Tutorial Title: Statistical Analysis of Microarray Studies

Instructor: Emmanuel Lazaridis

Length: one-half day

Tutor Qualifications:

Dr. Lazaridis is faculty in the Interdisciplinary Oncology Program of the University of South Florida (USF). He co-directs the activities of the Biostatistics Core of the H. Lee Moffitt Cancer Center and Research Institute at USF, where he also serves on the Scientific Review Committee. Dr. Lazaridis directs the USF Pilot Center for Mathematical-modeling of Image Data Across the Sciences (MIDAS), which was established in 1999. The mission of the MIDAS center is to create interdisciplinary collaborations among imaging, quantitative and biological scientists, in order to develop new analytic models of image-related data and to train researchers of various disciplines in modeling techniques. He joined the cancer center from Indiana University School of Medicine (IUSM), where he was affiliated with the Department of Medicine, the Regenstrief Institute for Health Care and the Herman B. Wells Center for Pediatric Research. He is a graduate of the statistics program of the University of Chicago.

Dr. Lazaridis has successfully competed for funding as Principal Investigator in studies of aging, cancer and statistical methods through the NIH, the American Cancer Society, and the Department of Defense. In addition he has developed substantial and productive collaborations with other medical researchers. He directs the research computing section of the Molecular Oncology Program Project (PO1-CA-82533), which seeks to develop more effective therapies for human cancer based on a better mechanistic understanding of tumor cell survival and drug resistance. Dr. Lazaridis participates in the Biomarker Development Laboratory at Moffitt (U01-CA-84973), where he develops technologies for biomarker testing in lung cancer. As co-PI he directs all biometrics research for the studies Molecular Fingerprint of STAT3 Regulated Genes for Early Detection of Human Cancer (DAMD-17-98-1-8659) and similarly as co-investigator for Decoding Fingerprints Portending Colon Cancer Metastasis (UO1-CA85052). Both of these studies employ spotted cDNA and oligonucleotide microarrays.

In past work, Dr. Lazaridis has collaborated with investigators in diabetes research (*Diabetes Care, Journal of General Internal Medicine; Archives of Family Medicine*), pharmacology (*American Journal of the Medical Sciences*), cancer research (*Blood; Bone Marrow Transplantation*) and aging (*Journals of Gerontology: Medical Sciences; Facts and Research in Gerontology*). In addition to his collaborative, analytic and scientific expertise, he has published articles on Bayesian models and applications (*Statistics in Medicine; Communications in Statistics: Stochastic Models; Computer Methods and Programs in Biomedicine*), and regarding innovations in instructional methods and theory (*Academic Medicine; Journal of Statistics Education*). Recent articles include “Linking Image Quantitation and Data Analysis in Modern Biological Experimentation”, “Statistical Contributions to Molecular Biology”, and “Introduction to Microarray Experimentation and Analysis.”

His short course, "Statistical Methods in Molecular Biology: RNA and Protein Analyses", first sponsored by the American Statistical Association at the 2000 Joint Statistical Meetings, received rave reviews from the attendees, with 52% grading the course an overall A, and 34% an overall B. 64% of the statisticians attending the JSM short course rated his knowledge concerning the topics of instruction at the A level, and an additional 28% at a B level. Dr. Lazaridis regularly offers continuing education to molecular biologists, statisticians and other quantitative analysts in areas of microarray, proteomics, flow cytometry and other molecular biology technologies and their associated analytic methods.

Tutorial Presentation:

This tutorial will focus on analysis of RNA expression data derived from microarray images. Both oligonucleotide and spotted arrays will be covered. Examples illustrating specific analytic techniques, as

well as the interrelationship between imaging and statistical methods, will be highlighted. Some examples will be drawn from the literature, while others will come from microarray studies at the H. Lee Moffitt Cancer Center & Research Institute, exploring aspects of leukemia, colon and breast cancer biology. The course will introduce an ontology of analytic methods, briefly cover data mining techniques, and concentrate on methods for hypothesis-driven research. Recent developments in visualization, analytic probe selection, principal component and SVD analyses, Bayesian statistics, and latent class modeling will be presented and discussed. The material of the proposed tutorial continues to develop at a rapid pace. Because of his background, research interests and expertise, the instructor is at the forefront of the development of statistical methods for the analysis of experimental molecular biology data. Important new developments may be incorporated into the tutorial as time permits.

A primary focus of this tutorial is to explain specific analytic methods and to illustrate their application to microarray data. The tutorial follows an ontology of microarray analytic methods developed with support from the American Statistical Association. The course is given in sections, each concentrating on a specific family of analytic methods. My experience is that these sections range from 30 to 50 minutes in length. Typically, each topic is introduced through one or more examples that illustrate the kinds of questions the method can address. The introductory examples are followed by presentation and description of the technique. At this stage, some specialized knowledge of mathematics and computer programming is important if one is to understand the course in its entirety. For example, in the context of principal components and related dimension reduction techniques, a full understanding of the methods requires some familiarity with linear algebra. This said, the instructor does strive to communicate the idea behind the method graphically for attendees without a strong technical background. The technical presentation is followed by descriptive application to the examples. This is followed by general discussion with the audience. The time for each section is allocated as follows: 30% introductory examples, 20% methods, 30% descriptive application, 20% general discussion. Thus, a 40 minute section would include 12 minutes of introductory examples, 8 minutes of formal methods presentation, 12 minutes demonstrating and describing an analysis, and 8 minutes of discussion. Because this tutorial is proposed to follow at one-half day format, the timing of the sections will be closely monitored; however, the instructor recommends that attendees be given the opportunity for continued discussion following the conclusion of the formal program.

The course will be presented from CD-ROM. Slides will be in MS Powerpoint. If possible, an Internet connection should be provided for the instructor's computer, for access to computational resources at the Moffitt Cancer Center. Although no published textbook is appropriate for this course, the instructor has developed a handout that will be revised as appropriate for attendees of this tutorial. Access to data and methods will be given to participants through the MIDAS Center web site after the course has completed. The instructor is an accomplished web programmer who also oversees implementation of biometrics web materials for the cancer center.

Intended Audience:

This tutorial will be geared towards quantitative analysts from varied backgrounds who are already substantially familiar with one or more microarray technologies. It will be example driven, so as to be accessible to the widest possible audience while still retaining rigor in presentation. Participants must have prior familiarity with microarray technologies, since this tutorial will only briefly cover technology and only in the context of data analysis seeking to uncover quality control issues. Familiarity with basic molecular biology concepts will be assumed, but a handout will cover those concepts for participants with weak biology backgrounds. College mathematics and computer science experience is needed to completely understand the implementation of algorithms to be presented, but the course is designed so that attendees without such background will still benefit.

The instructor has designed and conducted professional training activities for non-statistical analysts, biologists and medical doctors at Indiana University and at the University of South Florida. He is supported by the American Statistical Association to conduct training in microarray data analysis to statisticians at the

annual Joint Statistical Meetings (2000, 2001). A primary goal of this tutorial for the instructor is to continue his outreach beyond the statistics community to the larger community of applied data analysts. Thus, participation is sought particularly from computer scientists, molecular biologists, bioinformaticists and other non-statistician analysts, and the tutorial is designed with this audience in mind.

Learning Outcomes:

On completion of this tutorial, attendees will be cognizant with data analytic issues associated with microarray studies, and understand application, benefits and limitations of available methods to analyze data from these studies.

Outline:

These are the available modules from which this tutorial will be constructed. Depending on the exact timing of the course (starting / ending times, scheduled breaks), and feedback from the ISMB 2001 committee, and new developments in the field, sections may be added or omitted, or their presentation order changed.

Section 0: Introduction to this tutorial (10 minutes)

Section 1: Imaging issues in microarray studies (25 minutes)

Section 2: Combining information on oligonucleotide arrays (20 minutes)

Section 3: Quality control in spotted cDNA slides (20 minutes)

Section 4: An ontology of inferential methods (15 minutes)

Section 5: Models geared towards hypothesis testing (30 minutes)

Section 6: Mining data through clustering – options, benefits and limitations (25 minutes)

Section 7: Dimension reduction through linear components models (40 minutes)

Section 8: Latent class models (25 minutes)

Section 9: Issues in experimental design (20 minutes)

Section 10: Quantitation adjustments in microarray data analysis (15 minutes)

Section 11: Linking images to data analysis – rationale and technology (20 minutes)

Section 12: Forthcoming analytic developments for microarray studies (20 minutes)

While this outline indicates the material to be covered and the relative time allocations recommended by the instructor, it does not give a complete description of the course content. Included as an appendix to this proposal is a chapter written by the instructor and his colleagues for the text *Biostatistical Methods in Molecular Biology* (Humana Press, in press). This text gives brief descriptions of many of the methods that will be covered in this course. It will also give the committee a feel for the philosophical convictions of the instructor. These are key elements of the design of this tutorial, which concentrates on hypothesis-driven analytic methods as opposed to database informatics and visualization techniques.