

Topics in Statistical Genetics and Bioinformatics Stat 597D, Spring 2001

Francesca Chiaromonte (chiaro@stat.psu.edu, 5-7075, Thomas 411)

David Hunter (dhunter@stat.psu.edu, 3-0979, Thomas 310)

Tue, Thur 2.30–3.45pm 223 Thomas

This course will present modeling approaches and computational statistical methods, focusing on their application to problems in genetics and bioinformatics.

We will provide some theoretical background on topics such as maximum likelihood estimation, Markov chains, and multivariate analysis, with a particular focus on techniques for the synthesis of high-dimensional, large data sets. Computational statistical techniques we will cover include the EM algorithm, Newton's method, and scoring for maximum likelihood estimation and some Markov Chain Monte Carlo methods, including Gibbs sampling. Among the genetics applications we plan to cover are:

- Gene counting algorithms. Computational methods, including EM and scoring, for estimation of allele frequencies.
- Hypothesis testing. The transmission/disequilibrium test and permutation tests for linkage disequilibrium.
- Pedigree Analysis. Calculation of likelihoods for pedigree data, applications to problems such as paternity testing and risk prediction, and mapping of Mendelian and complex traits.
- Evolutionary trees. Applications of maximum likelihood and continuous-time Markov chains to reconstruction of evolutionary trees.

On the bioinformatics front, we will concentrate on the analysis of global gene expression data from microarrays, covering issues related to:

- Sources of error, preprocessing and normalization of expression data.
- Reduction: identification of relevant projections of the data through classical spectral and singular value decompositions, modern exploratory tools (e.g. projection pursuit, "guided tours"), and exhaustive reduction schemes.
- Classification and Clustering: identification of relevant partitions of genes and/or conditions through hierarchical and non-hierarchical clustering techniques, and maximum likelihood fit of mixture models.
- Working with a Response: reduction and classification when a response variable is observed along with expression; sufficient dimension reduction and variable selection methods for regression.

Time permitting, we will also discuss some literature on how to use and combine expression, sequence and biochemical information to reconstruct gene networks.

Although emphasis will be placed on applications, our main aim is to present a methodological toolkit to graduate students in statistics, computer science/engineering, and the life sciences. Some familiarity with probability and statistical concepts (at the senior undergraduate level) and some experience with computation (low-level languages and statistical packages) will be important.

We will use a few statistical texts as references, explore a selection of recent statistics, bioinformatics and genetics literature, and host guest lectures and discussion session. If the make-up of the class allows for it, we will encourage students from different fields to form work-teams and perform analyses of published or new data.

Grading will be based on approximately biweekly assignments and a final project involving literature review and data analysis (using computation).