Research plan

William Noble Grundy
Assistant Professor
Department of Computer Science
Columbia University

Work in my lab currently focuses on three areas: DNA microarray analysis, gene finding, and protein homology detection. Following are brief descriptions of these projects. Additional information may be found on my website, http://www.cs.columbia.edu/ bgrundy.

Gene prediction and functional annotation. A biologist with access to an array of genomic data of various types wants to answer two primary questions. First, where are the genes within the complete genomic sequence? Answering this question accurately involves not only recognizing the boundaries of the genes, but also locating their constituent elements, including exons, promoter regions, regulatory binding sites, and other key features. Given these elements, it is straightforward to infer the amino acid sequence of the protein derived from each gene. The second question involves classifying each protein according to its function. Such classification is usually accomplished by inferring homology, or evolutionary relatedness, between the protein of interest and some protein of known function.

We are working on a project that is divided into two phases, corresponding to the two tasks of gene prediction and functional annotation. In the first phase, we will develop and apply a gene finding system. This system is designed to be scalable and flexible with respect to the gene features it models, the machine learning algorithms it employs, and the range of experimental data from which it learns. Using as the core machine learning algorithm a discriminative training method based upon hidden neural networks, we will first validate the system by applying it to the complete *C. elegans* genome, comparing its predictions to those made by other gene finders. We will then retrain the system for the more difficult task of recognizing genes in human DNA.

The second phase of this project will consist of two parts. First, the software framework used for the gene finding system from phase one will be generalized to model families of related proteins. The resulting system will be capable of functionally classifying new proteins; the models employed, however, will be fundamentally sequence-based. In order to learn from non-sequential data such as mRNA expression data from microarray experiments, we will also develop functional classification techniques using support vector machines (SVMs). The statistics calculated by the sequence-based modeling system will function as one set of features used by the SVM system. Additional features will come from, for example, DNA microarray experiments, the upstream promoter region of each gene, similarity scores to known protein families, and three-dimensional structural information. The resulting discriminative classification system will provide excellent protein recognition capabilities.

The long-term goal of this research is to produce a system that not only functionally classifies genes, but also provides explanations for those classifications. The system developed here will be capable of providing functional explanations based upon the features of a single gene. Furthermore, the probabilistic nature of the models produced here will allow them to be incorporated into a future system that will be capable of producing contextual functional explanations on the level of complete pathways or even the entire cell.

Microarray expression analysis. With David Haussler and others at the University of California, Santa Cruz, I have developed a new method of functionally classifying genes using gene expression data from DNA microarray hybridization experiments. The method is based on the theory of support vector machines. SVMs are considered a supervised computer learning method because they exploit prior knowledge of gene function to identify unknown genes of similar function from expression data. SVMs avoid several problems associated with unsupervised clustering methods such as hierarchical clustering methods and self-organizing maps. SVMs have many mathematical features that make them attractive for gene expression analysis, including their flexibility in choosing a similarity function, sparseness of solution when dealing with large data sets, the ability to handle large feature spaces, and the ability to identify outliers.

We are currently working on improvements and extensions to this method. In particular, we are interested in developing SVM classifiers that exploit multiple data types simultaneously. In addition to microarray expression data, we have used phylogenetic profiles derived from whole-genome sequence comparisons, as well as features derived from the upstream, promoter region of each gene to be classified. This work involves the collection of appropriate features as well as the design of SVM kernel functions that incorporate prior or

empirically derived knowledge about those features.

Proten homology detection. Sequencing projects produce large quantities of biological sequence data; however, this data is only useful insofar as the functions of individual sequences are understood. Wet lab techniques for determining the function of a protein sequence can be time-consuming and expensive. Computational methods that infer sequence homologies (i.e., evolutionary relationships) can frequently replace wet lab experiments. I have developed several protein homology detection algorithms, Meta-MEME and Family Pairwise Search, that complement and improve upon existing techniques.

The Family Pairwise Search algorithm uses a linear combination of pairwise scores from a fast, heuristic algorithm to achieve better homology detection performance than sophisticated model-based techniques. Currently, we are investigating the use of this type of scoring system within the SVM learning algorithm. The pairwise comparison scores may be treated as similarity scores in an abstract feature space, thereby allowing the SVM to learn to discriminate protein family members within this space.