Digital Signal Processing and Communications Laboratory
Electrical and Computer Engineering Department

The challenge of knowledge discovery in Genomic Sciences

Luis Arturo Medrano-Soto

UCLA-DOE Institute for Genomics and Proteomics
Los Angeles, California

Friday November 18, 2005
3:00pm CRBL 305

The current exponential growth of biological databases has exceeded by far the pace at which we can interpret and extract useful knowledge from such a deluge of data. High throughput technologies now allow the generation of gigabytes of information in single experiments. In this talk I will present several concepts and examples to illustrate this point and the corresponding challenge posited to computational genomics.

Standard clustering methodologies are commonly used in bioinformatics to study biological phenomena that involve mathematically homogeneous data (e.g. gene expression patterns). In this case, the "distance" among entities (genes) is easily interpretable. On the other hand, if we need to classify genes by the simultaneous analysis of heterogeneous attributes (e.g. expression levels, chromosomal proximity, molecular function, phylogenetic distance, mode of regulation, etc.), an alternative approach is required because the concept of distance has no useful interpretation under these circumstances.

Here I will introduce an alternative Bayesian method based on mixture models (referred to as BClass), which allows an integrated analysis of heterogeneous biological data. Various statistical distributions are used to model the continuous/categorical data commonly produced in biological experiments. We calculate the posterior probability of each entry (e.g. gene) to belong to each element (group) in the mixture. In this way, an original set of heterogeneous variables is transformed into a set of purely homogeneous characteristics represented by the probabilities of each entry to belong to the groups. The number of groups in the analysis is controlled dynamically by rendering the groups as 'alive' and 'dormant' depending upon the number of entities classified within them. Using standard Metropolis-Hastings and Gibbs sampling algorithms, we constructed a sampler to approximate grouping probabilities. Since this method eliminates the requirement of defining similarity measures, it is especially suitable for data mining and knowledge discovery in biological databases.

Reference:
Medrano-Soto, A. Christen JA and Collado-Vides, J. (2005). BClass: a Bayesian Approach Based On Mixture Models for Clustering and Classification of Heterogeneous Biological Data. J. of Stat. Soft. 13(2):1-18.

Sponsored by : IEEE
For more information contact Dr. Gerardo Rosiles , grosiles@utep.edu