— Invited Talk —

Computational Analysis of High-throughput Genomics Data Using Hidden Markov Models and Support Vector Machines

David Haussler
haussler@cse.ucsc.edu
Computer Science Department, 317-A Applied Sciences Bldg.
University of California, Santa Cruz, CA95064 USA

The complete genomic sequence for several key model organisms is now available, and the human genome sequence will be nearly finished by next summer. Hidden Markov models (HMMs) have been used to find genes in genomic DNA produced by these genome projects, and to detect remote homologs of the proteins made by these genes. Protein homologs are used to help classify genes/proteins by structure or function. Here we look at new statistical tools to accomplish these classification tasks. These tools are called support vector machines (SVMs). SVMs classify data by embedding it in a high-dimensional space, and finding a separating hyperplane in that space. In recent work with Tommi Jaakkola and Mark Diekhans, we have shown that HMMs can be used to embed proteins in a high dimensional space, and then SVMs can be used to structurally classify them in that space. A study of 33 protein families showed that in nearly all cases this method worked better than key previous methods that have been used for this task. In recent work with Manny Ares, Michael Brown, Bill Grundy and others, we have shown that, in principle, SVMs can be used to functionally classify a gene based on mRNA expression measurements for the gene under different experimental conditions. Here we used mRNA expression data for yeast genes obtained from DNA microarrays at the Pat Brown laboratory at Stanford. This method can only work if the differences in mRNA expression levels between two genes over several experiments reveal the functional distinction between these genes that you are trying to recognize. We are still at the very early stages of investigating this issue, and of refining the method. Both our structural and functional classification methods are far from foolproof. However, we are encouraged by their potential. Further information can be found on the links from my home page, http://www.cse.ucsc.edu/~haussler.