Research Interaction Team Plan

Team Name: Statistical Analysis of DNA Microarrays

Team Director(s): Eric Slud, possibly along with other faculty

Research Focus: The rapidly growing field of Bioinformatics is a rich source of statistical data on 'gene expression levels': measurements obtained simultaneously on large numbers (often many thousands) of genes or DNA fragments for each of a number (usually a few dozen or less at a time) of cell-lines, tissue-sources, or individuals. The datsets are large, highly cross-classified (since much additional information about the genes and/or individual mnay be available), and extremely hard to model because the sample sizes are small (dozens) while the dimensionality is high (thousands) and the dependence relationships (essentially, information regarding the co-firing of different genes) are often badly understood at the biological level. *Another key feature of this field is that there is a great deal of publicly available data on the Web.* Techniques and statistical fields relevant to the analysis and modelling of such data include: pattern recognition and classification, clustering, (generalized) linear models and random-effects extensions, principal components methods, nonparametric regression, spatial data analysis, data-resampling and bootstrapping, and the list goes on. At present, the efforts to understand such data are largely descriptive and atheoretical. The objective of this Research Team is to investigate various modelling ideas with respect to statistical goodness of fit and stability of gene- and cell-line clusters generated. There is room here for work combining specifications and modelling ideas from various approaches which have been tried previously; for computational data analyses which will lead to purely empirical model-choices; for exploration of new probabilistic dependence ideas which can be used to attack the biological idea of co-firing genes along coherent biochemical pathways; and for algorithms defining clusters using fitted models.

Graduate Prerequisites: A graduate student should have completed Stat 700 and should preferably be familiar with Statistical Computing at least at the level of Stat 430 (SAS programming) or Stat 798C (Splus and SAS) or AMSC 660 (miscellaneous software including MATLAB).

Undergraduate Prerequisites: An interested undergraduate should have had at least one course in Mathematical Statistics (e.g. Stat 401 or 420) and some experience — either in courses or projects — with numerical computing or data analysis.

Graduate Program: Graduate students will be involved in formulating the research questions, in searching out available literature references and publicly available software, and in implementing model-fits and goodness-of-fit evaluations.

Undergraduate Program: Undergraduate students will be involved primarily in experimenting with ideas for generating and comparing the validity and stability of clusters from fitted models on public data.

Meeting Schedule: We will run an intensive seminar in the spring of 2002. Initially this will involve informally summarizing the main ideas and approach of papers, a weekly activity continuing the AMSC 699 seminar run on the same topic in Fall '01, with reading list on the Web at **http://www.math.umd.edu/∼evs/genom/genom.pdf**) and trying to implement as many of them as possible on a few key publicly available datasets (posted at various web-sites by previous biomedical investigators). Some of our meetings will take plan at workstations for illustrative interactive sessions, e.g. in Splus data analyses. As we get some experience, we will formulate our own models and clustering ideas for comparison.

In addition, or as a temporary replacement for some these meetings in the Spring, we may attend and discuss the short-course/seminar on DNA sequence models and algorithms which will be given in the Spring '02 Tuesday afternoon Stat Workshop by Dr. Stephen Altschul of NIH.