BE.010 Spring 2005
Session #10 notes

Bioengineering Applications in Computer Science

Outline of this session:
- definition of machine learning and of its relevance to computational genomics research

Introduction

Professor of EECS David K. Gifford is the head of the Computer Genomics Group and a member of the Computer Science and Artificial Intelligence Laboratory (CSAIL) at MIT.

Machine Learning

Machine learning is the process by which a machine uses a sample training set to learn and then to generalize the data that it receives based on experience. Let us take handwriting analysis as an example. Machine learning would involve the development of a computer algorithm to recognize and interpret a person's handwriting based on a particular sample set. Although this can be done with relative ease in the human brain, this form of artificial intelligence is very difficult to program in computers.

One solution to the problem would be the memorization of large amounts of training data in hopes that all possible combinations of letters are covered. However, this method is not very effective due to the limited memory space of all machines. The goal of any type of machine learning is to get beyond the mere data and to draw general conclusions.

Expression Profiling Using DNA Microarray Technology

Recently, the emergence of many technologies that allow for the study of organisms at a genomic level has led to the demand for an efficient way to process the enormous amount of information that is retrieved from expression profiling. The goal is to create reasonably accurate models of genes in order to generalize their individual levels of expression.

Microarray technology makes use of data generated by genome projects to understand what types of genes are expressed in a particular cell in the organism at a particular time and under the influence of certain conditions. A microarray consists of a glass slide about one inch or less onto which DNA molecules are attached at fixed locations. A particular microarray may contain tens of thousands of such spots. Different gene expression levels can be measured by comparing cells exposed to different conditions.

Microarray technology uses fluorescent dyes to indicate the presence of messenger RNA. Experiments generate colored images which are then submitted for analysis. Based on fluorescence intensities and colors of each individual spot, the expression levels of particular genes can be determined through the use of image analysis software. All images must be properly scaled to make different arrays comparable.

Gifford Lab: Computational Genomics

The Gifford lab is currently developing computational algorithms to analyze the results of microarray research. Their goal is to develop a 3-D model of cellular activities and function based on different levels of gene expression.

Professor Gifford's lab is currently collaborating with Professor of Biology Richard A. Young's lab at the Whitehead Institute of Biomedical Research to develop a Genetic Regulatory Modules (GRAM) algorithm that can identify collections of genes that share common regulators and thus common expression profiles. GRAM combines the information from the microarray experiments previously discussed with protein-binding data to create genome-wide regulatory networks that describe cellular function. (MIT faculty website- www.psrg.lcs.mit.edu/people/gifford.html)

There are between 1,000 and 1,200 regulators in the human body. At the experimental level, yeast cells are exposed to various stressful conditions such as acid, heat, high/low peroxide to determine which regulators would bind to a particular protein. We determine the nucleic acid fragments that are bound to the protein and thus use the genome to figure out which regulators are binding.

Current Drawbacks

The problems of the GRAM algorithm are similar that of machine learning in handwriting analysis described earlier. Currently, the GRAM algorithm must try every single one of the 3x10^9 base pairs in the human genome to determine which one best matches the likelihood of the data. The correlation between a particular combination of transcriptional regulators to genes expressed must be very strong in order to be valid. Thus, the algorithm is repeated many times to detect genes with increasingly greater links to particular regulatory gene combinations.