A Machine Learning Approach to DNA Microarray Bioinformatics

S. Y. Kung
Princeton University

Microarray bioinformatics is widely recognized to be a vital technology for drug design and disease classification. A collection of microarray experiments can yield a data matrix, whose rows standing for genes while columns for conditions (either independent or time-course). Note that a gene may participate in many pathways, each of which may be coactive under a subset of conditions. This motivates our study into a *bi-cluster analysis*, involving simultaneous classification of genes into functional groups and conditions into co-active categories.

The proposed machine learning system comprises three processing subsystems:

(1) *Feature extraction* to take into account the underlying biological coherence models.

(2) *Adaptive classification* techniques for bicluster discovery.

(3) *Multi-modality data fusion* for improving gene prediction accuracy.

The usefulness of bicluster discovery can be evaluated via its goodness for gene prediction. Various machine learning models - such as the support vector machine (SVM) and the decision-based neural networks (DBNN) – are considered. The biclustering technique is applied to various yeast gene groups, including (1) easy type (e.g. ribosomal genes) and (2) hard type (e.g. molecular activities genes). Preliminary simulations suggest that the proposed machine learning approach can yield highly competitive accuracy in terms of specificity, sensitivity, and/or precision.