Comparative Study of Clustering Techniques for Gene Expression Microarray Data

Sonia Leach[1,2] **Larry Hunter**[1]
sml@cs.brown.edu lhunter@nih.gov

[1] Molecular Statistics and Bioinformatics Section, Biometrics Research Branch, National Cancer Institute, 7550 Wisconsin Ave, Room 3C06, Bethesda MD 20892-9015, USA
[2] Department of Computer Science, Brown University, 115 Waterman Street, Providence RI 02901, USA

Keywords: clustering, gene expression microarrays

1 Introduction

Scientists can now monitor on a genomic scale the patterns of gene expression under varying environmental conditions. With this rapidly growing wealth of information comes the need for organizing and analyzing the data. One natural approach is to group together genes with similar patterns of expression. Several approaches have suggested various alternatives for similarity metrics and clustering algorithms [1, 2, 4, 7, 8]. However, the studies reported on the use of a particular choice of metric and clustering algorithm, with little comparison to other techniques. In this paper, we systematically investigate the relative merits of several well-known metrics and clustering algorithms as applied to gene expression data. Our intent is to provide biologists with a useful guide for choosing a technique to analyze gene relationships in their data, by alerting them to the strengths and limitations inherent in that choice.

2 Method and Results

The gene expression dataset used in this study is a subset of the publicly available collection of yeast expression data (http://www.cmgm.stanford.edu/pbrown/). The data are in the form of log-ratios of mRNA expression levels between an experimental and control sample of cells. Combining data from mitotic cell division, diauxic shift, and sporulation, we obtained transcript levels over 92 experimental conditions. We chose 1853 genes showing significant variation at the 1% level under a χ^2-based index. Of these genes, 45% have a MIPS functional classification (http://www.mips.biochem.mpg.de/). Four percent of the values in the dataset of 1853 genes and 92 experiments were missing; 1110 genes had at least one missing value. To investigate the effect of missing values, we created an additional dataset where the missing values for a particular gene were imputed using the average of the 30 nearest neighbors.

We compared several clustering techniques, including a bayesian mixture of models algorithm (AutoClass-C v3.2 [3]), K-means, and agglomerative hierarchical clustering. For the hierarchical method, we used a euclidean metric, correlation based metrics, and a mutual information based metric [8]. To apply the mutual information based metric, we discretized the data, using AutoClass-C to suggest the number of bins and bin boundaries. For linkage algorithms, we considered both complete linkage (furthest neighbor) and the variant of average linkage algorithm as used in [4] when the use of such an average is appropriate given the metric. To compensate for missing values, euclidean distance between two genes is weighted by $\sqrt{92/N}$ where N is the number of non-missing values common to the two genes. To estimate the number of clusters to use in K-means, we consider a mean square error ratio [5], a log-odds probability ratio, and a Monte Carlo Cross Validation technique [6] for a range of values of K. We found that the three estimators tend to agree on the optimal value of K.

Using a correlation metric and an average-link variant of hierarchical clustering, [4] found that for this clustering combination, large groups of genes with similar expression patters clustered together, and that genes of similar function tend to cluster together. Using the non-hierarchical methods, we were able to compare the consistency of the clusterings across many different methods. We compared four clusterings found with Autoclass, using four different model term types, and two different K-means clusterings. Intersecting these six classifications, we observed that large groups of genes are preserved across clustering methods and that these groups tend to have a consistent functional classification. The same clusters were also visible in many of the hierarchical clusters, which leads us to believe that there are easily identifiable classes that will be found regardless of the choice of metric or clustering algorithm. In particular, both AutoClass and K-means can easily identify clusters with high *linear* correlation. The choice of metrics used in hierarchical clustering attempted to capture other correlations as well.

For the hierarchical clusterings, we found that the choice of the average-link variant used in [4] with the euclidean or the correlation based metrics tends to produce long chains of clusters. This makes it difficult to extract meaningful classes from the dendrogram because cuts must be made deep in the tree to get non-singleton classes. This greatly complicates the cross-method comparisons and any measures that are calculated on a per-cluster basis. To capture negative linear correlation, we also used a metric based on the absolute value of correlation. Such a metric might be used to identify genes which share a common transcription factor binding site with opposing effect on the activation of the genes. As expected, we found that in addition to the easily identifiable positively correlated groups, the clusters of the dendrogram tend to include highly negatively correlated genes as well. Use of the mutual information based metric was an attempt to capture other kinds of relationships of genes, beyond the linear correlations. We found several groups which clearly show the linearly correlated genes, but in addition include genes whose expression pattern is not so obviously related. We are investigating the biological significance of the inclusion of these genes.

References

[1] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J., Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci. USA*, 96(12):6745–6750, 1999.

[2] Ben-Dor, A., Shamir, R., and Yakhini, Z., Clustering gene expression patterns, *J. Comput. Biol.*, 6:281–297,1998.

[3] Cheeseman, P., Kelly, J., Self, M., Stutz, J., Taylor, W., and Freeman, D., AutoClass: a Bayesian classificiation system, *Proc. Inter. Conf. Machine Learning '88*, Morgan Kaufmann, 54–64, 1988.

[4] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D., Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, 95(25):14863–14868, 1998.

[5] Hartigan, J.A., *Clustering Algorithms*, Wiley Press, 1975.

[6] Smyth, P., Clustering using Monte Carlo Cross Validation, *Proc. 2nd Inter. Conf. Knowledge Discovery and Data Mining*, 126–133, 1996.

[7] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R., Interpreting patterns of gene expression with self-organizing maps, *Proc. Natl. Acad. Sci. USA*, 96(6):2907–2912, 1999.

[8] Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L. and Somogyi, R., Large-scale temporal gene expression mapping of central nervous system development, *Proc. Natl. Acad. Sci. USA*, 95(1):334–339, 1998.