

Bio³ Seminar Series

Cell Lines, Microarrays, Drugs and Disease: Trying to Predict Response to Chemotherapy

Keith Baggerly, Ph.D

Associate Professor

Department of Bioinformatics and Computational Biology

M. D. Anderson Cancer Center, Texas



Abstract:

Over the past few years, microarray experiments have supplied much information about the dysregulation of biological pathways associated with various types of cancer. Many studies focus on identifying subgroups of patients with particularly aggressive forms of disease, so that we know who to treat. A corresponding question is how to treat them. Given the treatment options available today, this means trying to predict which chemotherapeutic regimens will be most effective.

We can try to predict response to chemo with microarrays by defining signatures of drug sensitivity. In establishing such signatures, we would really like to use samples from cell lines, as these can be (a) grown in abundance, (b) tested with the agents under controlled conditions, and (c) assayed without poisoning patients. Recent studies have suggested how this approach might work using a widely-used panel of cell lines, the NCI60, to assemble the response signatures for several drugs. Unfortunately, ambiguities associated with analyzing the data have made these results difficult to reproduce.

In this talk, we will describe how we have analyzed the data, and the implications of the ambiguities for the clinical findings. We will also describe methods for making such analyses more reproducible, so that progress can be made more steadily.

Friday, September 12, 2008 10:00-11:00 am

Conference Room E501, New Research Building

Refreshments will be provided at 9:45am

Sponsored by the Department of Biostatistics, Bioinformatics, and Biomathematics
This seminar series meets the 2nd and 4th Friday of every month.

Lombardi Comprehensive Cancer Center, Georgetown University Medical Center

Bio³ Seminar Series

Prediction Limits for Poisson Distribution

Valbona Bejleri, Ph.D

Associate Professor
Department of Mathematics
University of District of Columbia

Abstract:

Statistical prediction differs from standard confidence interval estimation. A point of interest in prediction is the estimation of the unknown values of the random variable, corresponding to the outcomes from the future experiment. We derive prediction limits for a Poisson process using both frequentist and Bayesian approaches. An algorithm of how to construct the optimal (smallest) frequentist upper prediction limit for a single future observation is presented. Our work is based on a Poisson model that uses a Poisson-binomial relationship. Bayesian prediction limits are also calculated. The relationship between prediction limits derived using Bayesian approach (with noninformative priors) and limits derived using frequentist approach is discussed. We show that there is no prior distribution which produces a two sided prediction interval which coincides with the frequentist prediction interval at both the upper and lower limit. Conditions under which Bayesian and frequentist limits agree are important in order to inform our choice of method. The area of application includes the prediction of rare events. An example with real life data will be presented.

Friday, September 26, 2008 10:00-11:00 am

Conference Room E501, New Research Building

Refreshments will be provided at 9:45am

Sponsored by the Department of Biostatistics, Bioinformatics, and Biomathematics

This seminar series meets the 2nd and 4th Friday of every month.

Lombardi Comprehensive Cancer Center, Georgetown University Medical Center

Bio³ Seminar Series

Statistical modelling in trans-study projection of genomic signatures

Mike West, Ph.D

Arts & Sciences Professor of Statistical Science
Department of Statistical Science
Duke University

Abstract:

Gene expression profiles of physiological states and clinical outcomes play increasing roles as biomarkers in both experimental and human observational studies. In moving towards clinical applications, key questions arise of how to link and combine such measures across contexts: from laboratory experiments with cultured cells, to animal model experiments, to human outcome studies and clinical trials. I will highlight these considerations and discuss statistical issues and some of the methodology we use to address these questions. I will draw mainly on studies in cancer genomics where in vitro laboratory results involve gene expression signatures of changes in human cells in response to genetic or environmental interventions, and the in vivo contexts is human breast cancer.

Statistical modelling involves large-scale regression and factor analysis for cross-study integration of gene expression signatures. Examples show how these methods can (a) reveal and quantify the greater complexity of patterns of expression underlying in vitro defined signatures when projected and formally analysis in vivo, (b) link in vitro signatures to a number of cancer-relevant biological pathways not initially represented in the experimental context, and (c) refine and improve gene expression signatures as biomarkers of clinical outcomes. Our analyses also rely on the use of formal statistical methods for biological pathway annotation, illustrating an overall strategy for integrative, trans-study analysis of gene expression data sets that has emerged from oncogene pathway projects and studies of micro-environmental responses in cancer.

Friday, October 10, 2008 10:00-11:00 am

Conference Room E501, New Research Building

Refreshments will be provided at 9:45am

Sponsored by the Department of Biostatistics, Bioinformatics, and Biomathematics

This seminar series meets the 2nd and 4th Friday of every month.

Lombardi Comprehensive Cancer Center, Georgetown University Medical Center

Bio³ Seminar Series

H-CLAP: Hierarchical Clustering within a Linear Array with an application in Genetics

Samiran Ghosh, PhD

Assistant Professor, Department of Mathematical Sciences
Indiana University-Purdue University, Indianapolis

Abstract:

In most cases where clustering of data is desirable, the underlying data distribution to be clustered is unconstrained. However clustering of site types in a linear array, as is often desired in studies of linear sequences such as DNA, RNA or proteins, represents a problem where data points are not necessarily exchangeable and are directionally constrained. Each position in the linear array is fixed, and could be either "marked" (i.e., of interest) or "non-marked". Here we describe a method for clustering of those marked positions. Since the cluster-generating process is constrained by locality inside such an array, traditional clustering methods need adjustment to be appropriate.

We develop a hierarchical approach. We adopt a Markov clustering algorithm for clustering, revealing any natural partitioning in the pattern of marked sites. The resulting clustering algorithm is named H-CLAP, Hierarchical Clustering in a Linear Array through recursive Partitioning.

The method employs domain-specific constraints and has several advantages compared to a standard agglomerative hierarchical clustering algorithm. We have tested the efficacy of our method on data sets, including two biological and several simulated.

Friday, October 24, 2008 10:00-11:00 am

Conference Room E501, New Research Building

Refreshments will be provided at 9:45am

Sponsored by the Department of Biostatistics, Bioinformatics, and Biomathematics

This seminar series meets the 2nd and 4th Friday of every month.

Lombardi Comprehensive Cancer Center, Georgetown University Medical Center

Bio³ Seminar Series

High-throughput flow cytometry data analysis: tools and methods in Bioconductor

Florian Hahne, PhD

Computational Biology Program, Division of Public Health Sciences,
Fred Hutchinson Cancer Research Center

Abstract:

Automation technologies developed during the last several years have enabled the use of flow cytometry high content screening (FC-HCS) to generate large, complex datasets in both basic and clinical research applications. A serious bottleneck in the interpretation of existing studies and the application of FC-HCS to even larger, more complex problems is that data management and data analysis methods have not advanced sufficiently far from the methods developed for applications of flow cytometry (FCM) to small-scale, tube-based studies. Some of the consequences of this lag are difficulties in maintaining the integrity and documentation of extremely large datasets, assessing measurement quality, developing validated assays, controlling the accuracy of gating techniques, automating complex gating strategies, and aggregating statistical results across large study sets for further analysis. In this seminar, we present a range of computational tools developed in Bioconductor that enable the automated analysis of large flow cytometry data sets, from the initial quality assessment to the statistical comparison of the individual samples.

Friday, November 14, 2008 10:00-11:00 am

Warwick Evans Conference Room, Building D, Medical Center Campus
Refreshments will be provided at 9:45am

Sponsored by the Department of Biostatistics, Bioinformatics, and Biomathematics
This seminar series meets the 2nd and 4th Friday of every month.

Bio³ Seminar Series

Model Building: Data with Random Location and Random Scale Effects

William S. Cleveland

Shanti S. Gupta Distinguished Professor, Purdue
Joint work with Lei Shu, Abbott Laboratories; Chaunhai Liu, Purdue;
and Lorraine Denby, Avaya Labs



Abstract:

General approaches and tools for model building will be presented for data with random effects, pervasive in medical studies where people are units with repeat measurements. Typically, fitted models have random location effects, but any time location effects are present, there is a high potential for random scale effects to be present; at the very least, it is wise to routinely check for scale effects.

Our stepwise model building approach identifies the error, scale, and location distributions, in that order; each subsequent step uses any previous identifications. Visualization tools are at the core of the identification methods. Also at the core is in-field null-power simulation, which applies to the specific data at hand and its specific finite sample. Null simulations allow us to judge if deviations from expected patterns warrant attention. Power simulations determine our ability to differentiate alternative models. Approaches and methods are illustrated by application to three data sets from customer opinion polling, nutrition, and hospital services.

Friday, December 12, 2008 10:00-11:00 am

Warwick Evans Conference Room, Building D, Medical Center Campus
Refreshments will be provided at 9:45am

Sponsored by the Department of Biostatistics, Bioinformatics, and Biomathematics
This seminar series meets the 2nd and 4th Friday of every month.

Bio³ Seminar Series

Application of Multivariate Procedures for Microarray Data Analysis

Nusrat Jahan, PhD

Department of Mathematics and Statistics
James Madison University

Abstract:

Advances in modern biotechnology have made it possible to study the expression levels of thousands of genes simultaneously. Microarray technology is producing enormous data but information content per gene is small due to lack of replicate samples. Huge volume of information with limited sample size, poses challenging problems for statistical analysis and interpretation. In microarray data analysis, univariate methods are primarily used for detecting differences among gene expression values between two or more conditions, resulting in huge number of leads. But microarray data are essentially multivariate in nature. In order to study the relationship among genes that have different behavioral pattern, multivariate analysis is needed. Multivariate techniques can discern patterns among the differentially expressed genes. These techniques are useful for identifying groups of genes that are behaving in a coordinated pattern and also outlier genes. In our study we look at the case where number of genes is large but sample size is very small. Different multivariate procedures are applied to narrow down the number of leads obtained from various univariate analyses. Finally we identify outlier genes that are able to distinguish between two known different classes of samples.

Friday, January 9, 2009 10:00-11:00 am

Conference Room E501, New Research Building

Refreshments will be provided at 9:45am

Sponsored by the Department of Biostatistics, Bioinformatics, and Biomathematics
This seminar series meets the 2nd and 4th Friday of every month.

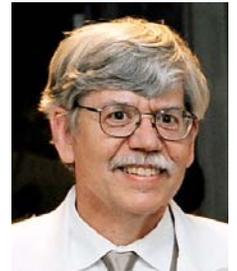
Lombardi Comprehensive Cancer Center, Georgetown University Medical Center

Bio³ Seminar Series

Screening for Alternative Transcript Splicing using Affymetrix Exon Array Data

Peter Munson, PhD

Chief, Mathematical & Statistical Computing Laboratory
National Institutes of Health, Bethesda, Maryland



Abstract:

Exons, introns and splicing of mRNA, discovered in the late 1970s, is a process which adds considerable complexity to the human genome. In addition to other forms of regulation of gene expression (chromatin, transcriptional, degradation, sequestration, etc) alternative splicing allows the gene to produce not just one, but from 2 to 10,000+ distinct protein products by selecting alternative subsets of exons, thus allowing the roughly 30,000 human genes to produce 100,000 or more proteins, as and where needed. The Affymetrix Human Exon 1.0 ST array includes about 1.4 million probesets, more than enough to cover an estimated 1 million human exons, and allows for measurement of gene expression and potentially, discovery of known and novel alternative splicing events.

Significant new statistical and bioinformatic problems are encountered when analyzing data from this microarray. Detecting alternative splicing requires detection of differences in differential expression between exons within a gene, making the analysis very sensitive to the underlying assumptions. Also, while the chip can indeed detect the presence of alternative splicing, it is not always possible to determine the particular mechanism of splicing, or exactly which exons are involved. We present several potential solutions to problems encountered in the analysis of such data. One solution is to re-design the array to include "junction" probes which hybridize only when the splice between two exons is present. Our experience with one such design will be reviewed.

Friday, January 23, 2009 10:00-11:00 am

Conference Room E501, New Research Building

Refreshments will be provided at 9:45am

Sponsored by the Department of Biostatistics, Bioinformatics, and Biomathematics
This seminar series meets the 2nd and 4th Friday of every month.

Bio³ Seminar Series

A Distribution for P-Values

Chang Yu, PhD

Assistant Professor, Director of Biostatistical Research, Generical Clinical Research Center, Vanderbilt University Medical Center

Abstract:

What is the distribution of the p-value under the alternative hypothesis? We describe the properties of a parametric distribution defined on the interval (0,1). This distribution includes the uniform as a special case. The functional form is derived as the distribution of the p-value in a statistical test of a pair of close hypotheses in a wide variety of settings. The distributional form is retained when it is compounded with a uniform or when the individual p-values are sampled from a variety of different hypotheses. We describe properties of the parameter estimate and the distribution of extreme order statistics. The distribution is fitted to data from a study of breast cancer patients comparing many genetic markers. The p-values generated in a microarray experiment comparing gene expressions can be considered a mixture of p-values under the null hypothesis and under a range of alternative hypotheses. The proportion under the null is of interest. Using the derived distributions, we provide a method to estimate this proportion under the framework of mixture models.

Friday, February 13, 2009 10:00-11:00 am

Conference Room E501, New Research Building

Refreshments will be provided at 9:45am

Sponsored by the Department of Biostatistics, Bioinformatics, and Biomathematics
This seminar series meets the 2nd and 4th Friday of every month.

Bio³ Seminar Series

Statistical Input into Science Policy

Mary A. Foulkes, Ph.D.

George Washington University
Department of Epidemiology and Biostatistics



Abstract:

A new emphasis on evidence-based policy presents unprecedented opportunities for statistical input, for statisticians to contribute to new efforts. There are, however, numerous substantive statistical contributions to policy from the past which will be reviewed. Public health issues, for example, in infectious diseases, have raised challenges and questions that statistical modeling, experimental design and novel analyses have addressed. Many new directions in science, such as genomics, and new capabilities, such as high throughput computing, require quantitative approaches often provided by bioinformatics, economics, or other disciplines, but may miss some essential statistical thinking. Science policy is at a tipping point where statistical thinking will become a necessary component; where communication of statistical issues will become an even more essential aspect of the discipline of statistics. The science of science policy as a new initiative of the Federal government, and particularly of NSF, will be reviewed.

Friday, February 27, 2009 10:00-11:00 am

Conference Room E501, New Research Building

Refreshments will be provided at 9:45am

Sponsored by the Department of Biostatistics, Bioinformatics, and Biomathematics

This seminar series meets the 2nd and 4th Friday of every month.

Lombardi Comprehensive Cancer Center, Georgetown University Medical Center

Bio³ Seminar Series

A Bayesian approach to adjust for diagnostic misclassification in Poisson regression

James Stamey, PhD

Associate Professor, Department of Statistical Sciences
Baylor University, Texas

Abstract:

Response misclassification of counted data biases and understates the uncertainty of parameter estimators in Poisson regression models. To correct these problems, classical procedures have been proposed but they rely on asymptotic distribution results and on supplemental validation data in order to estimate unknown misclassification parameters. We derive a new Bayesian Poisson regression procedure that accounts and corrects for misclassification for a count variable. Under the Bayesian paradigm one may use validation data, expert opinion, or a combination of these two approaches to correct for the consequences of misclassification. The Bayesian procedure proposed here yields an operationally effective way to correct and account for misclassification effects in Poisson count regression models. We also investigate a Bayesian variable selection procedure. We demonstrate the performance of the model and variable selection procedure in simulation studies. Additionally, we analyze two real data examples and compare our new Bayesian inference method that adjusts for misclassification to a similar analysis ignoring misclassification.

Friday, March 13, 2009 10:00-11:00 am

Conference Room E501, New Research Building

Refreshments will be provided at 9:45am

Sponsored by the Department of Biostatistics, Bioinformatics, and Biomathematics

This seminar series meets the 2nd and 4th Friday of every month.

Lombardi Comprehensive Cancer Center, Georgetown University Medical Center

Bio³ Seminar Series

Improving the Efficiency of the Logrank Test Using Auxiliary Covariates

Xiaomin Lu, PhD

Assistant Professor, Department of Epidemiology and Biostatistics
University of Florida

Abstract:

The logrank test is widely used in many clinical trials for comparing the survival distribution between two treatments with censored survival data. Under the assumption of proportional hazards, it is optimal for testing the null hypothesis of $H_0: \beta = 0$, where β denotes the logarithm of the hazard ratio. In practice, additional auxiliary covariates are collected together with the survival times and treatment assignment. If the covariates correlate with survival times, making use of their information will increase the efficiency of the logrank test. We apply the theory of semi-parametrics to characterize a class of regular and asymptotic linear estimators for β when auxiliary covariates are incorporated into the model, and derive estimators that are more efficient. The Wald tests induced by these estimators are shown to be more powerful than the logrank test. Simulation studies and a real data from ACTG 175 are used to illustrate the gains in efficiency.

Friday, March 27, 2009 10:00-11:00 am

Conference Room E501, New Research Building

Refreshments will be provided at 9:45am

Sponsored by the Department of Biostatistics, Bioinformatics, and Biomathematics
This seminar series meets the 2nd and 4th Friday of every month.

Bio³ Seminar Series

Penalized composite likelihood estimation for age-dependent branching processes using CFSE-labeling data

Ollivier Hyrien, PhD

Assistant Professor of Biostatistics and Computational Biology,
University of Rochester Medical Center

Abstract:

Recent advances in flow cytometry have resulted in the development of powerful bioassays to analyze the dynamics of cell populations. CFSE-labeling experiment is one such assay that has become a standard tool in the analysis of cell kinetics, especially for investigating the proliferation of lymphocytes. The statistical analysis of such data has received little attention, but it poses a number of methodological issues. In this talk, we will describe a framework that relies on mixture models and age-dependent branching processes to analyze CFSE-labeling data. A penalized composite likelihood approach will be proposed to estimate cell kinetics parameters. Large and finite sample properties of the proposed estimator will be discussed. The talk will be illustrated by an application to the proliferation of activated lymphocytes. This is joint work with Rui Chen and Martin S. Zand from the University of Rochester.

Friday, April 24, 2009 10:00-11:00 am

Conference Room E501, New Research Building

Refreshments will be provided at 9:45am

Sponsored by the Department of Biostatistics, Bioinformatics, and Biomathematics
This seminar series meets the 2nd and 4th Friday of every month.

Lombardi Comprehensive Cancer Center, Georgetown University Medical Center

