Comparison of Clustering Metrics and Unsupervised Learning Algorithms on Genome-Wide Gene Expression Level Data

Sonia Leach† **Lawrence Hunter**‡ **David Landsman**§

†Brown University
Box 1910
Providence RI 02912
sml@cs.brown.edu

‡National Cancer Institute, MS-9105
7550 Wisconsin Ave., Room 3C06
Bethesda, MD 20892-9015
hunter@msb.nci.nih.gov

§National Center for Biotechnology Information
National Library of Medicine
Bethesda MD 20894
landsman@ncbi.nlm.nih.gov

With the recent availability of genome-wide DNA sequence information, biologists are left with the overwhelming task of identifying the biological role of every gene in an organism. Technological advances now provide fast and efficient methods to monitor, on a genomic scale, the patterns of gene expression in response to a stimulus, lending key insight about a gene's function. With this wealth of information comes the need to organize and analyze the data.

One natural approach is to group together genes with similar patterns of expression. Several alternatives have been proposed for both the similarity metric and the clustering algorithm (Wen *et al.* 1998; Eisen *et al.* 1998). However, these studies used a specific metric-clustering algorithm pair. In our work, we aim to provide a more systematic investigation into the various metric and clustering algorithm alternatives. We also offer two methods to handle missing data.

The data sets include a single time course of rat spinal cord development, a single time course of a human cell growth model, and an aggregation of data from the yeast *S. cervisiae* under several experimental conditions. The data contains missing datapoints in cases of measurement error or inconclusive signal. We consider two techniques for handling missing datapoints, namely weighting by the number of valid points, and linear interpolation.

For similarity metrics, we compare a euclidean distance metric, a correlation metric, and a mutual information-based metric. The euclidean metric is commonly used due to its spatially intuitive interpretation of distance and ease of calculation. However, it might fail to recognize negative correlation, thus we use sample correlation to capture both positive and negative correlation. Not all the significant relationships between genes are modelled under either metric. In particular, both summarize the contributions along the whole trajectory, assuming that the type of correlation is constant throughout time. Two genes might be correlated positively within a certain range of their values and negatively related in another range. To capture this type of dependence, we consider a third metric based on mutual information.

We combine the metrics with unsupervised learning algorithms. We consider k-means, hierarchical agglomerative clustering and AutoClass (Cheeseman *et al.* 1988). The advantage of k-means is its simplicity. However, many iterations may be required and the number of clusters must be specified *a priori*. The hierarchical clustering algorithm attempts to cluster all elements (pairwise) into a single tree. The algorithm is fast and requires no *a priori* knowledge of the number of clusters. However, the cluster boundaries are usually manually extracted based on other known information, such as gene function. AutoClass (Cheeseman *et al.* 1988) can be viewed as a stochastic version of k-means which models the means as gaussians. AutoClass automatically searches for the number of clusters. However, as an approximation to the full-Bayesian classification, AutoClass can be slow on large problems.

We address questions like how different are the results using a particular combination of metric, clustering algorithm and missing value compensation? Which one is the best for this important application? How stable and reliable are the clusters? How well do the clusters match known functional classes? How sensitive are the methods to missing values?

The work by Sonia Leach was funded by a fellowship from the Rhode Island Space Grant Program under the National Space Grant College and Fellowship Program.

References

Cheeseman, P.; Kelly, J.; Self, M.; Stutz, J.; Taylor, W.; and Freeman, D. 1988. AutoClass: A Bayesian classificiation system. In *Proceedings of the Fifth International Conference on Machine Learning*, 54–64.

Eisen, M. B.; Spellman, P. T.; Brown, P. O.; and Botstein, D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences USA* 95:14863–14868.

Wen, X.; Fuhrman, S.; Michaels, G. S.; Carr, D. B.; Smith, S.; Barker, J. L.; and Somogyi, R. 1998. Large-scale temporal gene expression mapping of central nervous system development. *Proceedings of the National Academy of Sciences* 95(1):334–339.