Unsupervised Clustering Analysis of Gene Expression

Haiyan Huang, Kyungpil Kim

The availability of whole genome sequence data has facilitated the development of high-throughput technologies for monitoring biological signals on a genomic scale. The revolutionary microarray technology, first introduced in 1995 (Schena et al., 1995), is now one of the most valuable techniques for global gene expression profiling. Other high-throughput genomic technologies, such as Serial Analysis of Gene Expression (SAGE) (Velculescu et al., 1995), mass spectrometry for protein identification (Henzel et al., 1993) and ChIP-chip for DNA binding (Ren et al., 2000), have also been widely used for different purposes in current biological and medical research.

With the consequent explosion of various genomic data, a general question facing the biologists and statisticians is how to organize the observed genomic data into meaningful structures. Cluster analysis methods have been widely explored for this purpose; that is to cluster biological objects sharing common characteristics into discrete groups. Such analyses allow the researchers to develop an integrated understanding of underlying biology. In general, clustering methods can be divided into two categories. The hierarchical methods produce nested clusters; the non-hierarchical methods divide a dataset of N objects into M clusters, with or without overlap (Hartigan 1975; Eisen et al. 1998; Tamayo et al. 1999).

In an expression matrix, each gene corresponds to one row and each condition/sample to one column. Common tasks in clustering analysis of expression data include i) grouping genes by their expressions over conditions/samples, ii) grouping conditions/samples based on the expression of genes, and iii) finding subgroups of genes and conditions/samples such that the identified genes share similar expression patterns over a specified subset conditions/samples. The first analysis has been widely performed in current biological research for discovering and understanding gene functional relationships. The second analysis is generally used in biomedical research, especially in clustering disease samples to diagnose disease types or disease progress. The third type of analysis, also known as a bi-clustering or two-way clustering, is useful in the circumstances when both genes and conditions/samples are expected to contribute to the uncovering of meaningful cluster patterns. We refer to Eisen et al. (1998) and Madeira and Oliveira (2004) for more descriptions on these analyses in biology. In the following, we illustrate different methods, with its advantages and pitfalls, for the first type of analysis.

Hierarchical clustering analysis of Microarray expression data

In hierarchical clustering, relationships among objects are represented by a tree whose branch lengths reflect the degree of similarity between objects. Compared to non-hierarchical clustering methods, hierarchical methods give a lot more object relationship information. In particular, the hierarchical dendrogram can help visualize the object relationship structure between and within clusters.

The demonstration dataset is a yeast sporulation dataset. Chu et al. (1998) measured gene expression in the budding yeast *Saccharomyces cerevisiae* at seven time points during sporulation using spotted microarrays and identified seven distinct temporal patterns of induction. 39 genes representing each of these seven patterns have been used to define model expression profiles in that study. **Fig. 1(a)** plots the gene expression profiles of these seven patterns, named as Metabolic, Early I, Early II, Early-Mid, Middle, Mid-Late and Late according to the functional behavior of the genes at different time points. **Fig. 1(b)** shows the hierarchical clustering results on these 39 representative genes. We see that the identified 7 clusters in **Fig. 1(b)** are largely consistent with the known gene function categorizations.

However, in hierarchical clustering, the dendrogram must be cut manually at some places to obtain clustering results. This means that it tends to concentrate on local clusters instead of global expression pattern. Furthermore, due to this local manner, some small errors in the early stages of cluster assignment can be drastically amplified in the final result.

Non-hierarchical clustering analysis of SAGE data

The non-hierarchical clustering algorithms, in particular the K-means clustering algorithm, run fast and consume less memory compared to hierarchical clustering algorithms. Due to the use of global properties of data, the clustering quality of a non-hierarchical method can also be advantageous over hierarchical methods in some circumstances.

The demonstration dataset is a SAGE dataset. SAGE is one of the effective techniques for comprehensive gene expression profiling. The result of a SAGE experiment, called a SAGE library, is a list of counts of sequenced tags. Each tag is of 10-11 base pairs long and acts like a bar-code, containing the information to distinguish between transcripts. Ideally, each tag is uniquely mapped to a gene, and thus its counts reflect the corresponding gene's expression level. The SAGE data can be naturally modeled by Poisson distributions due to the data nature; they are generated by "sampling," which results in counts. Accordingly, Cai et al. (2004) developed two Poisson-based measures and employed them into a K-means clustering procedure to group tags with similar count profiles across libraries. We will apply this method to a small SAGE dataset for illustration.

The dataset consists of 125 annotated tags with counts in 10 SAGE libraries from developing retina taken at 2 day intervals, ranging from embryonic to postnatal and adult (Blackshaw et al., 2004). These 125 tags fall into 4 functional groups based on their known behavior at different developmental stages (Table 1). **Fig. 2** shows the clustering results by the modified K-means clustering procedure in Cai et al. (2004) (named *PoissonC*). The results are largely consistent with the known functional categorizations of the genes. Actually 106 out of 125 tags are correctly clustered. Furthermore, when un-annotated genes are involved, the functions of annotated genes can be used to predict the functions of un-annotated genes in the same cluster. However, we should note that non-hierarchical algorithms can not give relationship information for genes within a cluster, or in other words, it can not distinguish

between genes within the cluster.

The dataset used here is a subset of the one used in Huang et al. (2006) for demonstrating the advantages of *PoissonC* over the K-means clustering procedure using Pearson correlation or Euclidian distance as similarity measures. We want to point out that the main advantage of *PoissonC* is that both the data magnitude and shape of expression curves are considered in grouping the genes. While Euclidian distance only cares about the magnitude and Pearson correlation can be overly sensitive to the shape. An appropriate similarity measure is critical for a successful cluster method.

Principal component analysis

Principal components analysis (PCA) (Jolliffe, 1986) is a statistical technique for determining key features of a high dimensional dataset in order to simplify analysis. Recently, it has been explored as a method for clustering gene expression data, as it is fast and can handle large datasets.

We use the same SAGE dataset to demonstrate the effectiveness of PCA. We first apply PCA to the normalized data. That is that we rescale and center the counts for each tag such that each tag has zero mean and unit variance of the counts across libraries. When the known four classes are viewed in the space of the first three principal components (PC's) (**Fig. 3(a)**), the four classes are reasonably well-separated in the Euclidean space. Especially we see a very clear separation between the early and late genes.

However, the first a few PCs are not guaranteed to always give good clustering results, since the PCs are determined by the empirical variance of overall data rather than class information (Yeung and Ruzzo 2001; Komura et al. 2005). This argument can be demonstrated by applying PCA to the original SAGE data. When no normalization is performed on the data, we see that the genes can not be well separated when they are viewed in the Euclidian space of the first three PCs (**Fig. 3(b-I)**). To understand this result, we further view the genes in the space of the first two PCs (**Fig. 3(b-II)**) and the space of the second and third PCs (**Fig. 3(b-III)**). From **Fig. 3(b-II),** we see that the first PC is not helping to cluster the genes correctly in the Euclidian space and also there is no hint on what other measures could be used to effectively separate the genes. While in the space of the second and third PCs (**Fig. 3(b-III)**), we see that the slope of the line connecting each point to original could be a good measure to cluster the genes. Motivated by this, we project the points onto the unit circle in the same space. The four classes are then reasonably well-separated (**Fig. 3(b-IV)**). The results in **Fig. 3(b-IV)** are comparable to the ones shown in **Fig. 3(a)**.

The above interesting observation can be explained by the following arguments: without normalizing the data, the most variance of data comes from the magnitude variation of expression levels among genes, which is the main information that the first PC has captured. However, for this particular dataset, "shape" of the expression curve, rather than the magnitude of expression, is a key feature in determining a class. Therefore the first PC is

not helpful in distinguishing between classes. This also explains the good result in **Fig. 3(a)**, where the variation of magnitude has been reduced by the normalization done before analyzing the data. But we should also note that the data normalization is not guaranteed to always help. When the data structure is complicated, there can be other sources of variations, which may reflect the within-class differences rather than between-class differences and could be not reducible by simple normalization. This type of variation, even though when it is big, should not be used in clustering. This raises an important but un-touched issue in PCA clustering analysis: how to estimate a biologically meaningful covariance matrix, instead of using the sample covariance matrix, to capture between-class variations according to the clustering purpose and data property. An initial attempt on this issue can be found in Jiang et al. (2006).

Summary

The clustering analysis can be applied much more broadly than we have described. Other immediate examples include the clustering analysis performed in comparative genomics and statistical genetics (Slack et al., 2005), i.e. carrying out a clustering analysis on multiple strains using whole strain comparison, genetic marker data and evolutionary data.

Generally, clustering techniques can work better with more background information. When enough prior knowledge is available, supervised clustering analysis can be performed. Supervised clustering, also regarded as classification, classifies the objects with respect to known reference data (Dettling and Bühlmann, 2002). For the issue discussed in PCA analysis above, supervised method could be beneficial to capture between-class variations with the amount of reference data or prior knowledge on data clusters available.

Table 1. Functional categorization of the 125 mouse retinal tags

	Function Groups				Total
	Early I	**Early II**	**Late I**	**Late II**	
Number of tags	32	34	32	27	125

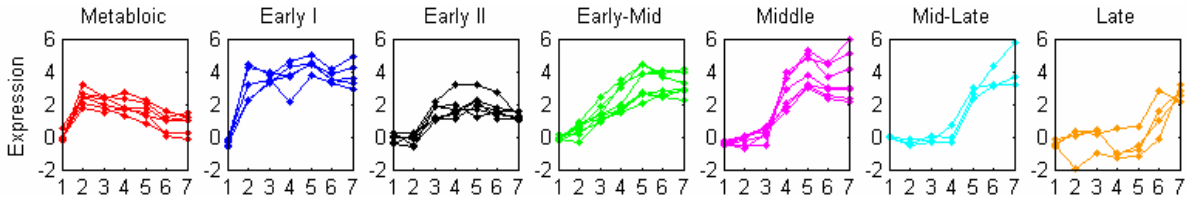

Fig. 1(a). Expression profiles of the 39 representative genes from 7 functional groups. The x-axis represents different time points of 0h, 0.5h, 2h, 5h, 7h, 9h, 11.5h.; the y-axis represents the normalized log-ratio expression levels.

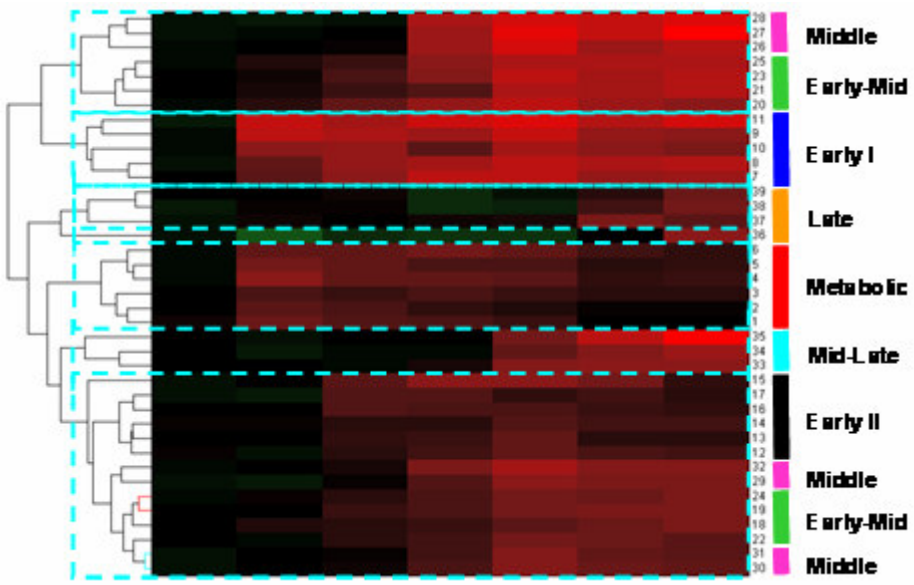

Fig. 1(b). Graphic display of hierarchical clustering (default setting: Pearson correlation; average linkage) results on the yeast sporulation dataset of 39 genes. Each gene is represented by a single row of colored boxes; each time point is represented by a single column. Seven separate clusters determined by the algorithm are indicated by blue dotted boxes. The known functional categorization of each gene is indicated by the colored bar on the right side (same color means same functional group).

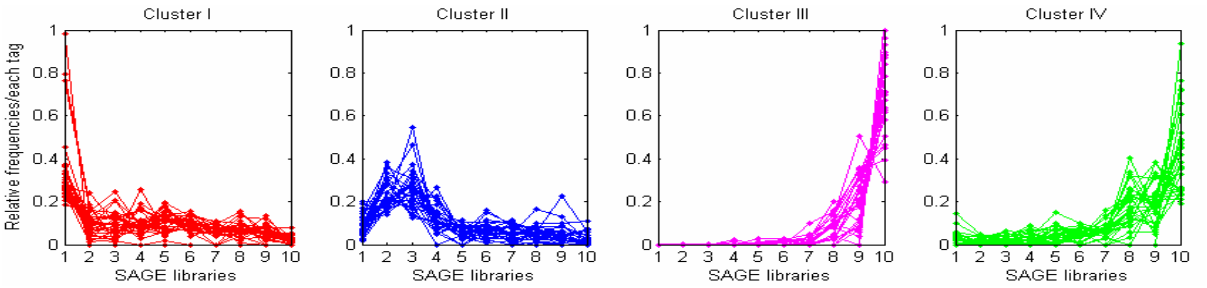

Fig. 2. Expression curves of genes from the four clusters determined by *PoissonC*. This result is largely consistent with the known functional categorization in Table 1. Cluster I corresponds to the group of Early I; Cluster II corresponds to group Early II; Cluster III corresponds to group Late I; Cluster IV corresponds to group Late II.

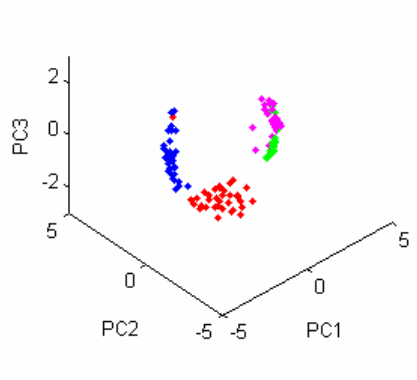

Fig. 3(a). Visualization of the known four classes from normalized SAGE data in the space of the first three PC's. Blue points represent genes from Early I group; Red points are Early II genes; Green points are Late I genes; Pink points are Late II genes. Please see Table 1 and Fig. 2 for more information on the data.

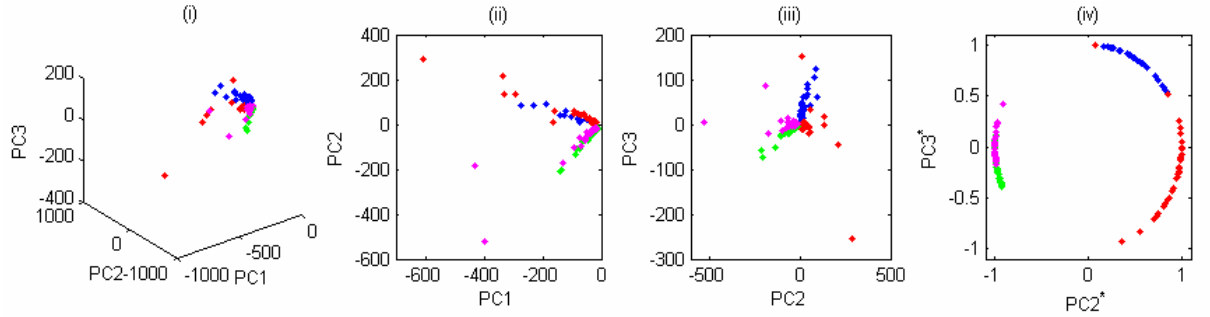

Fig. 3(b). Visualization of the known four classes from original SAGE data (un-normalized) in the space of the (i) first three PCs, (ii) first two PCs, and (iii) PC2 and PC3. In (iv), the points in (iii) are projected onto the unit circle for a better visualization. Blue points represent genes from Early I group; Red points are Early II genes; Green points are Late I genes; Pink points are Late II genes.

Reference

1. Blackshaw, S., Harpavat, S., Trimarchi, J., Cai, L., Huang, H., Kuo, W. P., Weber, G., Lee, K., Fraioli, R. E., Cho, S. H., Yung, R., Asch, E., Ohno-Machado, L., Wong, W. H. and Cepko, C. L. (2004), "Genomic Analysis of Mouse Retinal Development," *PLoS Biology,* 2, e247.
2. Cai, L., Huang, H., Blackshaw, S., Liu, J. S., Cepko, C. L. and Wong, W. H. (2004), "Clustering Analysis of SAGE Data Using a Poisson Approach," *Genome Biology*, 5, R51.
3. Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Bostein, D., Brown, P. O. and Herskowitz, I. (1998), "The Transcriptional Program of Sporulation in Budding Yeast," *Science*, 282, 699-705.
4. 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 of the United States of America*, 95, 14863-14868.
5. Hartigan, J.A. (1975), *Clustering Algorithms*, New York: John Wiley & Sons.
6. Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C., Grimley, C. and Watanabe, C. (1993), "Identifying Proteins From 2-Dimensional Gels by Molecular Mass Searching

of Peptide Fragments in Protein Sequence Databases," *Proceedings of the National Academy of Sciences of the United States of America*, 90, 5011-5015.

7. Huang, H., Cai, L. and Wong, W. H. (2006), "Clustering Analysis of SAGE Transcription Profiles Using a Poisson Approach," in *SAGE: Methods and Protocols*, ed. K. L. Nielsen, Humana Press Inc.
8. Jiang, K., Zhang, S., Lee, S., Tsai, G., Kim, K., Huang, H., Zhu, T. and Feldman, L. J. (2006), "Transcription Profile Analyses Identify Genes and Pathways Central to Root Cap Functions in Maize," *Plant Molecular Biology,* PUBLISHed.
9. Jolliffe, I. T. (1986), *Principal Component Analysis,* Springer-Verlag, New York.
10. Komura, D., Nakamura, H., Tsutsumi, S., Aburatani, H. and Ihara, S. (2005), "Multidimensional Support Vector Machines for Visualization of Gene Expression Data," *Bioinformatics*, 21, 439-444.
11. Madeira, S. C. and Oliveira, A. L. (2004), "Biclustering Algorithms for Biological Data Analysis: a Survey," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1, 24-45.
12. Man, M. Z., Wang, X. and Wang, Y. (2000), "Power-SAGE: Comparing Statistical Tests for SAGE Experiments," *Bioinformatics*, 16, 953-959.
13. Dettleing, M. and Bühlmann, P. (2002), "Supervised Clustering of Genes," *Genome Biology*, 3, research0069.1-0069.15.
14. Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P. and Young, R. A. (2000), "Genome-Wide Location and Function of DNA Binding Proteins," *Science*, 290, 2306-2309.
15. Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995), "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray," *Science,* 270, 467- 470.
16. Slack, A. T., Dohnt, M. F., Symonds, M. L. and Smythe, L. D. (2005), "Development of a Multiple-Locus Variable Number of Tandem Repeat Analysis (MLVA) for Leptospira Interrogans and its Application to Leptospira Interrogans Serovar Australis Isolates from Far North Queensland, Australia," *Annals of Clinical Microbiology and Antimicrobials*, 4, 10.
17. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrosky, E., Lander, E. S. and Golub, T. R. (1999), "Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation," *Proceedings of the National Academy of Sciences of the United States of America*, 6, 2907-2912.
18. Velculescu, V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W. (1995), "Serial Analysis of Gene Expressions," *Science,* 270, 484-487.
19. Yeung, K. Y. and Ruzzo, W. L. (2001), "Principal Component Analysis for Clustering Gene Expression Data," *Bioinformatics*, 17, 763-774.