The University of Chicago

Department of Statistics

Seminar

Chiara Sabatti
Department of Human Genetics, University of California at Los Angeles

"Genomewide Motif Recognition with a Dictionary Model"

****

Monday, April 8, 2002 at 4:00 PM
133 Eckhart Hall, 5734 S. University Avenue

ABSTRACT

Authors: AUTHORS: Chiara Sabatti and Kenneth Lange

Bussemaker et al. (2000, PNAS) proposed the simple idea of modeling DNA non coding sequence as a concatenation of words and gave an algorithm to reconstruct deterministic words from an observed sequence. Moving from the same premises, we consider words that can be spelled in a variety of forms (hence accounting for varying degrees of conservation of the same motif across genome locations). The overall frequency of occurrence of each word in the sequence and the parameters describing the random spelling of words are estimated in a maximum-likelihood framework using an E-M gradient algorithm. Once these parameters are estimated, it is possible to evaluate the probability with which each motif occurs at a given location in the sequence. These conditional probabilities can be used to monitor properties of genome sequences, such as neighboring occurrences of given motifs.
