation of cell growth, differentiation, apoptosis and specification of developmental fate. While the molecular mechanisms of TGF- $\beta$ /SMAD signaling pathway have been studied in detail, the global networks downstream of SMAD remain largely unknown. To address this question, we simultaneously performed chromatin immunoprecipitation followed by microarray analysis (ChIP-chip) and mRNA expression profiling to identify TGF- $\beta$ /SMAD regulated and synchronously coexpressed gene sets in ovarian surface epithelium. Intersecting the ChIP-chip and gene expression data yielded 150 direct targets, which were classified into 2 up-regulated and 2 down-regulated groups based on their temporal changes in expression after TGF- $\beta$  activation. We developed a novel data-mining method driven by random forest algorithm to model the SMAD transcriptional modules in the target sequences, which enabled us to predict novel combinatorial modules that predict up- or down-regulatory patterns of TGF- $\beta$ /SMAD target genes. The predicted SMAD modules contain SMAD binding element and up to 2 of 7 other transcription factor binding sites (E2F, P53, LEF1, ELK1, COUPTF, PAX4 and DR1). The combinatorial presence of SMAD and at least two of the other transcription factors can discriminate different synexpression target genes of SMAD. Together, our results further the understanding of interactions between the SMAD and other transcription factors at specific target promoters, and provide the basis for more targeted experimental verification of the co-regulatory modules.

**12:00 noon. *Floor Discussion***

## **I-18**

### **Analysis of DNA Barcode Data**

*Organizer: Javier Cabrera, Rutgers University, & Woollcott K. Smith, Temple University*

**10:30 a.m. *DNA Barcoding: Linking Biodiversity, Biotechnology, and Bioinformatics,***  
**David Schindel**, Consortium for the Barcode of Life

DNA barcoding was proposed in 2003 as a technique for assigning biological specimens to known species using a short gene sequence from a standardized position in the genome. For higher animal groups, a 650 base-pair mitochondrial region was selected as the standard barcode region. Since that time, more than 200,000 barcode sequences have been obtained from 25,000 species. The barcoding approach and its large database have generated a number of analytical challenges:

- What sample sizes per species are needed to make the database a reliable reference library?
- What analytical techniques should be used to assign specimens to known species?
- Should the same analytical techniques be used for well-known species and for taxonomic groups in which undiscovered species are likely to be found?
- Are there new approaches to data display and visualization that could aid the enterprise?

The Consortium for the Barcode of Life (CBOL), an international initiative of the Smithsonian Institution, approached Rutgers University's Center for Discrete Mathematics and Theoretical Computer Science (DIMACS) for help in meeting these challenges. DIMACS has been a principal leader in CBOL's Data Analysis Working Group since early 2005. Results of the effort will be reported at the Second International Barcode Conference in Taipei in September 2007.

**11:00 a.m. *Weighted Chinese Restaurant Process for clustering barcodes*, Javier Cabrera**, Rutgers University, **John Lau**, U Bristol, **Albert Y Lo**, Hong Kong UST

One of the main uses of DNA barcode data is to classify specimen's between known species and to detect new species not yet discovered. All species include some level of variation among individuals and in some cases this variation takes the form of splits among local populations and even subspecies. One important challenge is to develop an unsupervised classification rule that will identify clusters within individual species that rise above background variation and therefore might represent subspecies or other significant biological units. We proposed a new algorithm for clustering DNA strings that is based on the Weighted Chinese Restaurant Process. We will illustrate the algorithm with examples from real barcode data.

**11:30 a.m. *Model-based species identification using DNA barcodes*, Bogdan Pasaniuc**, Sotirios Kentros, Ion Mandoiu, University of Connecticut

Recently, DNA barcoding (i.e., sequencing a short standardized region of the genome) has emerged as an important tool for species identification and discovery. Drawing on existing work on model-based DNA and protein motif finding, in this work we develop statistical models of within-species variation in barcode sequences and fast methods for assigning new barcodes to known species. In particular, we evaluate the use of profile weight matrices, inhomogeneous Markov Chains, and Hidden Markov models in the context of species identification. We also explore efficient methods for computing statistical significance of assignments in the form of p-values. Preliminary experimental results on real barcode datasets show that model-based methods yield high identification accuracy with a highly scalable runtime.

**12:00 p.m. *Floor Discussion***

## **I-19**

### **Best of SIAM Data Mining 2007**

*Organizer: Vipin Kumar, University of Minnesota*

*Session Chair: Arnold Goodman, University of California, Irvine*

**10:30 a.m. *Harmonium-Based Models for Semantic Video Representation and Classification*, Jun Yang, Yan Liu, Eric Xing, Alexander Hauptmann**

**10:50 a.m.** *Less is More: Compact Matrix Decomposition for Large Sparse Graphs*, Jimeng Sun, Yinglian Xie, Hui Zhang, Christos Faloutsos

**11:10 a.m.** *ROAM: Rule- and Motif-Based Anomaly Detection in Massive Moving Object Data Sets*, Xiaolei Li, Jiawei Han, Sangkyum Kim, Hector Gonzalez

**11:30 a.m.** *Multi-way Clustering on Relation Graphs*, Arindam Banerjee, Sugato Basu, Srujana Merugu

**11:50 a.m.** *Fast Newton-type Methods for the Least Squares Nonnegative Matrix Approximation Problem*, Dongmin Kim, Suvrit Sra, Inderjit Dhillon

---

**12:10 noon – 2:00 p.m.**

**LUNCH**

---

---

---

**2:00 p.m. – 3:40 p.m.**

---

**I-20**

**Methods for Integrating Disparate Molecular Datasets for Prediction**

*Organizer: Peter S. White, Children's Hospital of Pennsylvania*

**2:00 p.m.** *Integrative Genomics of Human Neuroblastoma: Promises and Pitfalls for Translation to the Clinic*, John Maris, Children's Hospital of Pennsylvania

Neuroblastoma is an important pediatric cancer that continues to cause significant morbidity and mortality. We have been interested in understanding the genetic basis of the disease in order to improve both treatment planning as well as to develop novel therapeutic approaches. Whole genome approaches are being used to discover genes involved in predisposition to the disease—both classic familial predisposition genes as well as others that have much lower effect size and likely cooperate. Genome scale approaches have been used on tumor tissues at both the DNA and RNA level to define prognostic signatures, and chromosomal copy number aberrations have been shown to be highly predictive of outcome. Finally, these techniques when coupled with functional screens are being used to identify and prioritize potential therapeutic targets. Maximum

leveraging of the high dimensional datasets derived in this work will be realized by ongoing efforts for data integration. Recent examples of successful translation to the clinic, as well as problems with throughput and the bottleneck of clinical trial initiation will be discussed.

**2:30 p.m. *Data Integration for Disease Pathway Mapping*, Michael Krauthammer,**  
Yale University

The integration of genetic and molecular network data is a promising approach for identifying disease pathways in multifactorial disorders. We use a multi-layered data integration strategy. First, we combine molecular network resources for identifying high-confidence molecular interaction data. BioPAX, an OWL compliant network exchange standard, facilitates this integration effort, by providing a coherent framework for Description Logic (DL)-based inference over network resources. We then combine network data with high-throughput genetic information from linkage studies of multifactorial disorders. Such studies usually imply dozens, if not hundreds, of potential disease candidates. By integrating such genetic information with molecular interaction data, it is possible to prioritize functionally related gene candidates that map to a common network region. We present visualization and computational strategies that help in the identification of these regions, and that assist in the elucidation of novel disease pathways.

**3:00 p.m. *Ontology Similarity Networks and Database Interoperability*, Yves Lussier,**  
University of Chicago

The metathesauri (UMLS, NCI, etc.) are based on an axiom of isomorphism between concepts from different ontologies. Beyond isomorphic networks of metathesauri, we have build similarity measures that allow for interoperability across related - but not identical/isomorphic-metadata. We will explore the useful associative properties emerging from similarity networks and stochastic semantic networks between concepts of otherwise heterogeneous ontologies. More specifically, using ontology similarity networks, we will show queries retrieving related biomedical data across heterogeneous databases such as microarrays datasets and imaging datasets.

**3:30 p.m. *Floor Discussion***

## **I-21**

### **Roles of Protein Intrinsic Disorder in Interaction Networks**

*Organizer: Keith Dunker, Indiana University*

**2:00 p.m. *The Interplay Between Alternative Splicing and Intrinsic Disorder and Its Implications on Multicellular Complexity*, Pedro Romano,** Indiana University & Purdue University, Indianapolis

Alternative splicing of pre-mRNA generates two or more protein isoforms from a single gene, thereby contributing to protein diversity in eukaryotes. Intrinsic disorder, that is, lack of an equilibrium three-dimensional structure, empowers protein for signaling and regulation functions. Despite intensive efforts, an understanding of the protein structure-function implications of alternative splicing is still lacking. We hypothesize that mapping alternatively spliced segments to regions of intrinsic protein disorder enables functional and regulatory diversity while avoiding structural catastrophe. We analyzed a set of differentially spliced genes encoding structurally characterized human proteins containing both structured and intrinsically disordered amino acid residues. We show that, in 81% of the examples, alternative splicing was associated with fully and partially disordered protein regions and concomitantly led to functional profiling. Alternative splicing was shown to occur predominantly in areas of pre-mRNA encoding for disordered regions in the corresponding proteins with very high significance (p-value <  $10^{-18}$ ). Associating alternative splicing with protein disorder enables the highly abundant time-specific, tissue-specific and diverse modulation of protein function needed for cell differentiation and for the evolution of multicellular organisms.

**2:20 p.m. *Proteomics Studies of Intrinsically Unstructured Proteins (IUPs) in Mammalian Cells*, Charles Galea**, St. Jude Children's Research Hospital, Memphis, TN

Intrinsically disordered (or unstructured) proteins (IDPs/IUPs) have emerged recently as a distinct protein structural class that is extensively involved in signaling and regulation. The extent to which IDPs populate the human proteome is estimated to approach 40%. While IDPs can readily be identified through sequence analysis, it has been more difficult to study the actual IUPs that are predicted to exist and presumably function in cells. To address this issue, we have applied biochemical methods and a variety of mass spectrometry-based proteomics methods to isolate and identify large numbers of IUPs from mammalian cells. We will discuss our results regarding the types of IUPs we detect, which allows us to speculate about the roles these proteins play in normal cells in culture. We are also applying these tools to studies of cells in different states (e.g. dividing versus arrested) to gain insight into how the intrinsically unstructured proteome changes in response to signals (e.g. to arrest division in response to DNA damage). We will discuss our progress toward gaining an understanding of the dynamics of the IU proteome.

**2:40 p.m. *Structure and Function of Rapidly-Evolving Proteins Involved in Yeast Protein-Interaction Networks*, Celeste Brown**, University of Idaho

Sequence data from whole genome studies indicate that the vast majority of genes that encode protein sequences are evolving under purifying selection, which constrains the rate of protein evolution. A small fraction of proteins in each genome appear to be evolving by other mechanisms, either neutrally or in response to positive selection for amino acid substitutions, which lead to a faster rate of protein evolution. Structural and functional characterization of these rapidly-evolving proteins are sporadic. Previous studies on the evolution of proteins involved in protein-interaction networks indicate that the more interactors a protein has, the slower its evolutionary rate. These studies,

however, systematically eliminated proteins that are evolving at the highest rates by comparing distantly related organisms. In this talk, we describe recent research on the structure and function of rapidly-evolving *Saccharomyces cerevisiae* proteins and relate these to their involvement in protein interaction networks. A random sample of fast- and slow-evolving proteins indicate that fast-evolving proteins are predicted to contain a greater proportion of amino acids in loops and a lesser proportion in helices than are slow-evolving proteins. No differences were detected between fast- and slow-evolving proteins in the percent of predicted disordered residues nor the maximum length of disordered regions in a protein. Fast-evolving proteins were found more commonly in the cell nucleus than slow, and slow-evolving proteins were found more often associated with various membranes of the cell. The molecular functions of these proteins and biological processes in which they participate are highly connected to their cellular localization: fast-evolving proteins often bind DNA and are involved in DNA metabolism, and slow-evolving proteins often have transporter activity and are involved in transport. Twice as many fast-evolving proteins as slow-evolving had unknown functions. Although the median number of interactors was lower in the fast-evolving than the slow-evolving proteins, this difference was not significant.

**3:00 p.m. *Exploring Relationships Among Protein Disorders, Network Hubs, and Disease Targets in Protein Interactome Studies*, Jake Chen**, Indiana University & Purdue University, Indianapolis

With yeast 2-hybrid and protein affinity-based multi-protein complex pull down coupled with mass spectrometry analysis, there is a massive influx of high-throughput protein interactome data in eukaryotic organisms including *S. cerevisiae*, *D. melanogaster*, *C. elegans* and *H. Sapiens*. The collection of all protein interactions in a cell, also known as the protein interactome, has been shown to be essential in helping to interpret functional genomics and proteomics results by providing a molecular signaling network map—despite the obvious concerns for low resolution and noisy data. The human interactome, once completed, would provide many insights on how to identify protein targets and biomarkers for human diseases. In this talk, we describe recent research in our group exploring structural foundations of protein hubs—a characteristics of all scale-free networks such as the Internet and social networks—and disease relevant proteins. Recent studies on the existence and importance of protein hubs, and various hypotheses how they may have evolved have been presented. We further describe our findings that network protein hubs are enriched with specific sequence features around protein "disorder"—predicted structural disorder, sequence repeats, low complexity regions and also long chain length. These sequence features provide extended interaction surface and adaptability, a mechanism for the same region to bind many distinct partners. On the basis of these results, an alternative evolutionary model for network evolution has been developed. We further show evidence that protein disordered regions are associated with key disease proteins, which in turn are known hubs in protein sub-network. Finally, we describe a case study in Alzheimer's disease for identifying essential disease-relevant proteins from recently compiled human proteomics data (in <http://bio.informatics.iupui.edu/HAPPI>).

**3:30 p.m. Floor Discussion**

## **I-22**

### **IASC Session: Current Issues in Statistical Biocomputing**

*Organizer: Michael G. Schimek, Medical University of Graz, Austria, & Masaryk University, Brno, Czech Republic*

**2:00 p.m. Pathway-Specific Predictors for Gene Expression Data, Alexander Ploner,** Karolinska Institute, Stockholm, Sweden

The analysis of microarray expression data with measurements on tens of thousands of genes, but rarely more than a couple of hundred samples, is a hard problem. All the same, many biomedical studies only look at gene expressions on their own (unsupervised classification) or attempt at most to relate the expression values to a few key parameters like clinical outcome (supervised classification, differential expression).

Biological pathways like those found at KEGG represent current knowledge about interactions between genes and gene products and can be used in different ways to structure the analysis. On the simplest level, pathways provide a functional characterization of the predictors or candidate genes found in a discovery step as outlined above; typically, this approach tests whether a specific pathway is overrepresented among genes that have been found to be of interest. More ambitious methods test whether a pathway as a whole is regulated between biological states.

Our work with pathways has been motivated by the depressing finding that among seven large studies on gene expression in cancer patients, only one was able to predict patient survival better than a flipped coin. In contrast to current approaches that build predictive expression signature based on all available genes and purely statistical criteria, we propose to build separate predictive signatures for a large number of available and relevant biological pathways: while each individual signature may have little predictive power on its own right, we can select and combine them to build an improved global predictor.

The pathway-specific predictors suffer less from high-dimensionality; they also allow biological interpretation of the results, though care is required in case of overlapping pathways sharing one or several genes between them. We demonstrate an implementation of our approach based on PLS, using a large cohort of breast cancer samples.

**2:30 p.m. A Median Absolute Deviation Method for Detecting Copy-Number Changes in Array-CGH Data, Eva Budinska,** Masaryk University, Czech Republic, **Eva Gelnarova,** Masaryk University, Czech Republic, **& Michael G. Schimek,** Medical University of Graz, Austria, & Masaryk University, Czech Republic

Array CGH experiments have become a powerful technique for analyzing changes in DNA, by comparing a test DNA to a reference DNA. These microarray experiments

produce a huge amount of data and special statistical techniques are required for detecting alterations. We introduce a new, rather simple method for the detection of breakpoints to find gene copy number changes in array CGH data. In the first part of the analysis, the quantile smoothing approach proposed by Eilers and Menezes (2005) is used as an important step of the data pre-processing. Based on the assumption of rank order dependence of the probes and the jump character of gene copy number changes (in log ratios) breakpoints are detected. The method is sequential and is based on monitoring changes in variability of the distribution of log ratios using a moving window of fixed width. The variability of distribution of log ratios is estimated in each window by a modified version of the median absolute deviation. The idea of the method is that the variability is increased in windows that cover breakpoints. When the variability of the window exceeds some critical level, the breakpoint is detected. The critical level is derived as quantile of the empirical distribution of variability of the dataset. The last step of the analysis includes the reduction of false positive breakpoints via cluster analysis of segments. The number of clusters is estimated using silhouette information. Performance of the method is demonstrated using simulated and publicly available data sets. Our approach is compared to three other methods.

**3:00 p.m.** *Building Phenotypic Hierarchies From Nested Effects of Gene Perturbations*, **Florian Markowetz**, Princeton University, **Olga G. Troyanskaya**, Princeton University, & **Rainer Spang**, Max-Planck-Institute for Molecular Genetics, Berlin, Germany

Functional genomics has a long tradition of inferring the inner working of a cell through analysis of its response to various perturbations. Observing cellular features after knocking out or silencing a gene reveals which genes are essential for an organism or for a particular pathway. A key obstacle to inferring genetic networks from perturbation screens is that phenotypic profiles generally offer only indirect information on how genes interact. To approach this problem, we will show how to infer features of the internal organisation of the cell from the nested structure of perturbation effects in high-dimensional phenotyping screens. We propose statistical methods to infer a quasi-order of gene perturbations from noisy subset relations between observed effects. We discuss a Bayesian score for inferring quasi-orders from data and introduce heuristic inference methods suitable for large-scale phenotyping screens. We show the applicability of our methodology for two real world datasets, an RNAi study of immune response in *Drosophila melanogaster* and a compendium of gene knock-outs in the yeast *Saccharomyces cerevisiae*.

**3:30 p.m.** *Floor Discussion*

---

**3:40 p.m. – 4:00 p.m.**

**BREAK**

---

**4:00 p.m. – 5:40 p.m.**

---

## **I-23**

### **Immunological Proteomics**

*Organizer: Keith Baggerly, MD Anderson Cancer Center*

**4:00 p.m.** *The Biology of Immunological Assays: A Statistician's Perspective on Measuring Protein*, **Keith Baggerly**, MD Anderson Cancer Center

Over the past several years, microarrays and related assays have caught on for the measurement of both mRNA (expression arrays) and DNA (SNP chips, CGH arrays). However, much of the biology of interest occurs at the protein level, and this stage of the DNA/RNA/Protein triad has proved more resistant to measurement. In this talk, we will describe why this step is hard, present the underlying biology of classical protein assays such as ELISAs and Western blots, and describe how these are being extended to produce array-like measurements now.

**4:30 p.m.** *Statistical Analysis of Reverse-Phase Protein Arrays*, **Shannon Neeley**, Rice University

Reverse-Phase Protein Arrays (RPPAs, aka protein lysate arrays, tissue lysate arrays, or lysate arrays) are recently developed tools for measuring protein expression levels in large numbers of samples. These assays are for the most part massively parallelized versions of enzyme-linked immunosorbent assays (ELISAs). In their massive parallelization, these assays are similar to cDNA microarrays (for mRNA) and CGH assays (for DNA). However, while those assays make thousands of measurements on a single sample ("forward-phase"), RPPAs measure one thing on hundreds of samples ("reverse-phase").

In this talk, we will describe issues we have encountered in modeling this data. In particular, we will address quantification, data processing, and describe some of the tools we have developed for this purpose. Open questions will be identified.

**5:00 p.m. *Issues in the Processing of Humoral Immune Response Array Data From Microarrays*, Debashis Ghosh, University of Michigan**

A major area of recent interest has been the development of autoantibody, and more generally humoral response, signatures for detection and prognosis of cancer. There are many advantages afforded to such an approach; however, there are new analytical challenges that arise in the analysis of such data. We highlight these problems, along with some basic tools we have used for the analysis of such data. Examples from a variety of cancer studies will be used throughout.

**5:30 p.m. *Floor Discussion***

## **I-24**

### **Inferring Genetic Networks from Genomics Data**

*Organizer: Grace S. Shieh, Academia Sinica, Taiwan*

**4:00 p.m. *Computational Challenges for Modelling and Simulating Biological Pathways*, Satoru Miyano, University of Tokyo**

If the language for modeling and describing biological pathways would not be rich, we would lose a lot of valuable knowledge and information on biological systems produced and reported. Placing this understanding as our basis of development, we have been developing an XML format Cell System Markup Language (CSML) and Cell System Ontology (CSO) 3.0 (<http://www.csml.org/>) and a modeling and simulation tool Cell Illustrator (CI) 3.0 (<http://www.cellillustrator.com/>).

CI 3.0 employs the notion of Hybrid Functional Petri Net with extension (HFPNe) as its architecture. HFPNe was defined by enhancing some functions to hybrid Petri net so that various aspects in pathways can be intuitively modeled, including integer, real, string, boolean, vector, objects, etc. The architecture of CI 3.0 is designed so that users can get involved with modeling and simulation in a biologically intuitive way with their profound knowledge and insights, and they can also be benefited from some public/commercial pathway databases. Recently, we have developed a method for automatic parameter estimation for HFPNe models by developing a theory of data assimilation that will be implemented as a function of CI.

Some XML formats are proposed to be a standard format for biological pathways. However, these formats provide only partial solutions for the storage and integration of biological data. The aim of CSML 3.0 is to provide a really usable XML format for visualizing, modeling and simulating biological pathways. CSML 3.0 is defined as an integrated/unified data exchange format which covers widely used data formats and applications, e.g. CellML 1.0, SBML 2.0, BioPAX, and Cytoscape. With CSML3.0, we

have also developed automatic conversion tools which convert SBML 2.0 to CSML 3.0 and CellML 1.0 to CSML 3.0 automatically. Cell Illustrator 3.0 fully supports CSML 3.0 as its base XML. Thus, every model in SBML 2.0 and CellML 1.0 can be executable on Cell Illustrator 3.0. Further, CSML 3.0 has focused on Hybrid Functional Petri net with extension (HFPNe) architecture, extended HFPN with object notion, for more advanced biological pathway representations. In short, objects that constitute biological pathways are treated as "generic entity" of HFPNe architecture and any relations among objects are treated as "generic process" on the HFPNe architecture. In addition, all tags and attributes in CSML 3.0 have the mapping rule from CSO 3.0, which is the ontology based terminology definition with OWL. The format consists of the following four features and the details are available from the web site: (a) A model with entity set and relation set on HFPNe. (b) Submodels of (a). (c) Views for each sub-model in (b). (d) Various biological terms, e.g. organism, cell type, tissue type, and molecular events, are introduced, most of them are inherited from BioPAX Level 2 and some of them are originally developed for dynamic simulation.

**4:30 p.m. *Inferring Genetic Regulatory Networks Using Genomics Data*, Hongyu Zhao**, Yale University

Transcription regulation is a fundamental biological process, and extensive efforts have been made to dissect its mechanisms through direct biological experiments and regulation modeling based on physical-chemical principles and mathematical formulations. Recent advances in high throughput technologies have provided substantial amounts and diverse types of genomic data that reveal valuable information on transcription regulation, including DNA sequence data, protein-DNA binding data, microarray gene expression data, and others. In this talk, we will present a Bayesian analysis modeling framework to integrate diverse data types, e.g. protein-DNA binding data and gene expression data, to reconstruct transcriptional regulatory networks. The usefulness of this general modeling approach is demonstrated through its application to infer transcriptional regulatory networks in the yeast cell cycle. This is joint work with Ning Sun and Ray Carroll.

**5:00 p.m. *Is Less More?: On Statistical Investigation for Large Biological Networks*, Henry Horng-Shing Lu**, National Chiao Tung University, Taiwan

Is it possible to develop simplified models to gain insights for large and complex biological networks? This talk will discuss our attempts to develop statistical methods for this purpose that include network reconstruction by Boolean networks, studies of yeast transcription factors and evolution of the yeast protein interaction network. Future developments regarding this direction will be discussed as well.

**5:30 p.m. *Floor Discussion***

**I-25**  
**Best of CAMDA 2006**

*Organizer: Patrick McConnell, Duke University*

**4:00 p.m.** *Bayesian Joint Disease-Marker-Expression Analysis Applied to Clinical Characteristics of Chronic Fatigue Syndrome*, [Madhuchhanda Bhattacharjee](#), University of Lancaster, U.K.,

From the wide range of clinical information available for the CFS study, after comparing with other possible alternatives, the Empiric variable (from Illness-class) was chosen as the most comprehensive summary of the disease. A further subset of illness related clinical variables were considered in various multi-locus association analyses where disease was explained either by a selected subset of markers, gene-expressions, or both of these. The additional clinical variables were used as stratifying factors to homogenize the study population and strata-specific marker and expression effects with respect to disease were studied. The WinBUGS software was used in implementation and parameter estimation.

**4:30 p.m.** *Integration of Genetic and Genomic Approaches for the Analysis of Chronic Fatigue Syndrome Implicates Forkhead Box N1*, [Steve Horvath](#), UCLA

Chronic fatigue syndrome (CFS) is a difficult disorder to characterize due to its combined physical and psychological effects. Studies on physiological effects have demonstrated involvement of the endocrine, immune, and muscular systems. We apply a weighted gene co-expression analysis to identify genes involved in chronic fatigue syndrome. Our analysis identified 184 candidate genes that had very similar expression patterns in fatigued individuals. Gene ontology analysis indicated that many of these genes were involved in muscle development and function, which is consistent with previous physiological findings.

**5:00 p.m.** *A Systems Genetic Analysis of Chronic Fatigue Syndrome: Combinatorial Data Integration From SNPs to Differential Diagnosis of Disease*, [Roumyana Kirova](#), Oak Ridge National Laboratory, TN

Understanding complex disease requires deep integration of information spanning multiple levels of biological scale. Small genetic variations among individuals act through biological networks, pathways and systems in both quantitative and qualitative fashion to produce higher order consequences manifest in disease susceptibility. Environmental factors interact with this backdrop of genetic variation to induce disease in predisposed individuals. The integration of such diverse information can be achieved computationally by simultaneous phenotyping and genotyping of a reference population of individuals. The CAMDA chronic fatigue syndrome data set is an excellent example of this approach to integrative systems biology of disease. Combinatorial algorithms can be applied to draw connections within and among the diverse data types. Differential diagnosis and development of therapeutic targets can be achieved by integrative analysis of diverse disease related data types.

**5:30 p.m.** *Floor Discussion*

**Saturday, 26<sup>th</sup> May 2007**

---

**8:30 a.m. – 10:10 a.m.**

---

**I-26**

**Providing Biological “Confidence” from Computational Analyses**

*Organizer: Gregory E. Gonye, Thomas Jefferson University*

**8:30 a.m. Pharmacovigilance Maps: Quantifying Associations Between Prescription Drugs and Adverse Reactions, Ronald Pearson**, Prosanos Corporation & Thomas Jefferson University

It has been estimated, on the basis of a meta-analysis of 39 U.S. hospital studies, that adverse drug reactions ranked between the fourth and sixth leading cause of death in 1994, behind heart disease, cancer and stroke, possibly behind pulmonary disease and accidents, but ahead of pneumonia and diabetes. This talk describes pharmacovigilance maps, a recently developed data mining method that can be applied to large spontaneous reporting databases like the FDA’s AERS database to quantify the association between prescription drugs and adverse drug reactions. The basis for this method is an urn model of the adverse event database, leading to an exact distribution for the number of records listing a given drug/adverse event combination under a null hypothesis of statistical independence. Adverse drug reactions can cause this number of records to be significantly larger than expected under the null hypothesis, while effective risk management programs can cause this number to be significantly smaller than expected under the null hypothesis. Specific examples of these and other results obtained from the AERS database will be presented.

**9:00 a.m. *Identification of Perturbation-Relevant Gene Regulatory Networks from Statistical Enrichment Analysis of Gene Co-Expression Clusters*, Rajanikanth Vadigepalli**, Thomas Jefferson University

High throughput experimental technologies, such as microarrays, in conjunction with the availability of large biological databases, such as fully sequenced and annotated genomes, offer an ideal situation for the use of computational methods and tools in the identification and analysis of gene regulatory networks. Present technological developments have resulted in rapidly growing public resources containing systematic data sets of various types: gene expression changes from microarrays; protein-DNA interaction and transcription factor activity data from protein binding assays, chromatin immunoprecipitation experiments, and DNA footprinting; protein-protein interactions from two hybrid experiments and coimmunoprecipitation; and genomic sequence and ontology information in public databases. Transcriptional Regulatory Network Analysis (TRNA) rests on use of these databases of relevant information and includes evaluation of the significantly enriched transcriptional regulatory elements, cellular processes and pathways. The hypotheses from TRNA are at multiple levels: on relevant regulatory factors and interactions, on the regulatory networks, and on the significant cellular processes and functions. In this context, we have developed a toolset named PAINT to automate TRNA. PAINT builds on distinct genome-scale databases to integrate promoter sequences and cross-reference annotation with pattern matching tools and statistical enrichment analysis methods. PAINT has been successfully employed in studies involving neuronal adaptation, circadian rhythms, retinal injury, blood cell development, bladder cancer, liver regeneration, hypoxia and pre-apoptosis. This talk will draw results from multiple case studies to illustrate how the statistical and bioinformatics aspects implemented in PAINT can be utilized to derive functionally relevant gene regulatory network hypotheses. The validated regulatory networks will then enable a targeted approach for novel therapeutic interventions.

**9:30 a.m. *Of microRNA Targets and microRNA Precursors*, Isidore Rigouts**, IBM Thomas J. Watson Research Center

In this presentation, I will discuss our work on the problem of microRNA target detection and microRNA precursor identification. In particular, I will describe rna22, a method that we have developed for addressing these two problems and discuss the predictions that it allows us to make. The method has several interesting properties including resilience in the presence of noise and lack of dependence on cross-species conservation. Our analyses suggest that some microRNAs may have as many as a few thousand targets, and that in higher organisms as much as 92% of the gene transcripts are likely under microRNA control through their 3'UTRs, 5'UTRs and amino-acid coding regions. With respect to the number of microRNA precursors in model organisms, our analysis indicates that it may be substantially higher than currently believed, likely ranging in the tens of thousands in some mammalian genomes. Representative results from experimental testing of our predictions across a variety of contexts will also be discussed in detail.

**10:00 a.m. Floor Discussion**

**I-27**

## **Statistical Learning and Management of Biological Data**

*Organizer: Parthasarathy Srinivasan, Ohio State University*

**8:30 a.m. Using Molecular Data to Estimate the Evolution of Species – Not Just the Phylogeny of Genes, Dennis Pearl, Ohio State University, & Liang Liu, Harvard University**

Recent work on the estimation of the posterior distribution of the evolutionary history of a gene given a set of molecular data has allowed for more realistic probability models and for more extensive problems to be addressed in a likelihood-based setting implemented through powerful programs like MrBayes. However, scientific interest generally lies in the evolutionary relationships of a group of species and not just in the phylogeny of genes. In this talk we present a Bayesian hierarchical model to estimate the phylogeny of a group of species using multiple estimated gene tree distributions such as those that arise in a standard Bayesian analysis of DNA sequence data. Our technique uses traditional Markovian models of sequence changes within a gene together with coalescent theory to explain genealogical signals from species trees to gene trees and from gene trees to sequence data. Our implementation of the method involves a "simple" adaptation of the MrBayes code followed by a post-processing algorithm using importance sampling. The method has been employed in several examples including the evolution of yeast, of finches, and of macaques.

**8:50 a.m. Machine-Learning Approaches for Drosophila Gene Expression Pattern Image Analysis, Jieping Ye, Arizona State University, & Sudhir Kumar**

The genetic analysis of spatial patterns of gene expression relies on the direct visualization of the presence or absence of gene products (mRNA or protein) at a given developmental stage (time) of a developing animal. The raw data produced by these experiments include images of the Drosophila embryos showing a particular gene expression pattern revealed by a gene-specific probe. The identification of genes showing spatial and temporal overlaps in their expression patterns is fundamentally important to formulating and testing gene interaction hypotheses. In this talk, I will present a computational framework for finding genes with overlapping expression patterns. The first step is to develop a learning system to identify automatically the developmental stage by image analysis. Once the developmental stage is determined, expression patterns are compared for spatial overlaps. Experiments on a collection of expression pattern images from the Berkeley Drosophila Genome Project (BDGP) illustrate the effectiveness of the proposed system.

**9:10 a.m. I-Sites 2007: Protein Local Structure Motifs Have Covariant Sequence Patterns, Chris Bystroff**

Amino acid sequence probability distributions, or profiles, have been used successfully to predict secondary structure and local structure in proteins. Profile models assume the statistical independence of each position in the sequence, but the energetics of protein folding is better captured in a scoring function that is based on pairwise interactions, like a force field. I-sites motifs are short sequence/structure motifs that populate the protein structure database due to energy-driven convergent evolution. Here we show that a pairwise covariant sequence model predicts local structure significantly better than a profile-based model. We present an updated local structure motif library which improves prediction of most types of local structure. We show that pairwise covariant knowledge-based potentials are statistically feasible and a better model for the folding of short peptide sequences.

**9:30 a.m. *Integration and Analysis of Complex, Heterogeneous and Interrelated Biological Data – Towards a New Generation of Biological Databases*, Golan Yona, Technion – Israel Institute of Technology & Cornell University**

Recent advances in biotechnologies enable biologists to measure different aspects of biological systems, from sequencing complete genomes, to determining the structures of proteins, detecting interactions on a genomic scale, and measuring RNA and protein expression on a cellular level. This massive flow of new biological data poses a grand challenge in computational biology - the challenge of data integration.

Biological entities are strongly related and mutually dependent on each other, and to fully understand the role of an individual entity one has to know which other entities are related to it. For example, the function of a specific gene is determined from its biological context: the set of genes it is related to, its subcellular location, the set of interactions it forms, the pathways it participates in, and so on. Therefore, there is a growing need for a unified biological resource that would corroborate and integrate data from different resources and aspects of biological systems.

In this talk I will describe the Biozon system ([biozon.org](http://biozon.org)), an extensive knowledge resource of heterogeneous biological data that links different biological objects, spanning entities from the molecular level to the organism level (including sequence, structure, protein-protein interactions, pathways, expression data and more) Biozon is based on a graph schema and the integration results in a highly connected graph structure that provides a more complete picture of the known context of each object, which cannot be determined from any one source. Informally, Biozon can be described as Amazon and Google, combined together and applied to the diverse biological knowledge domain.

Biozon merges the holdings of more than a dozen molecular biology collections, including SwissProt, KEGG, PDB, BIND, and others, and augments this data with novel in-house derived data such as sequence or structure similarity, predicted interactions, and predicted domains. Currently, Biozon holds more than 100 million biological documents and 6.5 billion relations between them.

Biozon allows complex searches on the data graph that specify desired interrelationships

between types (for example, 3D structures of all proteins that interact with the protein BRCA1). Moreover, Biozon has a fuzzy search engine that extends complex searches to include homologous sequences or structures as a search step, or even genes with similar expression profiles. One can search, for example, for all proteins that are known to take part in a specific pathway or proteins with similar expression profiles (associated with the corresponding mRNA sequences) to these proteins. Biozon also integrates first-of-a-kind biological ranking system which resembles the methods implemented in Google.

**10:00 a.m. *Floor Discussion***

## **I-28**

### **Analysis and Visualization of Microarray Gene Expression Data Using Excel, SAS, and S-Plus**

*Organizer: Richard M. Heiberger, Temple University*

**8:30 a.m. *An Office-Software- and Menu-Driven Interface to Advanced Statistics in Biological Sciences*, Erich Neuwirth, University of Vienna, Austria**

Making state-of-the-art research level statistics packages accessible to researchers in other subject disciplines is a challenging task. Statistics packages like R in essence are high level programming languages and do not offer the same level of userfriendliness as standard office packages. Integrating statistics packages into standard office programs therefore is a worthwhile undertaking. Our project demonstrates how R (and also Scilab) can be embedded in Microsoft Excel. We will discuss and demonstrate different choices for the user interface depending on how much the user is supposed to be exposed to these packages. We will show how to use Microsoft Excel as the user interface for the statistical engine of R and how to incorporate the results into a spreadsheet and into other office documents. We also will discuss the mechanisms for embedding R (and Scilab) into other Windows applications and give examples how this can be achieved relatively easy using our freely available toolkit.

**9:00 a.m. *Distribution of microarray analysis graphics*, Robert.C.Gagnon, Glaxo-Smith-Klein**

I describe an analysis of thousands of genes from a series of microarray gene expression experiments. The analysis and distribution of its results is based on several software systems, each used in the area of its strength as perceived by the primary user of that phase of the analysis. Microsoft Excel, a system which is on everyone's desk, is used for assembling multiple spreadsheets of data for the purpose of collection, storage, and manipulation. Multifactor analysis of variance is conducted in SAS, the most prevalent analysis system. Genes of interest are visualized using S-Plus trellis, the most flexible graphics system. Genes of interest and their annotations are stored back in Excel, with links to respective visualizations, stored as .jpg files generated from the S-Plus system.

**9:30 a.m. *Disseminating Statistical Methodology and Results via R and Excel: Two Examples*, Balasubramanian Narasimhan, Stanford University**

It is often the case that a collaboration with a colleague in need of statistical help leads to the development of methodology or tool for doing the statistical analysis. Increasingly, the implementation of the method or tool is in the widely used open source package R. The accessibility of the code or tool that is developed can be significantly enhanced by wrapping the implementation in an interface such as Excel that is widely used and familiar to users. In this talk I will present two examples, one involving the analysis of differentially expressed genes in microarray experiments and another involving classification and prediction for microarray experiments. Both are widely used software packages in microarray gene expression analysis, where once again the computation is done in R but the data and results are available in Excel.

**10:00 a.m. *Floor Discussion***

---

**10:10 a.m. – 10:30 a.m.**

**BREAK**

---

---

**10:30 a.m. – 12:10 p.m.**

---

## **I-29 Metrology and Inference of Complex Biological and Engineered Systems**

*Organizer: Z.Q. John Lu, National Institute of Standards & Technology (NIST)*

**10:30 a.m. *Standards in Microarray Gene Expression Experiments*, Marc Salit, NIST**

The use of standards in gene expression measurements with DNA microarrays is ubiquitous – they just aren't yet the kind of standards that have yielded microarray gene expression profiles that can be readily compared across different studies and different laboratories. They also aren't yet enabling microarray measurements of the known,

verifiable quality needed so they can be used with confidence in genomic medicine in regulated environments.

This presentation will provide an overview of some of the standards in use, some that are in development, and some emerging approaches that will support the maturation of this technology.

**11:00 a.m.** *Bayesian Semiparametric Methods for Pathway Analysis*, **Inyoung Kim**, Yale University

Pathways are sets of genes that serve a particular cellular or physiological function. The genes within the same pathway are expected to function together and hence may interact with each other. It is, therefore, of scientific interest to study their overall effect rather than each individual effect. Limited work has been done in the regression settings to study the effects of clinical covariates and large numbers of gene expression levels on a clinical outcome. In this paper we propose a Bayesian MCMC method for identifying pathways related to a clinical outcome based on the regression setting. A semiparametric mixed model (Liu, *et al.*, 2006) is used to build dependence among genes using covariance structure with Gaussian, Polynomial, and Neural network kernels. The clinical covariate effects are modeled parametrically but gene expression effects are modeled nonparametrically. All variance components and nonparametric effect of genes are directly estimated using Bayesian MCMC approach. We compare our Bayesian MCMC approach with the method proposed by Liu *et al.* (2006) which was developed by connecting a least squares kernel machine with a linear mixed model. We show that our approach is comparable with the Liu *et al.*'s approach based on type I error and power using simulation. Our simulation study also indicates that our approach has smaller mean squares error than the other method for estimating parameters. An example of type II diabetes dataset (Mootha *et al.*, 2003) is used to demonstrate our approaches. This is joint work with Herbert Pang and Hongyu Zhao.

**11:30 a.m.** *Sensitivity Analysis Methodology for a Complex System Computational Model*, **Jim Filliben**, NIST

Complex systems are of many types—biological (the human brain), physical (the World Trade Center collapse), social (the U.S. domestic aircraft transportation network), and informational (the Internet). Many such complex systems are successfully advanced by the judicious construction of computational models to serve as predictive surrogates for the system. The use of such models increasingly serves as the backbone for characterizing, modeling, predicting and ultimately optimizing the systems themselves.

The results of such efforts gain credence only after a proper V&V (verification and validation) of such computational models. This task itself is extremely difficult and can frequently be achieved only in a matter of degrees. In practice, an essential component in this V&V is an appropriate sensitivity analysis of the computational model. Gaining an appreciation of the dominant factors (and the ever-present interactions) of a computational model for a complex system is always an essential component in

accepting/rejecting such a model, and (after acceptance) in gaining deeper insights and understanding as to what actually drives the complex system itself.

This talk describes the methodology (experiment design and statistical analysis) which was brought to bear to carry out a sensitivity analysis for computational models for a specific complex informational system (a network). Such methodology has application to assessing models of other complex system types.

**12:00 noon.** *Floor Discussion*

## **I-30**

### **Metagenome Informatics**

*Organizers: Hongwei Wu & Ying Xu, University of Georgia*

**10:30 a.m.** *Assembling the Human Gut Biome*, **Mihai Pop**, University of Maryland

Most of what is known about the bacteria inhabiting the human gastro-intestinal tract is based on either culture-based studies or on the targeted sequencing of the gene encoding the small subunit of the ribosomal RNA. I will describe the first attempt at characterizing the gut micro-biome through high-throughput sequencing of DNA extracted directly from fecal samples. This metagenomic approach provided us with a relatively unbiased view of the genetic diversity, and functional potential within the GI tract. I will concentrate on both the methods used for the assembly and analysis of the sequence data, and on several of the most interesting findings of this study.

**10:50 a.m.** *Benchmarking the Fidelity of Metagenomic Sequences Using Simulated Datasets*, **Kostas Mavrommatis**, Joint Genome Institute

Metagenomics is a rapidly emerging field of research for studying microbial communities. To evaluate methods currently used to process metagenomic sequences, we constructed three simulated datasets of varying complexity by combining sequencing reads randomly selected from 113 isolate genomes sequenced at the DOE Joint Genome Institute. These datasets were designed to model real metagenomes in terms of complexity and phylogenetic composition. Sampled reads were assembled using commonly used genome assemblers and genes were predicted using available gene finding pipelines. The phylogenetic origins of the assembled contigs were predicted using similarity and sequence composition binning methods. We explore the effect of simulated community structure and method combinations on the fidelity of each processing step by comparison to the isolate genome assemblies and gene calls. A number of limitations with the existing methods were revealed. The simulated datasets and tools to facilitate the analysis have been made available online, and should facilitate standardized benchmarking of tools for metagenomic analysis.

**11:10 a.m. *On Incorporating Genomic Neighborhood Information for the Prediction of Orthologous Genes*, Hongwei Wu, University of Georgia**

Most of the existing methods for the prediction of orthologous gene groups, including those clustering-based and those phylogenetic tree-based, only utilize the sequence similarity information of genes. For prokaryotic genomes, the genomic neighborhood information, e.g., operons and directons, reflects the functional relatedness among genes, and can therefore complement the sequence similarity information and be used for the prediction of orthologous genes. We have observed that orthologous gene pairs are more likely than non-orthologous gene pairs to have companionate homologous gene pairs in their genomic neighborhoods. Inspired by this observation, we have developed a computational method to predict orthologous gene groups for prokaryotes based on co-occurrences of homologous gene pairs in the same genomic neighborhoods. Given a set of prokaryotic genomes, we have first clustered directons into non-overlapping directon groups, and then cluster homologous genes belonging to the same directon group into non-overlapping gene groups. Our primary analyses through comparisons with the COG, KO and Pfam systems have shown that our method can not only reveal the functional relatedness among different gene groups, but is also promising to provide more accurate and specific predictions for the orthologous relationships among genes.

**11:30 a.m. *Exploring Marine Metagenomics via the CAMERA (Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis) Project: An Initial Analysis of Protein Family Diversity Found in the Global Ocean Sampling Study*, WeiZhong Li, Adam Godzik, John Wooley, & the CAMERA team, University of California, San Diego**

The volume of genomic data in public repositories for DNA and RNA sequence data continues to double at almost a yearly frequency and recently passed 100 gigabases of genomic sequence. Metagenomics will further accelerate this trend by sequencing samples of the entire environment *en masse* using novel sequencing technologies. To provide an avenue for studying complex ecosystems, for example, the environmental sequence data can be associated with a considerable range of metadata such as the temperature, chemistry (available nutrients), weather, time and spatial location of the sampling site. A new internet is also emerging, one in which dedicated fiber optical circuits allow researchers to connect at bandwidth speeds 1000 times greater than today's traditional shared internet. The CAMERA project is a robust implementation of technology supporting biological data analysis. Our website currently provides three marine metagenomic datasets, including around 8 million sequence reads and their assorted metadata from the Global Ocean Sampling expedition, along with sequence comparison tools. The resource will routinely add additional environmental datasets and new software tools and provides the means to conduct rigorous analysis of metagenomic data. New strategies for analysis are required to explore the vast collection of sequences – already more than in GenBank - being discovered in metagenomics and provided through CAMERA, including the largest current subset, the Global Ocean Sampling (GOS) obtained by the J. Craig Venter Institute. To explore the diversity of protein families in the GOS ORFs, we performed a hierarchical clustering analysis on the GOS ORFs using the newly modified CD-HIT algorithm and identified

34,803 protein families of >20 non-redundant ORFs. We also used the same clustering strategy on the NCBI non-redundant (NR) protein databases and identified 14,216 families. We found that out of the 34,803 GOS families, 22,377 have significant similarities to known proteins in NR that can be detected by BLAST. Using more sensitive homology detection methods, such as PDB-BLAST and FFAS, we identified an additional 4,252 families as having remote but statistically significant similarities to known proteins with the remaining 8,174 families predicted to represent novel protein families. Considering the need for extensive computational analysis of large metagenomic data sets, we note that using alternative clustering strategies, this entire project took about 1.3% of the computational effort used in the depositors' original analysis.

**12:00 noon. *Floor Discussion***

## **I-31**

### **Tutorial on the Bridge Between Classical & Systems Biology: Systems Models of Nature, Uncertainty, and Real Collaboration**

*Organizer: Arnold Goodman, University of California, Irvine*

**10:30 a.m. *A Starter Kit for Systems Biology: First Useful and Usable System Model of Protein Cycle, Pursuit of Uncertainty, and Interdisciplinary Collaboration, Arnold Goodman***, University of California, Irvine

Paul Silverman (*The Scientist* 18(10): 32-33) is first to explicitly envision need to create a new model of the protein cycle and explore it for uncertainty. Author led development of first useful and usable system model of the protein cycle, and introduced uncertainty as a missing factor in cell behavior (*The Scientist* 19(12): 20-21). Model has 5 interactive stages with inputs, operations and outputs -- and is being adapted to cell division cycle.

Far too much biology, chemistry and physics must succeed to determine cell behavior. Until disproved, cell behavior is composed of “dogma-determined”, “determinably-unknown”, and “describably-uncertain” parts. Key cell variables are random, their equations are random, and statistics is needed for the uncertainty. Cell biology progress depends heavily upon interdisciplinary collaboration and value-added problem solving.

Value-Added Problem Solving: Problem Agreement (Basis), Data and Theory, Models and Software, Solution Agreement, and Value Added (Goal) in terms of new insights, effective decisions and/or productive actions. Critical success factors of Collaboration in problem solving are: Commitment, Communication, Community and Control.

The dream of “systems biology” will require useful and usable system models, active pursuit of uncertainty, interdisciplinary collaboration and value-added problem solving.

**12:00 noon.** *Floor Discussion*