PHD SEMINAR ANNOUNCEMENT
Department of Statistics

"Statistical Inference for Multi-color Optical Mapping Data"

THURSDAY, July 22, 2004, 3:00 pm
Eckhart Hall, Room 110, 5734 S. University Avenue

Liping Tong
Department of Statistics, University of Chicago

ABSTRACT

We consider a study design in which the data consist of noisy observations of multiple copies of a DNA molecule of interest. Each copy is independently marked with colors at recognition sites (some specific short sequences of nucleotides, such as CCCTTTT, along a DNA molecule). Different colors make distinctions of different types of recognition sites. The relative positions of the observed colors are measured on each copy of the molecule.

Our main goal is to construct a physical map (a sequence marked with recognition sites and their relative positions) from the observations of multiple copies of the DNA molecule of interest. In addition, we would also like to assess the uncertainty in the estimated map and estimate error rates, which is useful for analyzing and refining the biochemical steps in the mapping procedure. This problem arises in the development of a new optical mapping method by Professor Laurens Mets and colleagues at the University of Chicago.

First, we propose statistical models for different sources of errors and use maximum likelihood estimation to construct a physical map and estimate error rates. To overcome the difficulties arising in the maximizing process, a hidden Markov Chain is proposed, and the EM algorithm is used for maximization. In addition, a simulated annealing procedure is applied to maximize the profile likelihood over the discrete space of sequences of colors. A limitation of maximum likelihood estimation is that it does not provide an easily interpretable assessment of uncertainty in the discrete parameter space of sequence of colors. To overcome this difficulty, we place prior probabilities on the parameters, and use Markov Chain Monte Carlo to perform Bayesian inference, which allows comparison of posterior probabilities of parameters of interest.

Finally, we apply the methods to the bacteriophage lambda genome, which contains about 50k nucleotide pairs along the linear double-stranded DNA molecule.