PubH 5470-2 Special Topics

Spring 2002, 3 credits, A/F or S/N

Title: Statistics in Genetics and Molecular Biology

Instructors: Drs. Cavan Reilly, Hegang Chen and Wei Pan. `Email:` {`cavanr, hegangc, weip`}`@biostat.umn.edu`.

Catalog Description: It has been well recognized that statistics is going to play an increasingly important role in analyzing large amounts of data resulting from the exciting development in genetics and genomics. This course will introduce statistical applications in genetic mapping, DNA or protein sequence alignment, and analyses of gene expression data and design of microarray experiments.

Lengthy Description: With the exciting developments in genetics and genomics, the challenge now is how to extract useful information from the resulting large amounts of data. It has been recognized that statistics is going to play an increasingly important role. This course will introduce statistical applications in three important areas, genetic mapping, DNA or protein sequence alignment, and analyses of gene expression data from microarray experiments.

1. Genetic mapping: Initially we will discuss some statistical aspects of radiation hybrid maps as an introduction to physical mapping. Then we will provide a statistical treatment of linkage analysis (genetic mapping), and compare/contrast the two sorts of maps. The topics we plan on covering in linkage analysis include: marker quality, lod scores, algorithms for computing genetic likelihoods (e.g. Elston-Stewart), sample size considerations, penetrance, admixture, multi-point linkage, locus ordering, effects of violations of assumptions in genetic models, affected sib-pair methods, linkage disequilibrium and models/methods for complex traits.

2. Sequence alignment: We will discuss statistical and computational methods for sequence alignment and its relationship to database searching, identification of motifs and other patterns using either individual or aligned sets of sequences, and gene discovery. In addition to the techniques such as dynamic programming algorithms, the emphasis will be placed on statistical methods for aiding alignment. The topics include: expectation-maximization (EM) algorithm, multiple EM for motif elicitation, hidden Markov models, motif-based hidden Markov models, and the Gibbs sampler. Furthermore computer programs for database searches (e.g. BLAST) will also be discussed.

3. Microarray data analysis: After introducing microarray technologies, we will discuss various statistical approaches in detecting differentially expressed genes under two conditions. The data may be drawn from microarray experiments with or without replicates of spots or arrays. This includes simple techniques, such as two-sample t-tests and nonpara-

metric rank tests, and some more sophisticated methods like SAM. We will then introduce several multivariate techniques, including various clustering (i.e. unsupervised learning) and classification (i.e. supervised learning) techniques, and their applications to microarray data. If time permits, additional topics to be covered include sample size/power calculations and more recent applications of gene expression data.

Some biological and statistical background will be reviewed when needed.

For whom intended: This course is designed for second-year biostatistics/statistics graduate students who want to learn genetics/genomics, and biology graduate students who have statistical background and want to learn statistical applications.

Prerequisites: Statistics at the level of PubH 5450–5452 or equivalent or permission of instructor. Some background with molecular biology is desirable.

Objective: At the end of the course, the student should have an appreciation of the basic problems facing genomic scientists, a basic understanding of statistical and computational methodologies in genetics and genomics, and a working biological and statistical knowledge in the three areas covered.

Textbooks: There is no required textbook. There will be some readings drawn from journal articles. Course notes will be distributed in class. The following are recommended references:

1. Vogel: to-be-filled
2. Ott, J. *Analysis of Human Genetic Linkage.* The John Hopkins University Press, 1999.
3. Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. *Biological Sequence Analysis.* Cambridge University Press, 1998.
4. Pevzner, PA. *Computational Molecular Biology, An Algorithmic Approach.* MIT Press, 2000.

Evaluations: Course evaluation will be based on homework assignments and projects. There may be no exam.