Banff International Research Station
for Mathematical Innovation and Discovery

Computational and Statistical Genomics

BIRS Workshop 06w5076
July 08–13, 2006
www.stat.berkeley.edu/~biolab/BIRS06

ORGANIZERS

Jenny Bryan (Department of Statistics; University of British Columbia)
Sandrine Dudoit (Division of Biostatistics; University of California, Berkeley)
Sündüz Keleş (Department of Statistics and Department of Biostatistics and Medical Informatics; University of Wisconsin, Madison)
Katherine S. Pollard (Department of Statistics and Genome Center; University of California, Davis)

MEALS

Continental Breakfast: 07:00–09:00, 2nd floor lounge, Corbett Hall, Sunday–Thursday
* Buffet Lunch: 11:30–13:30, Donald Cameron Hall, Sunday–Thursday
* Buffet Dinner: 17:30–19:30, Donald Cameron Hall, Saturday–Wednesday
Coffee Breaks: As per daily schedule, 2nd floor lounge, Corbett Hall
Beverages and small assortment of snacks available in 2nd floor lounge, Corbett Hall, on a cash honour-system.
* **N.B.** *Please remember to scan your meal card at the host/hostess station in the dining room of the Donald Cameron Hall for each lunch and dinner.*

MEETING ROOMS

All invited lectures are held in 159 Max Bell.
The Max Bell Building is accessible by the bridge on the 2nd floor of Corbett Hall. Hours 06:00–00:00 (midnight).
LCD projectors, overhead projectors, and blackboards are available for the presentations.
All lectures are 40 minutes long, followed by a 5-minute question period.
Please note that the meeting space designated for BIRS is in the lower level of the Max Bell Building, Rooms 155–159. Please respect that all other space has been contracted to other Banff Centre guests, including any food and beverage in those areas.

SCHEDULE

Day 0 — Saturday, July 08

16:00	Check-in, Front Desk, Professional Development Centre – open 24 hours
17:30–19:30	Dinner
20:00	Informal gathering, 2nd floor lounge, Corbett Hall

Day 1 — Sunday, July 09

07:00–08:45	Breakfast
08:45–09:00	Introduction and welcome to BIRS by BIRS Station Manager, 159 Max Bell
	Population Genomics
09:00–09:45	Rachel Brem, *Genome-wide evolutionary rates in lab and wild yeast*
09:45–10:30	David Hinds, *Statistical issues in analysis of large scale genetic association studies*
10:30–11:00	Coffee Break
11:00–11:45	Sündüz Keleş, *Integrating quantitative information from ChIP-chip experiments into motif finding*
11:45–14:00	Lunch ... and World Cup Final ;)
	Regulomics and Proteomics
14:00–14:45	Imola Fodor, *Host-pathogen science: Statistical opportunities*
14:45–15:30	Hongyu Zhao, *Bayesian error analysis model for reconstructing transcriptional regulator networks*
15:30–16:00	Coffee Break
16:00–16:45	Ingo Ruczinski, *Of mutants and men: Computational and statistical tools relevant for the exploration of the protein folding process*
16:45–17:30	Discussion
17:30–19:30	Dinner
20:00–21:30	Poster Session

Day 2 — Monday, July 10

07:00–09:00	Breakfast
	Microarray Gene Expression Analysis
09:00–09:45	Terry Speed, *A method for detection of alternative splicing from exon array data*
09:45–10:30	Fitnat Yildiz, *Molecular basis and consequences of phenotypic variation in Vibrio cholerae*
10:30–11:00	Coffee Break
11:00–11:45	Rebecka Jornsten, *MIXL - Mixture modeling with multiple levels and flexible parameterizations*
11:45–13:00	Lunch
13:00–14:00	Guided tour of The Banff Centre; meet in 2nd floor lounge, Corbett Hall
	Microarray Gene Expression Analysis
14:00–14:45	Aseem Z. Ansari, *CSI analysis: Defining the sequence recognition profile of DNA binding molecules*
14:45–15:15	Coffee Break
15:15–16:00	Steve Horvath, *Connectivity, module-conformity, and significance: Understanding gene co-expression network methods*
16:00–16:45	Discussion
17:30–19:30	Dinner
20:00–21:30	Software Demo Session

Day 3 — Tuesday, July 11

07:00–09:00	Breakfast
	Phylogenomics and Comparative Genomics
09:00–09:45	Wyeth Wasserman, *TBA*
09:45–10:30	Adam Siepel, *Searching for novel functional elements in mammalian genomes by comparative genomics*
10:30–11:00	Coffee Break
11:00–11:45	Bret Larget, *A Bayesian approach to gene tree concordance*
11:45–13:30	Lunch
13:30–17:30	Free Afternoon
17:30–19:30	Dinner

Day 4 Wednesday, July 12

07:00–09:00	Breakfast
	Integromics
09:00–09:45	Mark Segal, *Validation in genomics: CpG island methylation revisited*
09:45–10:30	Tim Hughes, *Decoding the vertebrate regulome*
10:30–11:00	Coffee Break
11:00–11:45	Jason Lieb, *Quantitation of chromatin contributions to transcription factor targeting*
11:45–13:00	Lunch
13:00–13:15	Group photo, front steps of Corbett Hall
	Integromics
13:15–14:00	David Rocke, *Statistical methods for identification of differentially expressed gene groups and pathways*
14:00–14:45	Joaquín Dopazo, *From genes to functional blocks in the study of biological systems*
14:45–15:15	Coffee Break
15:15–16:00	John Quackenbush, *Stochasticity and networks in genomic data*
16:00–16:45	Discussion
17:30–19:30	Dinner

Day 5 Thursday, July 13

07:00–09:00	Breakfast
	Statistical Computing and Bioinformatics
09:00–09:45	Duncan Temple-Lang, *Integrating software for data analysis*
09:45–10:30	Torsten Hothorn, *mboost: A package for model-based boosting*
10:30–11:00	Coffee Break
11:00–11:45	Discussion
11:45–12:00	Closing remarks by organizers
12:00–13:30	Lunch

*** Checkout by 12 noon.**

*** N.B.** *Participants for 5-day workshops are welcome to use the BIRS facilities (2nd floor lounge of Corbett Hall, Max Bell Meeting Rooms, Reading Room) until 15:00 on Thursday, although participants are still required to checkout of the guest rooms by 12:00 noon.*

Banff International Research Station

for Mathematical Innovation and Discovery

Computational and Statistical Genomics

July 08–13, 2006

ABSTRACTS: INVITED LECTURES

www.stat.berkeley.edu/~biolab/BIRS06/lectures.html

Speaker: Professor Aseem Z. Ansari (Department of Biochemistry and The Genome Center of Wisconsin, University of Wisconsin, Madison)
Title: *CSI analysis: Defining the sequence recognition profile of DNA binding molecules*
Abstract: A central goal of synthetic biology, chemical biology, and molecular medicine is the design and creation of synthetic molecules that can target specific DNA sites in the genome. Such molecules can be harnessed to regulate biological processes such as transcription, recombination, and DNA repair. However, a major hurdle in the design of sequence-specific DNA binding molecules is the inability to comprehensively define the full range of their DNA sequence recognition properties and therefore predict all their potential target sites in the genome. Toward this goal we have developed a high-throughput approach that provides a comprehensive profile of the binding properties of DNA binding molecules (1). The approach is based on displaying every permutation of a duplex DNA sequence (up to 10 positional variants) on a micro-fabricated array. The entire sequence space is interrogated simultaneously and the affinity of a DNA binding molecule for every sequence is obtained in a rapid, unbiased, and unsupervised manner. Using this platform we have determined the full molecular recognition profile of an engineered small molecule as well as a eukaryotic transcription factor. The approach also yielded unique insights into the altered sequence recognition landscapes due to cooperative assembly of DNA binding molecules in a ternary complex. Solution studies strongly corroborated the sequence preferences identified by the array analysis.

1) C. L. Warren, et al. (2006) PNAS 103, 867–872.

Speaker: Professor Rachel Brem (Department of Molecular and Cell Biology, University of California, Berkeley)
Title: *Genome-wide evolutionary rates in lab and wild yeast*
Abstract: Wild organisms, when introduced into the laboratory, often undergo selection for easier growth and reproduction. This initial adaptation, and further genetic changes during propagation in the lab, may affect a wide range of genotypes and phenotypes, such that the biology of a lab organism may no longer reflect that of wild populations. This concern has been discussed in the recent literature for the model organism S. cerevisiae. We studied a common lab strain of yeast in the context of the two wild yeast isolates whose whole-genome sequences are currently available. We found that one of the two wild strains, an isolate from a California vineyard, exhibited a higher rate of protein evolution than the lab strain. Protein evolutionary rates along the recent lineages of closely related yeast strains were similar, consistent with a model in which no single strain has been through a uniquely severe bottleneck or regime of positive selection in the recent past. Our work provides preliminary evidence that the lab strain has not recently accumulated a load of deleterious, protein-coding mutations over and above what is observed in natural yeast habitats. This suggests that on the genomic scale, laboratory strains can be considered a reasonable model for wild yeast, a conclusion which has implications for the genetics of many domesticated and experimental organisms.

Speaker: Dr. Joaquín Dopazo (Centro de Investigación Príncipe Felipe, Valencia, Spain)
Title: *From genes to functional blocks in the study of biological systems*
Abstract: With the popularisation of high-throughput techniques, the necessity of procedures for the biological interpretation of the results has increasing enormously. The procedures used are inefficient, based on thresholds imposed on the experimental values that do not take into account the functional correlations existing among the genes. New procedures more inspired in systems biology criteria are under development. Here we present a threshold-independent test for functional annotation which do not depend on the type of the data for obtaining significance values and consequently can be applied in a large number of genome-scale studies. We exemplify its application in evolution, microarray gene expression data and interactomics data.

Availability: A web server that performs the test described and other similar can be found at: `www.babelomics.org`.

Speaker: Dr. Imola K. Fodor (Bioinformatics and Biostatistics, Lawrence Livermore National Laboratory)
Title: *Host-pathogen science: Statistical opportunities*
Abstract: Host-pathogen interaction studies provide a unique perspective into infectious diseases that still pose major public health concerns worldwide. While technological advancements (e.g. microarrays, protein arrays, phenotype arrays, mass spectrometry) facilitate single experiments, the complexity of the problem (e.g. variation within individuals and among heterogeneous populations, evolving pathogens, integration of results obtained from disparate platforms or experiments using differing protocols) renders the systematic evaluation of pathogens and their effects on the host extremely challenging. Statistical opportunities will be presented, with the long term goal of improving disease signature detection and developing therapeutic interventions. Collaboration with the domain scientists is essential for progress.

This work was performed under the auspices of the US Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, with support from the Department of Homeland Security (Biological Countermeasures Program). UCRL-ABS-222457.

Speaker: Dr. David Hinds (Perlegen Sciences)
Title: *Statistical issues in analysis of large scale genetic association studies*
Abstract: Analysis of large association studies is challenging due to the extremely small expected proportion of rejected null hypotheses. A consequence is that results can be strongly influenced by very low frequency deviations from asymptotic distributions, due either to unusual biology or experimental artifacts. I will discuss some work we have done in this context for large scale oligonucleotide array-based association studies.

Speaker: Professor Steve Horvath (Departments of Human Genetics and Biostatistics, University of California, Los Angeles)
Title: *Connectivity, module-conformity, and significance: Understanding gene co-expression network methods*
Abstract: Many real networks have been found to contain clusters (modules) of highly interconnected nodes. In gene networks, modules may correspond to pathways. Within modules, highly connected 'hub'-nodes have often been found to be biologically or clinically interesting targets. Intramodular connectivity has arisen as an important concept for identifying biologically or clinically significant genes in a gene network. This paper presents theoretical and empirical results that show how intramodular connectivity, the clustering coefficient and other fundamental network concepts are related to each other. We consider undirected networks that can be represented by a symmetric, non-negative 'adjacency' matrix. In our applications involving gene co-expression networks, we find that the adjacency matrices of sub-networks comprised of module genes are approximately 'factorizable' which means that the local properties of each

node are determined by its ‘conformity’. In general, conformity is highly correlated to the node connectivity. We define the concept of node conformity for general networks and use it to relate standard and novel network concepts to each other.

We derive several theoretical results for gene co-expression networks that are constructed on the basis of gene expression profiles across microarrays. By decomposing the gene expression profiles of genes inside a module we arrive at the notion of a module eigengene, which effectively summarizes the gene expression profile of the entire module. We explain the meaning of fundamental network concepts inside a module in terms of the corresponding module eigengene. For example, we show that the module conformity of a gene is determined by the correlation between the gene expression profile and the module eigengene.

We provide a simple model for explaining why intramodular hub genes have often been found to be biologically significant. Starting with an external microarray sample trait, we define a measure of gene significance by the absolute value of the correlation between a gene expression profile and the sample trait. Similarly, we define a measure of module relevance by the absolute value of the correlation between module eigengene and the sample trait. We prove that in relevant modules there is a linear relationship between intramodular connectivity and gene significance. We illustrate our theoretical findings by studying the properties of a brain cancer gene network. Further apply our methods to a yeast gene co-expression network.

Joint work with Jun Dong.

Speaker: Dr. Torsten Hothorn (Friedrich-Alexander-Universität Erlangen-Nürnberg)
Title: *‘mboost’: A package for model-based boosting*
Abstract: Classically, boosting or functional gradient descent algorithms for optimizing various empirical risk functions have been implemented using relatively complex base-learners such as regression trees. The resulting regression fit is typically a ‘black box machine’, i.e., one can predict the outcome based on the covariate status of a new observation but the fitted functional form is too complex to be readable by human beings. Recently, boosting algorithms for fitting generalized linear or additive models have been suggested. The key innovation is the application of componentwise linear models or smoothing splines which allows us to reformulate the regression fit in terms of classical linear or additive models. In the former case, the regression coefficients can be interpreted in the usual way. Moreover, those boosting algorithms have been demonstrated to be useful for variable selection in high-dimensional situations as they typically occur in omic-statistics.

Joint work with Peter Bühlmann (ETH Zurich).

Speaker: Professor Tim Hughes (Banting and Best Department of Medical Research, University of Toronto)
Title: *Decoding the vertebrate regulome*
Abstract: We would like to be able to explain the causes of vertebrate gene expression patterns in terms of combinations of cis- and trans-regulatory activities. There are indications that this goal is attainable in yeast, which has 200 transcription factors and one or a few conserved elements in the promoter of each gene. However, vertebrate genomes encode >1,500 DNA-binding transcription factors, and contain roughly one million nonexonic elements that are candidate cis-regulators. Precise binding specificity is known for only a small minority of the transcription factors, and an even smaller minority of the conserved elements have been associated with a known or putative binding factor.

We are taking two general approaches towards generating new laboratory data that we believe will aid in decoding the vertebrate regulome. The first is a brute-force attempt to determine the DNA-binding specificity of as many transcription factors as possible in the mouse, which we believe will enable us to classify potential cis-regulatory elements and determine which combinations are predictive of observed gene expression patterns. The second approach is to examine the phylogeny of gene expression patterns, in an attempt to take advantage of the fact that both cis-regulatory elements and gene expression patterns display varying degrees and varying patterns of conservation among vertebrate species. Although both of these projects are in early stages, I will present evidence that both of them are feasible and productive, indicating that this work will greatly enhance our understanding of the activity and evolution of vertebrate transcriptional regulatory networks.

Joint work with Mike Berger, Martha Bulyk, Gwenael Badis-Breard, Esther Chan, Xiaoyu Chen, Quaid Morris, Anthony Philippakis, Shaheynoor Talukder.

Speaker: Professor Rebecka Jornsten (Department of Statistics, Rutgers University)
Title: *MIXL - Mixture modeling with multiple levels and flexible parameterizations*
Abstract: Model-based clustering is a popular tool for summarizing high-dimensional gene expression data. By grouping genes according to a common experimental expression profile we can gain new insight into the biological pathways that steer biological processes of interest. Clustering of gene profiles can also assist in assigning functions to genes that have not yet been functionally annotated.

Model-based clustering has to-date primarily been applied in a single-level setting; that is, a gene profile is defined across all experimental factor levels, regardless of whether one or more factors are studied. This can lead to a very inefficient model representation. For example, consider a two-factor experiment with factors ”time” and ”cell-line”. If a particular time-pattern is common to both cell-lines with the exception of an offset of overall gene expression, such gene clusters need only be distinguished by one offset-parameter, not a complete time and cell-line parameter vector. Furthermore, model selection in model based clustering has almost exclusively focused on determining which variables/experiments differ between the clusters, leaving the mean structure within a cluster open to subjective interpretation.

We propose a mixture model with multiple levels, MIXL, where we take the multi-factor experimental design explicitly into account in the mixture model parameterization. We also discuss several flexible parameterizations for model selection within a cluster. We demonstrate the multi-level clustering approach on several simulated data sets and a proliferating cell-line data set. We discuss how a multi-level approach can assist in detecting biologically relevant groups of genes that may be missed with a less efficient parameterization. We use our multi-level parameterization as a basis for mining the expression data for annotational context and regulatory motifs.

Speaker: Professor Sündüz Keleş (Department of Statistics and Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison)
Title: *Integrating quantitative information from ChIP-chip experiments into motif finding*
Abstract: Identifying binding sequences of transcription factors, which are generally 5-20 base pairs (bps) in length, within long segments of non-coding DNA is a challenging task. Recently developed ChIP-chip technology utilizing tiling arrays is especially promising since it provides high resolution genome-wide maps of the interactions between transcription factors and DNA. A two step paradigm is commonly used for performing motif searches based on ChIP-chip data. First, candidate bound sequences that are in the order or 500-1000 bps are inferred from the experimental ChIP-chip data. Then, motif searches are performed on these sequences. These two steps are generally treated as disconnected and the quantitative nature of the ChIP-chip information is ignored in the second step. More specifically, all candidate bound regions are assumed to equally likely have the motif(s) of interest and the motifs are assumed to reside at any position of the bound regions with equal probability. We develop a conditional two component mixture model that relaxes both of these assumptions by adaptively incorporating quantitative information from ChIP-chip experiments into motif finding. The performance of the new and existing methods are compared using simulated data and ChIP-chip data from the recently available ENCODE studies. Initial results indicate that adaptive utilization of the ChIP-chip data provides substantially better sensitivity and specificity for motif finding.

Joint work with Heejung Shim.

Speaker: Professor Bret Larget (Department of Statistics and Department of Botany, University of Wisconsin, Madison)
Title: *A Bayesian approach to gene tree concordance*
Abstract: Bayesian phylogenetics involves the estimation of evolutionary relationships from genetic data. It is not unusual for different genes to support different evolutionary histories. The two most common strategies to deal with this are to either combine all of the data into a single analysis that assumes a single common tree or to make separate independent estimates for each gene. We describe an approach for the estimation of several gene trees between these two extremes that accommodates the possibility of multiple different gene trees, but also incorporates information from all genes through a Dirichlet process prior distribution on the map from genes to trees. We use a novel two-stage Markov chain Monte Carlo approach for calculations for this problem and demonstrate the results on an example from yeast genomics.

Speaker: Professor Jason Lieb (Department of Biology, University of North Carolina, Chapel Hill)
Title: *Quantitation of chromatin contributions to transcription factor targeting*
Abstract: Sequence motifs that are potentially recognized by DNA-binding proteins occur far more often in genomic DNA than do observed in vivo protein-DNA interactions. To determine how chromatin influences the utilization of particular DNA binding-sites, we compared the in vivo genome-wide binding location of the yeast transcription factor Leu3 to the binding location observed on the same genomic DNA in the absence of any protein cofactors. We found that the DNA-sequence motif recognized by Leu3 in vitro and in vivo was functionally indistinguishable, but Leu3 bound different genomic locations under the two conditions. Accounting for nucleosome occupancy in addition to DNA sequence motifs significantly improved the prediction of protein-DNA interactions in vivo, but not the prediction of sites bound by purified Leu3 in vitro. Measurements of nucleosome occupancy in strains that differ in Leu3 genotype show that low nucleosome occupancy at loci bound by Leu3 is not a consequence of Leu3 binding. These results permit, for the first time, quantitation of the epigenetic influence that chromatin exerts on DNA binding-site selection, and provide evidence for an instructive, functionally important role for nucleosome occupancy in determining patterns of regulatory factor targeting genome-wide. In addition, we present evidence that nucleosome positioning can be important for specifying environment-specific targets. We show Rap1, a master regulator of yeast metabolism, binds to an expanded target set upon nutrient depletion despite decreasing protein levels and no evidence of posttranslational modification. Whole-genome measurements of nucleosome occupancy and the distribution of the chromatin regulator Tup1 provide evidence for a mechanism of dynamic target specification that coordinates the genome-wide distribution of intermediate-affinity DNA sequence motifs with chromatin-mediated regulation of accessibility to those sites.

Cheol-Koo Lee (Lieb lab), and Xiao Liu and Joshua A. Granek in the laboratory of Neil D. Clarke (Genome Institute of Singapore, `clarken@gis.a-star.edu.sg`) did the Leu3 experiments and contributed to the analysis. Michael J. Buck (Lieb Lab) performed the Rap1 experiments and analysis.

Speaker: Professor John Quackenbush (Dana-Farber Cancer Institute and the Harvard School of Public Health)
Title: *Stochasticity and networks in genomic data*
Abstract: Two trends are driving innovation and discovery in biological sciences: technologies that allow holistic surveys of genes, proteins, and metabolites and a realization that biological processes are driven by complex networks of interacting biological molecules. However, there is a gap between the gene lists emerging from genome sequencing projects and the network diagrams that are essential if we are to understand the link between genotype and phenotype. 'Omic technologies such as DNA microarrays were once heralded as providing a window into those networks, but so far their success has been limited. Although many techniques have been developed to deal with microarray data, to date their ability to extract network relationships has been limited. We believed that by imposing constraints on the networks, based on associations reported through articles indexed in PubMed, we could more effectively extract biologically

relevant results from microarray data and develop testable hypotheses that could then be validated in the laboratory. Using literature networks as constraints on a Bayesian network analysis of microarray data, we show that we are able to recover evidence for a wide range of known networks and pathways, even in experiments not explicitly designed to probe them.

With a putative gene-interaction network, the problem of producing viable models of the cell remains. While systems biology approaches that attempt to develop quantitative, predictive models of cellular processes have received great attention, it is surprising to note that the starting point for all cellular gene expression, the transcription of RNA, has not been described and measured in a population of living cells. To address this problem, we propose a simple (and obvious) model for transcript levels based on Poisson statistics and provide supporting experimental evidence for genes known to be expressed at high, moderate, and low levels. Although what we describe as a microscopic process, occurring at the level of an individual cell, the data we provide uses a small number of cells where the echoes of the underlying stochastic processes can be seen. Not only do these data confirm our model, but this general strategy opens up a potential new approach, Mesoscopic Biology, that can be used to assess the natural variability of processes occurring at the cellular level in biological systems.

Speaker: Professor David M. Rocke (Division of Biostatistics, University of California, Davis)
Title: *Statistical methods for identification of differentially expressed gene groups and pathways*
Abstract: There are a number of reasons why it may be important to search for differential expression of groups of genes and pathways rather than for differential expression of individual genes. If biological samples are taken at a fixed time following an intervention, a transcriptional cascade may occur at different speeds in different individuals. At a fixed time, the differential transcription will then lie in different genes within the pathway for different individuals, thus resulting in a signal that occurs in one gene in the pathway for some individuals and other genes for other individuals. Polymorphisms and differences in physiology can result in differential expression of one gene from a class (e.g., MAP Kinases) in one individual and other genes from the same class in other individuals. Up- or down-regulation may be broad across a class of genes but with a signal that is too diffuse and weak to be detected in the results from individual genes. The greater power from aggregation of results may increase the sensitivity.

The earliest methods for handling this general problem class involved computing whether the gene group at issue is over-represented in a set of, for example, significantly differentially expressed genes. If a statistical test is used for each gene on the array, and if the genes are ordered by the p-value of the test, then one can define a cross tabulation by (in the gene group)/(not in the gene group) and (p ¡ c)/(p ¿= c), where c is some cutoff in the p-value, and then test for association between the gene group and the significance cutoff. Since the cutoff is arbitrary, one can instead conduct a test for all cutoffs simultaneously, which turns out to be equivalent to a Kolmogorov-Smirnov test of the distribution of p-values in the gene group against the distribution in genes not in the group. This is sometimes called Gene Set Enrichment Analysis (GSEA).

These procedures are robust to a certain extent, because they only depend on the ranking of the p-values, but there are several disadvantages. In the first place, the fact that the actual test results are not used but only the ranking is not always an advantage, particularly if the statistical test for each gene producing the p-value is itself robust. In the second place, the procedure requires a single test result per gene. In some cases, as in the illustrative example, results are available for each gene and for each individual, and some means of using all of this information needed to be found. In Rocke et al. (2005), a new method of analyzing gene groups and pathways was introduced, called the Test of Test Statistics (ToTS) method, in which a set of test statistics, such as in the case of the example a t-test for the dose-response slope, is tested for a positive or negative bias using the Wilcoxon one-sample test. This proved to be much more powerful than tests of individual genes. In order to make the procedure robust to correlations in the tests, a re-sampling based method is used to determine significance, rather than the usual asymptotic p-values for the Wilcoxon test. We have subsequently investigated a number of variant methods for summarizing the gene groups, and found some even more effective ones.

Speaker: Professor Ingo Ruczinski (Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University)
Title: *Of mutants and men: Computational and statistical tools relevant for the exploration of the protein folding process*
Abstract: The protein folding process is an example of a highly specific, spontaneous, nanometer-scale self-assembly process. It also lies at the heart of an increasing number of disease states, such as Alzheimer's, Parkinson's and Huntington's disease. Lastly, and perhaps most centrally, it is the process by which the information encoded in our genes is expressed in the form of the functional catalysts that promote and control every aspect of cellular life. Thus motivated, the study of protein folding has exploded in the last decade, with literally thousands of biochemical research groups participating in its exploration.

Of particular interest to the community of protein folding aficionados is the structure that the protein adopts in the folding transition state (the state with highest free energy), as that state forms the barrier that ultimately defines the folding pathway. Moreover, there is some evidence that it is this state from which disease-causing aggregation often occurs. Unlike the structure in the initial, unfolded state and the final, folded state however, the structure in the transition state cannot directly be assessed by experimental means such as X-ray crystallography or Nuclear Magnetic Resonance (NMR)as X-ray crystallography or Nuclear Magnetic Resonance (NMR) Spectroscopy. Therefore, researchers have to rely on different means to assess structure and mechanism in the folding process. The most common approach is by introducing mutants at different sequence loci. Using the kinetic data derived by those experiments, information can be obtained on which parts of a protein are participating in interactions in the transition state, and which are not, yielding critical information about the nature of this defining conformation.

In this presentation we discuss the statistical analysis of the kinetic data generated by the wet labs, the inference about the protein folding pathway, discuss protein stability and folding kinetics in vitro and in silico, and show how this information is also relevant for example for genome-wide protein structure prediction, and potentially, drug design.

Speaker: Professor Mark R. Segal (Department of Epidemiology and Biostatistics, University of California, San Francisco)
Title: *Validation in genomics: CpG island methylation revisited*
Abstract: Data structures frequently encountered in genomics settings are characterized as "large p, small n" reflecting the collection of numerous feature measurements on a small number of samples. Both microarray and sequence-based predictor studies are of this flavour. When a linked binary phenotype (outcome) is also available, two analysis strategies are employed: (1) evaluation of individual feature differences between outcome classes, and (2) development of (ideally interpretable) multi-feature classification rules for discriminating between classes. The impact of p¿n on validation for both sorts of analyses has been widely recognized, with many recent statistical developments in multiple testing and classifier construction and selection respectively. Here we showcase some of these ideas in the context of predicting CpG island methylation based on DNA attributes, including sequence patterns and predicted structure. While the published analyses (Bock et al., 2006, PLoS Genetics) were seemingly stringent with respect to validation, employing Bonferroni corrections and stratified cross-validation, there are subtleties surrounding the manner in which these are applied that warrant further scrutiny.

Speaker: Professor Adam Siepel (Department of Biological Statistics and Computational Biology, Cornell University)
Title: *Searching for novel functional elements in mammalian genomes by comparative genomics*
Abstract: Vast amounts of comparative sequence data have added a new dimension to genomics, making it possible to search genome-wide for sequences that show the signatures of natural selection and hence are likely to be functional. Certain classes of functional elements, such as conserved protein-coding genes, can be identified quite accurately from comparative sequence data; others, such as conserved RNA genes and regulatory elements, and lineage-specific functional elements of various kinds, are harder to find, but comparative prediction methods are rapidly improving. I will present recent work on the identification

of functional elements in mammalian genomes, including both new methods and new biological findings obtained using these methods. In particular, I will discuss the identification of conserved protein-coding genes and RNA secondary structures, and of sequences showing lineage-specific conservation or acceleration. In some cases, our methods have identified novel functional elements that have subsequently been confirmed experimentally. In other cases, they have helped to shed light on the amount and nature of functional DNA in mammals, especially in noncoding regions.

Speaker: Professor Terry Speed (Department of Statistics, University of California, Berkeley)
Title: *A method for detection of alternative splicing from exon array data*
Abstract: Analyses of EST data show that alternative splicing is much more widespread than once thought. The advent of exon and tiling microarrays means that researchers now have the capacity to experimentally measure alternative splicing on a genome. New methods are needed to analyse the data from these arrays. We present a method, FIRMA (Finding Isoforms using Robust Multichip Analysis), for detecting alternative splicing in exon array data. FIRMA has been developed for Affymetrix exon arrays, but could in principle be extended to other exon arrays, or to tiling array data. We have evaluated the method using simulated data, and have also applied it to a real data set consisting of a panel of 16 human tissues. FIRMA is able to detect muscle specific exons in several genes confirmed by reverse transcriptase PCR.

Speaker: Professor Duncan Temple Lang (Department of Statistics, University of California, Davis)
Title: *Integrating software for data analysis*
Abstract: As different communities and sub-fields of different ”omic” research thrive, the number of different software projects grows. It is obviously important to be able to rapidly experiment with, integrate and extend software to do new things efficiently. To this end, we will briefly talk about four different aspects of this integration problem that might be relevant to inter-disciplinary analysis, allowing statisticians to export functionality to other scientists and similarly import computational tools from others.

Firstly, we'll describe how we can have the computer create an R package to interface to arbitrary C/C++ by analyzing the declarations using R. We use this to generate an R interface - RwxWidgets - to the wxWidgets GUI toolkit which itself provides facilities for rapidly creating high quality, rich interfaces to statistical functionality and analysis.

We'll also describe a bridge between R and Perl that allows Perl programmers to transparently use R functions, and similarly allows R programmers to access Perl classes and methods. And finally, I'll mention the TypeInfo package we have developed with Robert Gentleman's group to facilitate making R functions available as general servers, e.g. via Web Services/SOAP or DCOM.

Speaker: Professor Fitnat Yildiz (Department of Environmental Toxicology, University of California, Santa Cruz)
Title: *Molecular basis and consequences of phenotypic variation in Vibrio cholerae*
Abstract: V. cholerae, the causative agent of the disease cholera, is a natural inhabitant of aquatic environments. The pathogen causes periodic, seasonal cholera outbreaks in regions where the disease is endemic and can spread worldwide in pandemics. The ability of V. cholerae to cause epidemics is linked to its survival in aquatic habitats. The capacity of V. cholerae to undergo a phenotypic variation event, that results in the generation of two morphologically different colony variants termed smooth and rugose, is predicted to be important for the survival of the pathogen in natural aquatic habitats. With the availability of the completely sequenced genome of V. cholerae, we employed a genome-wide approach to understand the molecular basis and molecular consequences of smooth-to-rugose phenotypic variation. To this end, we performed whole genome expression profiling studies of smooth and rugose phase variants and of the regulatory mutants. Analysis of the expression data revealed that ”rugosity” and ”smoothness” is modulated by second messenger cyclic di-guanylic acid (c-diGMP) and controlled by a complex hierarchy of positive and negative transcriptional regulators. In this talk, I will describe genetic and genomic characterization of networks controlling phenotypic variation in V. cholerae.

Speaker: Professor Hongyu Zhao (Department of Public Health, Yale University)
Title: *Bayesian error analysis model for reconstructing transcriptional regulator networks*
Abstract: 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. Despite these efforts, transcription regulation is yet not well understood because of its complexity and limitations in biological experiments. 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 lecture, we propose a Bayesian error analysis model to integrate protein-DNA binding data and gene expression data to reconstruct transcriptional regulatory networks. There are two unique aspects to this proposed model. First, transcription is modeled as a set of biochemical reactions, and a linear system model with clear biological interpretation is developed. Second, measurement errors in both protein-DNA binding data and gene expression data are explicitly considered in a Bayesian hierarchical model framework. Model parameters are inferred through Markov chain Monte Carlo. The usefulness of this approach is demonstrated through its application to infer transcriptional regulatory networks in the yeast cell cycle.

Joint work with Ning Sun and Ray Carroll.

Banff International Research Station

for Mathematical Innovation and Discovery

Computational and Statistical Genomics

July 08–13, 2006

ABSTRACTS: CONTRIBUTED POSTER SESSION

`www.stat.berkeley.edu/~biolab/BIRS06/posters.html`

Presenter: Kasper D. Hansen (Division of Biostatistics, UC Berkeley)
Title: *Genome-wide identification of alternative isoforms targeted by nonsense-mediated mRNA decay in Drosophila melanogaster*
Abstract: In order to investigate the nonsense mediated decay (NMD) mechanism in fly, we have constructed a microarray targeting all known alternative isoforms from FLYBASE. A novel deconvolution algorithm for estimating how the individual isoforms change under treatment is proposed and applied to UPF-1 knockdown data.

Joint work with Qi Meng, Marco Blanchette, Richard E Green, Fabian L. Gallusser, Jan Rehwinkel, Elisa Izaurralde, Sandrine Dudoit, Donald C. Rio, Stephen E. Brenner.

Presenter: Professor Ludwig A. Hothorn (University of Hannover, Germany)
Title: *DNA motif identification based on order restricted model selection*
Abstract: A DNA sequence of length k can be described by a 4 by k position specific weight matrix, where the frequencies of the bases A, C, G and T along the columns are counted. Using the maximum of the column values, a 2 by k contingency table results. Recently, van Zwet et al., (2005) identified DNA motifs by a certain order restriction for the entropy estimated from the 2 by k table. We formulate a motif by a specific order restricted hypothesis, so-called epidemic alternative. Therefore a motif is where a change point from the conservative part occurs downwards followed by a change point upwards after some positions. For this specific order restriction we use a model selection approach where for the null model (no motif) and all elementary alternative models (all possible motifs within a sequence of k) the model selection criteria: log-likelihood penalty term is estimated. This method is a modification of Anraku's (1999) model selection approach under total order restriction for Gaussian variables. An improved characteristics for motif identification was found using Gaussian change-point penalty term according to Ninomiya (2005). In a simulation study the correct model selection rates for a true null model, a true symmetric and true asymmetric motif for different sequence length, sample sizes, proportions of the conservative region and motif pattern are demonstrated. Based on a real data example, a R program is demonstrated. Even on a PC is the computation time short so that an implementation of this algorithm in more complex problems seems to be possible. The major advantage of this new approach is the outcome: the existence of a motif or not (global decision) and if a motif exists, the identification of a specific motif (local decision).

Joint work with X. Mi.

References

Anraku K. An information criterion for parameters under a simple order restriction. Biometrika 1999; 86:141-152.

Ninomiya Y. Information criterion for Gaussian change-point model, Statistics & Prob. Letters 2005; 72;

237-247.
vanZwet E, Kechris KJ Bickel PJ et al. Estimating motifs under order restrictions. Statistical Applications in Genetics and Molecular Biology 2005, 4: 1,1.

Presenter: Professor Katherine S. Pollard (Department of Statistics and Genome Center, University of California, Davis)
Title: *Genome-wide scan for the most evolutionarily accelerated regions in the human genome reveals a novel RNA sturctural gene expressed in early neocortical development*
Abstract: The developmental and evolutionary mechanisms of behind the emergence of human-specific brain features remain largely unknown. However the recent ability to compare our genome to our closest relative, the chimpanzee, provides new avenues to link genetic and phenotypic changes in the evolution of the human brain. We devised a ranking of of all regions in the human genome of at least 100bp that exhibit significant evolutionary acceleration, defined as being highly conserved in mammalian evolution up to the common ancestor of human and chimp, but extensively changed by substitutions in the human lineage since that ancestor. The most dramatic such human accelerated region, HAR1, was found in a novel RNA gene that is strongly expressed specifically in the developing human neocortex from 7 to 19 gestation weeks, a critical period for neuronal specification and migration of cortical neurons. Within the neocortex, HAR1 is strikingly co-expressed with reelin in Cajal-Retzius neurons in the subpial granular layer, a transient population of cells that is particularly prominent in humans and higher primates. Cajal-Retzius are of fundamental importance in specifying the six-layer structure of the human cortex. HAR1 and the other human accelerated regions provide exciting new candidates in the search for uniquely human biology.

Computational and Statistical Genomics

July 08–13, 2006

ABSTRACTS: CONTRIBUTED SOFTWARE DEMO SESSION

`www.stat.berkeley.edu/~biolab/BIRS06/softwareDemos.html`

Presenter: Professor Steve Horvath (Departments of Human Genetics and Biostatistics, University of California, Los Angeles)
Title: *Weighted gene co-expression network analysis using the R software*
Abstract: How to use custom-made R functions for weighted network analysis. Relevant R code and sample data sets can be found at the following webpage: `www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork`. Specifically, topics include the following: 1) network construction with the scale free topology criterion, 2) gene module detection with the topological overlap matrix, 3) definition of a gene significance variable based on a clinical trait, 4) module enrichment analysis, 5) understanding the relationship between intramodular connectivity and gene significance, 6) network based gene selection strategy. We will demonstrate how to screen for biomarkers in a brain cancer microarray data set.

Presenter: Professor John Quackenbush (Dana-Farber Cancer Institute and the Harvard School of Public Health)
Title: *MeV and AMP: Tools for analysis of gene expression data*
Abstract: DNA microarray analysis has become a standard technique in many laboratories. As the cost has fallen while reliability has increased, the challenge for most laboratories is no longer in generating the data but rather in analyzing it. While advanced statistical tools exist in the form through efforts such as Bioconductor, to most bench biologists these are unwieldy and challenging to use. To address these issues our group has been developing a suite of software tools aimed at delivering advanced data mining and data analysis tools to the bench biologist in an intuitive and easy to use format. In this presentation I will highlight two of these tools - the Automated Microarrray Pipeline (AMP), which provides normalization and preliminary analysis for gene expression data collected on Affymetrix Gene Chips, and MeV, an open-source software tool for data analysis and data mining.