Bioinformatics II, BIO597F/CSE598F/STAT597F (Spring 2002)

Francesca Chiaromonte, Statistics, chiaro@stat.psu.edu, Thomas 411, 5-7075.
Webb Miller, Computer Science and Engineering, webb@cse.psu.edu, Pond 326A, 5-4551.

Tue/Thur 2:30-3:45; 117 Thomas; *CSE schedule number:* 976593; 3 credits

This course is dedicated primarily to computational and statistical methods for exploring, synthesizing and understanding global gene expression data (e.g. from microarrays). The first part of the course deals with expression data alone, while the second considers techniques to merge expression information with data from other sources.

The course begins with a short introduction to microarray experiments, including the generation of global gene expression data on two or multiple experimental conditions, multiple experimental units, time courses. We will also touch upon data preprocessing (sources of error; normalization procedures to make readings from different chips comparable; centering and standardization; missing values) and devote some attention to gene filtering (what genes present significant changes in expression across experimental conditions or units, or along a time course). We will then concentrate on three main topics, each involving a specific class of multivariate statistical methods:

1. Identifying characteristic patterns that explain expression variation; Principal Components Analysis (Singular Value Decomposition) and the basics of dimension reduction techniques.

2. Parsing genes and/or experimental conditions or units, on the basis of expression profile similarity; hierarchical, partition, and mixture-based clustering algorithms, and some hints at algorithms for the simultaneous parsing of genes and conditions (units).

3. Investigating responses on experimental units (e.g. a classification into known groups, or a quantitative trait), and the role of genes' expression in predicting them; Regression modeling with under-resolution, Discriminant Analysis, dimension reduction techniques for regression, and some hints at other supervised classification algorithms.

The second part of the course will focus on methods that combine gene expression data with other types of biological information, such as expressed sequence tags, genomic DNA sequences, quantitative proteomics, and databases of known physical interactions. Techniques to be discussed range from scoring tandem mass spectra against a peptide database to attempting automatic reconstruction of gene networks.

The course has no pre-requisites, but some computational skills and/or familiarity with basic concepts in statistics and sequence analysis (e.g. Bioinformatics I, BIO/CSE597F) will help. Undergraduates must obtain consent of the instructors to register for the the course.

There will be no text book; lectures will combine methodological background description and presentation of analyses and results from recent articles. We will provide a list of reference books, distribute articles, and post class notes on the website.

Students will be divided in small groups that will work together on approximately four homework assignments and a final project. Homework assignments will include literature review, as well as computing and data analysis, and will be handed in as short reports produced by each group. In the final project, groups will be asked to work on a data set in an open-ended fashion, designing and performing an analysis (i.e. selecting questions, methods to address them, and appropriate literature references). Analyses by each group will then be presented to the class. In addition to homework and final project, there will be an in-class test on basic concepts towards the end of the semester. The grade will be based 30% on homework, 30% on the in-class test, and 40% on the final project.