STAT597A/CSE598F/BIOL597A: Bioinformatics II, Spring 2006

Meetings:	Tue-Thur 2.30-3.45pm, 365 Willard (location subject to change)
Instructors:	Francesca Chiaromonte, Statistics, chiaro@stat.psu.edu, 505 Wartik, ph 5-7075. Naomi Altman**,** Statistics, naomi@stat.psu.edu, 312 Thomas, ph 5-3791.
Web-site:	http://www.stat.psu.edu/~chiaro/BioinfoII_06 (will be active by Jan. 13)

The course is dedicated to statistical and computational methods for the design and analysis of global gene expression studies (e.g. from microarrays), and will cover the following topics:

Introduction, data preprocessing, experimental design, differential expression (Altman ~5 weeks)

1. Introduction to gene expression including basic biology concepts and terminology.
2. Introduction to statistical analysis including basic concepts, terminology and graphical tools. Introduction to the "R" programming environment and Bioconductor for gene expression analysis.
3. Affymetrix and “Spotted” arrays: array design and data preprocessing (normalization).
4. Designing a microarray experiment – from platform to data.
5. Identifying differentially expressed genes.

Multivariate analysis tools applied to global gene expression data (Chiaromonte ~8-9 weeks)

6. Identifying fundamental variation patterns in global expression data: Principal Components Analysis (Singular Value Decomposition) and the basics of dimension reduction techniques.
7. Clustering genes and arrays: Parsing genes and/or experimental conditions or units based on expression profile similarity. Hierarchical, partitioning, and mixture-based algorithms; heatmaps.
8. Investigating categorical and quantitative responses on experimental units, and the role of gene expression in predicting them: supervised dimension reduction, discriminant analysis, regression analysis with under-resolution, hints at other supervised classification algorithms.

Selected topics on multisource analyses and networks (Chiaromonte ~1-2 weeks time permitting)

9. Techniques to combine global expression data with other types of biological information, such as interspecies conservation, annotation of regulatory elements, functional annotation of genes, databases of protein interactions.
10. 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 bioinformatics (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 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 5 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 be presented in 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