EXTRACTING GLOBAL STRUCTURE FROM GENE EXPRESSION PROFILES

Charless Fowlkes[1], Qun Shan[2], Serge Belongie[3], and Jitendra Malik[1]
Departments of Computer Science[1] and Molecular Cell Biology [2], University of California at Berkeley; Department of Computer Science and Engineering, University of California at San Diego[3]

Abstract: We have developed a program, GENECUT, for analysing datasets from gene expression profiling. GENECUT is based on a pairwise clustering method known as *Normalized Cut* [Shi and Malik, 1997]. GENECUT extracts global structures by progressively partitioning datasets into well-balanced groups, performing an intuitive k-way partitioning at each stage in contrast to commonly used 2-way partitioning schemes. By making use of the *Nyström* approximation, it is possible to perform clustering on very large genomic datasets.

Keywords: gene expression profiles, clustering analysis, spectral partitioning

1. INTRODUCTION

DNA microarray technology empowers biologists to analyse 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 [Ripley, 1996]. Broadly speaking, the goal of clustering is to partition a set a feature vectors into *k* groups such that the partition 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* [Buhmann, 1995]. 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. A family of genes related by a series of small mutations might well exhibit this sort of structure, particularly when features are based on sequence data.

Clustering algorithms can also often be characterized as greedy or global in nature. The agglomerative clustering method used by [Eisen et al., 1998] 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 [Tamayo et al., 1999]. 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 [1997, 2000]. Normalized Cuts is a pairwise clustering that finds a partitioning of the data set into well-balanced groups. The resulting clustering minimizes a well-defined, global cost function. Experience in the field of computer vision, VLSI layout and parallel computing suggests that spectral graph methods [Chung, 1997] such as Normalized Cuts provide excellent results on a wide range of practical problems. In Section 2, we outline the NCut method for clustering and in Section 3, demonstrate the application of NCut to the Rosetta yeast gene expression dataset [Hughes et al., 2000].

2. CLUSTERING WITH NORMALIZED CUT

In this section we describe the NCut cost function, which provides a measure of cluster quality that takes into account both the within-group similarity and the between-group dissimilarity. We also outline the algorithm used for finding a clustering of the data that has low cost. The reader is referred to [Shi and Malik, 2000] and the references therein for additional detail.

2.1 The NCut Criterion

We use the Pearson correlation between vectors of expression data to capture the degree of similarity between two genes or two experiments. We will apply the same clustering algorithm to both the problem of clustering genes and that of clustering experiments so in this section we refer generically to the items being clustered. Let $\mathbf{W}_{\mathbf{ij}}$ be the Pearson correlation between the *ith* and *jth* data points. First consider the case of partitioning the dataset into two groups (bi-partitioning). Let **V** denote the complete set of data which is broken into subsets **A** and **B**. The NCut cost function is defined as

$$NCut\,(A,B) = \frac{cut\,(A,B)}{assoc\,(A,V)} + \frac{cut\,(A,B)}{assoc\,(B,V)}$$

where the *cut* and *association*, defined as

$$cut(A,B) = \sum_{i\in A, j\in B} wij \qquad assoc(A,V) = \sum_{i\in A, j\in B} wij$$

are graph-theoretic terms that quantify the cost of this partition (the cut) and the total connection of the subset to the whole set (the association). Normalizing by the association term makes NCut different from graph theoretic techniques based on min-cut (applied to genomic data by [Sharan and Shamir, 2000] which can generate highly unbalanced clusters and require elaborate post-processing ([Shi and Malik, 2000] provides a comparison).

While finding the **A-B** partition that minimizes the NCut criterion is an NP-hard optimization problem, it is possible to relax the constraints in order to obtain a closed form eigenproblem that yields high quality approximations. The problem is formulated in terms of minimizing the Rayleigh quotient,

$$\frac{y^T(D-W)y}{y^T Dy}$$

where **W** is the matrix whose entries are $\mathbf{W_{ij}}$, **D** is a diagonal matrix with $\mathbf{D_{ii}} = \mathbf{\Sigma\ W_{ij}}$ and **y** is a partition indicator vector. If we allow **y** to take on continuous values then the minimum is obtained by the second leading eigenvector of the generalized eigenvalue problem **(D-W)y=λDy.**

2.2 K-Way Partitioning

The NCut bi-partitioning technique has been applied to genomic expression data by [Xing and Karp, 2001] for a data set containing two clusters. However, for the analysis of a large compendium of expression data, we would expect there to exist far more than two clusters. Generalization to the case of more than two groups can be obtained in a number of ways. One method is to apply bi-partitioning recursively on **A** and **B**. Another method is to compute k leading eigenvectors instead of just the second one: this leads to a k-dimensional *embedding* that is amenable to clustering with simple central methods such as k-means [Duda and Hart, 1973]. The approach taken in our present work is a combination of these two methods. We perform a recursive k-way clustering where k is automatically chosen at each level to minimize the k-way NCut criterion defined as

$$NCut_k(A_1, A_2, \ldots, A_k) = \frac{1}{k}\sum_{i=1}^{k}\frac{cut(A_i, V - A_i)}{assoc(A_i, V)}$$

We find that this criterion constitutes an effective form of *model selection* and yields natural clusters while avoiding the artificial constraint of bi-partitioning or pairwise merging schemes.

2.3 Clustering Large Datasets

Our algorithm was prototyped in MATLAB where it takes less than a minute to cluster the 560 genes used in our experiments. For very large problems, the computation and memory requirements to solve the eigenproblem can become a limiting factor for interactive data analysis. To avoid these costs, we can exploit the *Nyström Approximation* which allows one to extrapolate the solution to a large clustering problem using a small subset of the data [Fowlkes et al., 2001].

This approximation exploits redundancy between rows of the $\mathbf{W}_{ij}$ matrix by choosing a small subset of the genes and computing their similarity to every other gene in the dataset. This thin strip of the matrix is then used to compute a direct numerical approximation to the eigenvectors needed for partitioning. The memory and processing expenses grow in proportion to the number of samples rather than the total number of data points so by using this approximation, our method should extend efficiently to the analysis of complete genomes with thousands of experiments.

3. RESULTS

We have built a system for interactively browsing the results of the NCut algorithm called GENECUT. The clustering results presented in this paper along with prototype software are available at **http://www.cs.berkeley.edu/~fowlkes/bio/**. In this section we present some results that indicate our algorithm is capable of finding clusters that exist in the data. A robust algorithm is extremely important since true clusters in a data set are unknown and poor clustering results could easily be misleading. While it is difficult to evaluate the performance of clustering algorithms quantitatively, we are able to point to clusters of well characterized genes which have closely related functions, suggesting that the algorithm is effective.

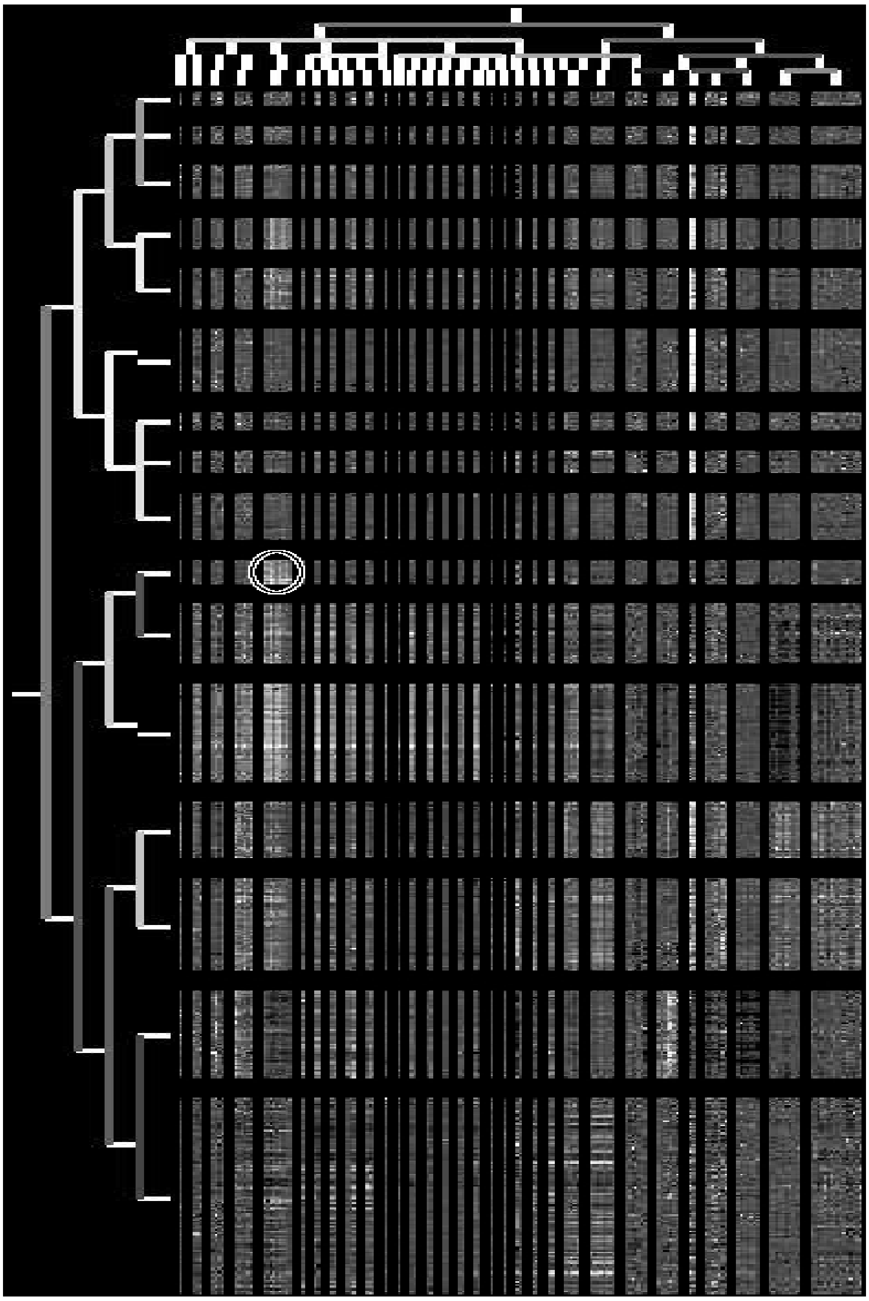

Figure 1. The result of performing a recursive, k-way partitioning on a set of 560 genes and 123 experiments. Genes are arrayed along the x-axis and experiments along the y-axis. The contents of the cluster indicated by the white circle are listed in Table 1 and Table 2. The color-coding on the tree indicates the cost of the associated k-way cut. The contents of other clusters are available for interactive exploration: **http://www.cs.berkeley.edu/~fowlkes/bio/**

Figure 1 gives a visual overview of the clustering analysis presented by GENECUT for the Rosetta gene expression dataset [Hughes et al., 2000]. The output of the clustering algorithm is presented in the form of a web page that allows the user to traverse up and down through the layers of the tree structure in both the experimental and gene dimensions. The user can click on clusters in the overview image in order to view the genes and experiments in that cluster. Gene descriptions include links to detailed descriptions and a link that invokes a BLAST search of the *Saccharomyoes* Genome Database using the 500 bp upstream sequence.

We expect that clusters of genes showing similar expression patterns are likely to share some conservative regulatory motif. The ability to do a BLAST query quickly is a first step towards seeking similar transcription factor binding sites. We are currently exploring DNA motifs associated with several of these clusters. Automatic identification of these putative motifs would clearly be helpful in experimental design.

Experiment #	**Description**	
9	erg2	Deletion
10	erg3	Deletion
107	hmg1	Deletion
61	Yer044c (haploid)	Deletion
29	ERG11 (tet promoter)	Shutdown
35	HMG2 (tet promoter)	Shutdown
73	Lovastatin	drug treatment
82	Terbinafine	drug treatment
71	Itraconazole	drug treatment

Table 1. Experiment cluster #5, an interesting group of experiments found by GENECUT (shown circled in Figure 1). This cluster contains experimental conditions relating to the sterol synthesis pathway.

Table 1 shows a cluster along the experimental axis that groups together a set of experiments that all involve perturbations of sterol biosynthesis. To extract global features from an experimental cluster like these sterol synthesis experiments, we sort the gene clusters by their normalized variances. We reason that the makeup of gene clusters with high variance across a particular experiment cluster is likely to be biologically relevant.

Table 2 lists the gene cluster that has the highest mean variance in expression level for the sterol synthesis experiments cluster. This gene cluster makes biological sense and also agrees with a visual examination of the datawset

Gene	System Name	Description
1	YHR007C	[*ERG11*] Cytochrome P450 (lanoterol 14 alpha-demethlase), essential for biosynthesis of ergosterol
110	YDR530C	[*APA2*] ATP adenylyltransferase II
169	YGL001C	[*ERG26*] C-3 sterol dehydrogenase, C-4 decarboxylase, required for ergosterol biosynthesis
195	YGR049W	[*SCM4*] Protein that suppressed temperature-sensitive allele of CDC4 when overexpressed
197	YGR060W	[*ERG25*] C-4 sterol methyl oxidase: enzyme of the ergosterol biosynthesis pathway
210	YGR175C	[*ERG1*] Squalene monooxygenase (squalene epoxidase), an enzyme of the ergosterol biosynthesis pathway
279	YJL113W	Unknown
337	YKRO53C	[*YSR3*] Sphingoid base-phosphate phosphatase, putative regulator of sphingolipid metabolism and stress response
344	YLL0112W	Protein with similarity to human triacylglycerol lipase
380	YML008C	[*ERG6*] S-adenosylmethionine delta-24-sterol-C-methyltransferase, carries out methylation of zymosterol as part of the ergosterol biosynthesis pathway
392	YMR015C	[*ERG5*] Cytochrome P450, delta 22(23) sterol desaturase, catalyses an intermediate pathway step in the biosynthesis pathway
434	YNL111C	[*CYB5*] Cytochrome b5
491	YOR237W	[*HES1*] protein implicated in ergosterol biosynthesis, member of the KES1/HES1/OSH1/YKR003W family of oxysterol-binding (OSBP) proteins
511	YOR394W	Member of the seripauperin (PAU) family (YPL282C and YOR394W code for identical proteins)
523	YPL272C	Unknown

Table 2. Gene cluster #10 found by GENECUT contains genes related to sterol biosynthesis. This cluster had the largest variance across experimental conditions for the set of experiments in experiment cluster #5

Many easily identified clusters discussed in [Hughes et. al. 2000] were also found by the GENECUT algorithm. This is notable since the two algorithms employed take quite different approaches (local agglomerative vs. global divisive). Figure 2 contrasts the genes found by our algorithm with those of [Hughes et. al. 2000] for the sterol gene cluster (our cluster #10). Genes that appear in the intersection of the two clusters are presumed to be related with high confidence while those which only appear in a single cluster require more experiments to pin down. Since the agglomerative clustering algorithm produces a dendrogram whose leaves are individual genes, the cluster shown is actually a manually selected sub-tree.

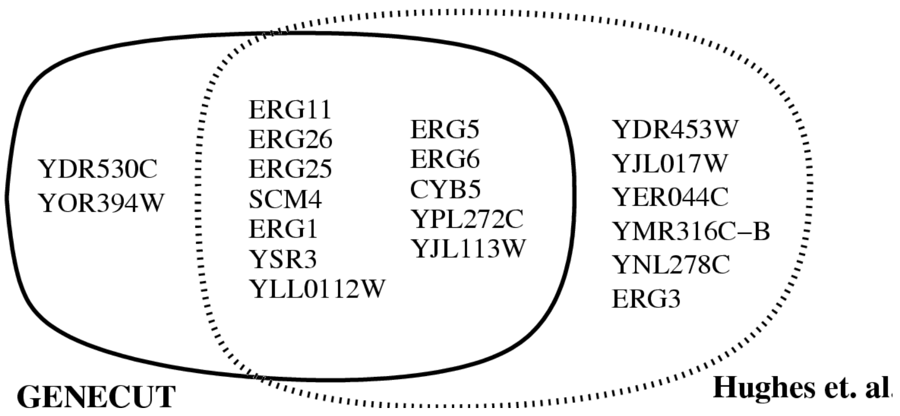

Figure 2. A comparison of the "sterol" cluster found by [Hughes et. al. 2000] (dotted circle) and that found by the GENECUT algorithm (solid circle). As with many other clusters, there is significant overlap.

4. CONCLUSIONS

In this report, we developed a novel application of the NCut algorithm to the problem of gene expression profile analysis. The algorithm performs favourably by focusing on the global features and recursively partitioning the dataset into clusters. We demonstrate the utility of NCut in extracting global features from an experiment cluster, and further explore regulatory sequences within the representative gene clusters. It may be possible to use this algorithm effectively in conjunction with hierarchical clustering tools in order to perform "harvesting" of dendrograms and allow rapid exploration of genomic data sets. We envision that this algorithm can ultimately be used as a general clustering tool in various areas of genomics research such as protein classification, DNA sequence data, and drug sensitivity profiling.

5. REFERENCES

Buhmann JM (1995) Data Clustering and Learning. In: Arbib MA (ed) The Handbook of Brain Theory and Neural Networks. MIT Press.

Chung FRK (1997) Spectral Graph Theory. American Mathematical Society.

Duda R and Hart P (1973) Pattern Classification and Scene Analysis. John Wiley & Sons.

Eisen MB et al. (1998) Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci 95: 14863 – 14868.

Fowlkes C, Belongie S and Malik J (2001) Spatiotemporal grouping using the Nyström approximation. In: Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.

Hughes TR and Marton MJ et al. (2000) Functional discovery via a compendium of expression profiles. Cell 102:109-126.

Ripley BD (1996) Pattern Recognition and Neural Networks. Cambridge.

Sharan R and Shamir R (2000) Click: A clustering algorithm with applications to gene expression analysis. In: Proc. Of ISMB, pages 307-316, AAAI Press.

Shi J and Malik J (1997) Normalized cuts and image segmentation. In: Proc IEEE Conf. Computer Vision and pattern Recognition, pages 731-737.

Shi J and Malik J (2000) Normalized cuts and image segmentation. IEEE Trans. PAMI 22: 888-905.

Tamayo P et al. (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and applications to hematopoietic differentiation. Proc. Natl. Acad. Sci. 96: 2907-2912.

Xing EP and Karp RM (2001) Cliff: Clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. In: Proc. Of the Nineteenth ISMB.