Extracting Global Structure from Gene Expression Profiles

Charless Fowlkes[1], Qun Shan[2], Serge Belongie[1], Jitendra Malik[1]

[1]Computer Science, UC Berkeley; [2]Molecular Cell Biology, UC Berkeley

DNA microarray technology empowers biologists to analyze thousands of mRNA transcripts in parallel, providing insights about the cellular states of tumor cells, the effect of mutations and knockouts, progression of the cell cycle, and reaction to environmental stresses or drug treatments. Gene
expression profiles also provide the necessary raw data to interrogate cellular transcription regulation networks. Efforts have been made in identifying cis-acting elements based on the assumption that co-regulated genes have a higher probability of sharing transcription factor binding sites.

There is a well recognized need for tools that allow biologists to explore public domain microarray datasets and integrate insights gained into their own research. One important approach for structuring the exploration of gene expression data is to find coherent clusters of both genes and experimental conditions. The association of unknown genes with functionally well-characterized genes will guide the formation of hypotheses and suggest experiments to uncover the function of these unknown genes. Similarly, experimental conditions that cluster together may affect the same regulatory pathway.

Unsupervised clustering is a classical data analysis problem that is still an active area of intensive research in the computer science and statistics communities (see Ripley [7] for an overview). Broadly speaking, the goal of clustering is to partition a set of feature vectors into k groups such that the partitioning is "good" according to some cost function. In the case of genes, the feature vector is usually the degree of induction or suppression over some set of experimental conditions. As of yet, there is no clear consensus as to which algorithms are most suitable for gene expression data.
Clustering methods generally fall into one of two categories: central or pairwise [1]. Central clustering is based on the idea of prototypes, wherein one finds a small number of prototypical feature vectors to serve as "cluster centers". Feature vectors are then assigned to the most similar cluster center. Pairwise methods are based directly on the distances between all pairs of feature vectors in the data set. Pairwise methods don't require one to solve for prototypes which provides certain advantages over central methods. For example, when the shape of the clusters are not simple, compact clouds in feature space, central methods are ill-suited while pairwise methods perform well since similarity is allowed to propagate in a transitive fashion from neighbor to neighbor.

Clustering algorithms can also often be characterized as greedy or global in nature. The agglomerative clustering method used by Eisen et al. [4] to order microarray data is an example of a greedy pairwise method: it starts with a
full matrix of pairwise distances, locates the smallest value, merges the corresponding pair, and repeats until the whole dataset has been merged into a single cluster. Because this type of process only considers the closest pair of data points at each step, global

structure present in the data may not
be handled properly.

Another unsupervised clustering approach that has been applied to gene expression analysis is the self-organizing map [11]. While this technique is useful for structuring data sets in some applications, the lack of an explicit "energy function" has made it difficult to analyze.

Our approach to clustering gene expression data is based on the Normalized Cuts (NCut) method introduced by Shi and Malik [9][10]. Normalized Cuts is a pairwise clustering algorithm that finds a partitioning of the data set into well balanced groups. The resulting clustering minimizes a well de ned, global cost function. Experience in the field of computer vision, VLSI layout and parallel computing suggest that spectral partitioning methods [2] such as Normalized Cuts provide excellent results on a wide range of practical
problems. Section 2 outlines the NCut method for clustering and Section 3 demonstrates the application of NCut to the Rosetta yeast gene expression dataset [6].