G8325: Topics in Advanced Statistics
Modern Statistical Learning and Computing
with Applications to Biology
Department of Statistics, Columbia University
Fall 2005
Tentative as August 18, 2005

Basic Course Information

• **Instructor:** Tian Zheng

• **Office hours:** Mondays 1-3PM, Room 1007 SSW

• **Email:** tzheng@stat.columbia.edu

• **Lecture Time:** TBA;
  First meeting will be **3-5PM Wednesday, September 7th, 2005**

• **Room:** Room 1025 (the conference room), SSW

• **Course website (access limited):**
  http://www.stat.columbia.edu/~tzheng/teaching/Topic_fall05/

Description

In this course we will explore some current statistical learning and computing methodologies with case studies from current computational biology literature. Through these case studies, we will also have a brief survey of the major research areas in statistical applications to biology. Specific topics on computing and simulations are to be covered along the way.

Developments in biotechnology over the past two decades provide biologists with new tools to study the biological activities of various organisms (including human beings) with more details, on a larger scale, and in real time. Data generated from such experiments are usually of unprecedented scales and complexities, the analysis task of which offer new challenges to the field of statistics. New methods have been developed to address these difficult tasks, which is also facilitated by recent advancements in computing powers.

The course will start with several introductory lectures on biology, introductory statistical computing, and important general concepts in statistical learning. Each week we will start with a statistical methodology topic and followed by discussion on a current paper from the computational biology literature. Topics to be covered include but not limited to: EM algorithms, Monte Carlo optimization, bootstrap methods, clustering and Classification, tree-based learning methods, etc. Biological topics to explore include but not limited to: gene expression analysis, haplotype inference, phylogeny analysis, sequence analysis, etc.

Textbook

No textbook required. We will use lectures notes and relevant journal articles, which will be posted on the class website.

Reference texts

More readings on most topics covered in this course can be found in the following references.


For readings on basic biological concepts, go to the Bookshelf from the National Center for Biotechnology Information (NCBI). You can find more than you can read.


Assignments and Grading

Each officially registered student is required to select a paper from the optional reading paper list on the course website, conduct a detailed “evaluation” of the statistical methods used in the selected paper and give a 25-min presentation on the “evaluation” results. Ways of “evaluation” may include a repetition of (or parts of) the analysis in the selected paper (upon availability of data), a simulation-based evaluation of the performance of the methods used in the paper, or comparison of the methods used with some relevant methods through simulation studies. Students are also encouraged to discuss possible improvements of the methods, extension of the same strategy to other research area (not restricted to biology), and propose new methods and strategies for the problem addressed in the given paper.

Grading is based on this paper evaluation project and in-class participation. Students can choose any statistical computational package. Lecture examples will be given using R.