APPLICATION OF CLUSTERING METHODS TO SPORULATION YEAST MICROARRAY DATA

Diego De Pablo

depablodiego@uma.es
Health Engineering. Málaga University.

This work presents a comparison between three main clustering methods applied to yeast sporulation data: K-means, hierarchical clustering and self-organizing maps (SOM). K-means was used to cluster genes based on expression patterns, selecting the optimal number of clusters using the elbow method. Hierarchical clustering allowed to analyze the structure of the data without predefining a number of clusters, using a dendrogram to identify significant groups. Finally, SOM was applied as an unsupervised clustering technique, with a visual representation reflecting the topology of the data. This comparison aims to evaluate the effectiveness of each method in gene clustering, similar to the work Comparisons and validation of statistical clustering techniques for microarray gene expression data

1 Introduction

The aim of this study is to apply clustering techniques to a DNA microarray dataset of *Saccharomyces cerevisiae* gene expression during sporulation, and compare the results with those from a separate analysis.

Temporal Patterns in Gene Expression During Sporulation

Gene expression during sporulation follows distinct temporal patterns, observe the figure 1, reflecting specific cellular events [1]. These include:

- **Metabolic Early:** Rapid induction at t0.
- **Early I and II:** Sustained expression from t0.5 to t2.
- **Early-Middle:** Peak expression around t5.
- **Middle:** Activation between t5 and t7, related to meiosis.
- **Mid-Late:** Increased expression from t7 to t9, linked to spore wall formation.
- **Late:** Induction between t9 and t11.5, associated with spore maturation.

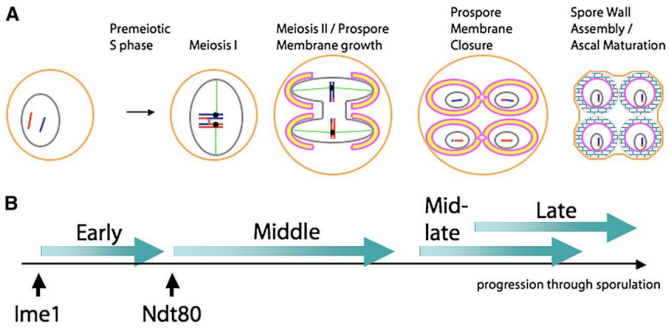

Fig. 1. A representation of the process of sporulation in budding yeasts

By utilizing DNA microarrays encompassing a significant portion of key genes, we can comprehensively investigate temporal gene expression patterns throughout the process of meiotic spore formation. However, this technique generates large datasets of unlabeled gene expression data, making pattern identification challenging. The abundance of expression levels for hundreds of genes at different time points presents an ideal scenario for applying clustering algorithms to uncover meaningful gene groups and understand the underlying transcriptional regulatory mechanisms.[2]

Application of Clustering Techniques to Analyze Sporulation Data

While hierarchical clustering (UPGMA) with correlation distance has been a popular choice in microarray studies (at least the begginnings of this century), it's important to recognize the diverse range of clustering algorithms available in pattern recognition and statistics. To classify genes based on their temporal expression profiles during yeast sporulation, we will employ various clustering algorithms, including hierarchical clustering, K-means, self-organizing maps (SOM), and Diana. This comparative analysis aims to identify distinct gene expression patterns and gain insights into the underlying transcriptional regulatory mechanisms.[2]

2 description of the methods

There is a wide variety of clustering techniques, in this work it was decided to focus on details that were not carried out in the paper Comparisons and validation of statistical clustering techniques for microarray gene expression data by the Datta brothers, giving priority to the following:

2.1 clustering techniques

- **Hierarchical Clustering with Correlation:** Hierarchical clustering with correlation is a method that groups data into a hierarchical structure rather than assigning a fixed number of clusters beforehand. It starts with each data point as a separate cluster and gradually merges the closest clusters until a single cluster remains.[3]
  - **Algorithm:** The "average" method is used to calculate the distance between clusters. This method computes the average distance between points in one cluster and points in the other cluster.[2]
  - **Common Method:** This approach, known as UPGMA, is a popular and straightforward method for hierarchical clustering.[2]
  - **Distance Metric:** The distance between genes is calculated using a correlation-based measure, where a higher correlation indicates greater similarity between the gene expression profiles.[2]
- **K-means Clustering:** Clustering is an unsupervised machine learning technique used to divide a dataset into distinct groups or clusters. Each cluster consists of data points that are more similar to each other than to those in other clusters. K-means is a popular clustering algorithm that partitions data into a predefined number of clusters, represented by centroids. The algorithm works iteratively to assign data points to the nearest centroid, updating centroids based on the mean of the points assigned to each cluster.[4]
  1. **Initialization:** k points are randomly selected from the dataset as initial centroids of the k clusters. These centroids represent the center of each cluster.
  2. **Assigning Points to Clusters:** Each data point is assigned to the cluster whose centroid is closest. The distance is usually calculated using the Euclidean distance.
  3. **Updating Centroids:** The positions of the centroids are recalculated as the average of all the points assigned to each cluster.
  4. **Repetition:** Steps 2 and 3 are repeated until the centroids no longer move significantly or a maximum number of iterations is reached.[4]
- **Self-Organizing Maps (SOM):** is an unsupervised neural network that learns to map high-dimensional data onto a low-dimensional grid. It identifies representative prototype vectors and establishes a continuous mapping from the input space to this grid. The grid, often visualized as a 2D map (see Figure 2), consists of neurons with associated weight vectors. These weight vectors are initially random but converge to represent clusters of similar data points during training.

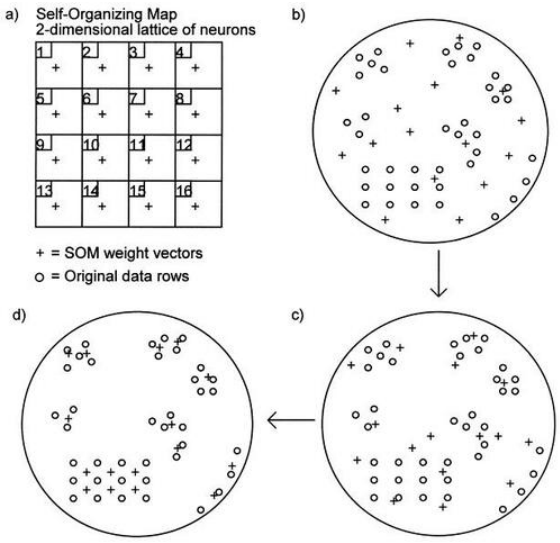

Fig. 2. The image illustrates the Self-Organizing Map (SOM) learning process. Panel (a) shows the initial SOM structure, (b) depicts the random initialization of weight vectors, (c) represents an intermediate stage of learning, and (d) shows the final configuration where weight vectors cluster around data points.

2.2 Additional Clustering Methods which are theoretically mentioned

- **DIANA (Divisive Analysis Clustering):** This method differs from hierarchical clustering in that it starts with all data in a single cluster and progressively divides it into sub-clusters.[2]
- **Fanny:** Utilizes fuzzy logic to generate a probability vector for each observation, assigning observations to clusters based on the highest probability. L1 distance (Manhattan distance) is typically used as the dissimilarity measure, offering robustness compared to Euclidean distance. [2]
- **Model-based Clustering:** Treats the data as arising from a mixture of distributions, allowing for a probabilistic interpretation of clusters.[2]
- **Hierarchical Clustering with Partial Least Squares:** Leverages partial least squares to identify gene relationships through their expression profiles, demonstrating its effectiveness as noted by Datta (2001)[2].

2.3 Validation Methods

To validate the obtained clusters, we can employ various strategies. One approach is to experiment with different numbers of clusters and observe how the average distance between points and their respective cluster centroids changes. This is known as the *elbow method* due to the characteristic elbow shape of the resulting graph. However, the elbow method alone does not guarantee optimal cluster formation.

To address this, we can utilize *silhouette analysis*, a cluster validation technique based on the silhouette coefficient. The silhouette coefficient measures the similarity of a data point to its own cluster compared to its similarity to neighboring clusters. It ranges from -1 to 1, with values closer to 1 indicating better cluster membership and values closer to -1 suggesting misclassification.

The silhouette coefficient is calculated as follows:

- a: The average distance between a point and all other points in its cluster.
- b: The average distance between a point and all points in the nearest different cluster.

$$\text{Silhouette Coefficient} = \frac{(b-a)}{\max(a,b)}$$

By calculating the average silhouette coefficient for all points within each cluster, we can assess the overall cluster validity. Combining silhouette analysis with the elbow method can help us determine the optimal number of clusters that minimize the overall distance between points and their clusters while maximizing the silhouette coefficient, suggesting well-formed and meaningful clusters.

3 most relevant results obtained comparing both methods

Bold paragraph. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

4 Conclusions

Bold paragraph. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

References

1. S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, and I. Herskowitz. The transcriptional program of sporulation in budding yeast. *Science*, 282(5389):699–705, 1998.
2. Susmita Datta and Somnath Datta. Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19(4):459–466, 2003.
3. Michael J. Guess and Scott B. Wilson. Introduction to hierarchical clustering. *Journal of Clinical Neurophysiology*, 19(2):144–151, March 2002.
4. Douglas Steinley. K-means clustering: a half-century synthesis. *The British Journal of Mathematical and Statistical Psychology*, 59(Pt 1):1–34, 2006.