Bioinformatics II, Spring 2004

Main Instructor: **Francesca Chiaromonte**, Statistics, chiaro@stat.psu.edu,Wartik 505, ph 5-7075.
Also: **Webb Miller**, Computer Science and Eng., webb@bio.cse.psu.edu, Wartik 501, ph 5-4551.
(Plus other guest lecturers).

Course web-site: http://www.stat.psu.edu/~chiaro/BioinfoII_04 (will be active later this week)

This course is dedicated primarily to computational and statistical methods for exploring, synthesizing and understanding global gene expression data (e.g. from microarrays). The main part of the course deals with expression data alone, and will cover material from various modules. These include:

1. Introduction to microarray experiments: spotted and Affymetrix arrays; generation of global gene expression data on two or multiple experimental conditions, multiple experimental units, time courses.
2. Data preprocessing: sources of error; normalization procedures to make readings from different chips comparable; centering and standardization; missing values; filtering. In the context of normalization, we may consider some ANOVA-type scheme and special experimental designs.
3. Identifying differentially expressed genes.

Next, we will concentrate on three main topics, each involving a class of multivariate statistical methods:

4. Identifying characteristic patterns that explain expression variation; Principal Components Analysis (Singular Value Decomposition) and the basics of dimension reduction techniques.
5. Parsing genes and/or experimental conditions or units, on the basis of expression profile similarity: hierarchical, partition, and mixture-based clustering algorithms.
6. Investigating responses on experimental units (e.g. a classification into known groups, or a quantitative trait), and the role of gene expression in predicting them: regression modeling with under-resolution, Discriminant Analysis, dimension reduction techniques for regression, and some hints at other supervised classification algorithms.

If time permits, the final part of the course will be devoted to a selection of topics from the following:

A. Techniques to combine gene expression data with other types of biological information, such as expressed sequence tags, genomic DNA sequences and alignments, quantitative proteomics, and databases of known physical interactions.
B. Techniques for the investigation 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) will help. Undergraduates must obtain consent of the instructors to register for 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 and use 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 select a data set, and work on it 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.

All Penn State and Eberly College of Science policies regarding academic integrity apply to this course. For details, see http://www.science.psu.edu/academic/Integrity/index.html