Lab 1. Clustering

Application of clustering methods to sporulation yeast microarray data

Alberto Montero Solera and Teresa Vega Martínez

13albertomontero@uma.es teresavegamar@uma.es
Ingeniería de la Salud. Universidad de Málaga.

1 Introduction

The analysis of gene expression data is a fundamental tool that allows us to understand complex biological processes. One such process is sporulation in yeast (*Saccharomyces cerevisiae*), a key cellular differentiation event in response to stress conditions. During this process, gene expression levels vary over time, enabling the identification of patterns associated with different phases of sporulation.

Clustering gene expression data is useful for identifying groups of genes that may be functionally related or co-regulated, providing a deeper understanding of the mechanisms underlying biological processes.

In this project, we will apply clustering methods to microarray data obtained during yeast sporulation to group genes with similar behaviors over time.

1.1 Yeast reproduction

Saccharomyces cerevisiae is a species of yeast that serves as a widely used model organism in both industry and research. This single-celled fungus can reproduce either asexually or sexually.

Yeasts alternate between two forms in their life cycle, one haploid and one diploid, and reproduce asexually through a process known as budding.

Under environmentally stressful conditions, yeasts can reproduce sexually, which is important as it introduces genetic variation into a population. This type of reproduction forces diploid yeast cells to enter the sporulation pathway, involving significant changes in gene expression and cellular structure.

The diploid cell undergoes **Meiosis I**, where homologous chromosomes pair up and are separated. Each haploid cell then undergoes a second division in which the sister chromatids are separated (**Meiosis II**).

Following meiosis, each haploid nucleus is enclosed by a membrane, forming a structure called an **ascospore**. Once the protective wall is created, the mother cell remains intact and can produce multiple ascospores, typically four.

Finally, the spores undergo maturation, during which they become metabolically dormant. This allows the spores to survive without nutrients for extended periods until favorable conditions for germination arise.

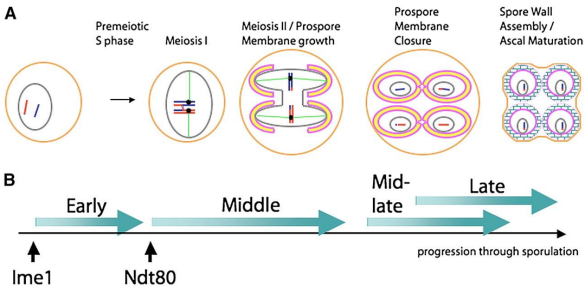

Fig. 1. Sporulation process in budding yeast

1.2 Gene expression during sporulation

Gene expression during sporulation in budding yeast is highly orchestrated, involving the sequential activation and repression of specific gene sets across different stages. These stages include distinct transcriptional phases: **early**, **middle**, **mid-late**, and **late**.

Early genes are involved in the initiation of meiosis, chromosome pairing, and recombination. These genes are regulated by specific transcription factors.

Middle genes are essential for meiotic divisions and spore formation. **Mid-late genes** play a role in spore wall formation. Finally, **late genes** are crucial for spore maturation, although the factors regulating these later stages are not fully understood. Around 150 genes have been identified as participants in the sporulation process.

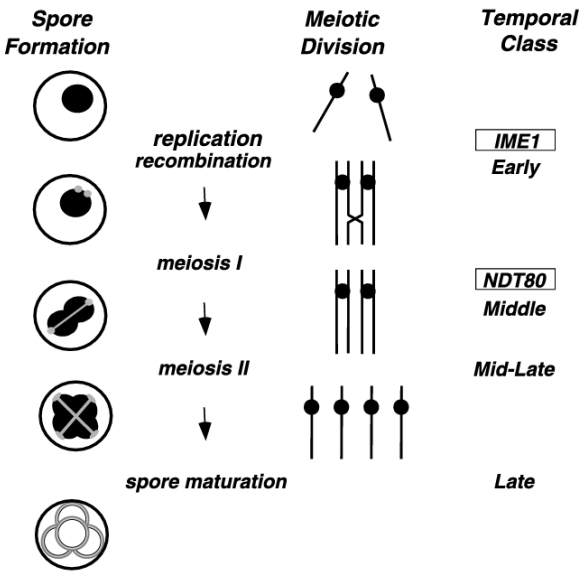

Fig. 2. Yeast reproduction

2 Algorithm and description

Different clustering methods were implemented to identify gene expression patterns in a microarray dataset collected during sporulation in yeast. The main goal of using multiple clustering algorithms was to evaluate and compare their ability to group genes with similar expression profiles over time. These patterns help reveal how genes are coordinated at various stages of the sporulation process. However, in this study, only the **k-means algorithm** will be applied and studied.

2.1 K-means Clustering

K-means is a partition-based clustering method where the number of clusters (K) is predefined, often initialized randomly, and the clusters are represented by centroids. The algorithm iteratively assigns each data point to one of the K clusters based on the mean expression profile of the cluster, aiming to minimize the within-cluster variance. It is fast and effective for large datasets, but it can be sensitive to the initial selection of cluster centers and may not always converge to the global optimum.

Expectation–Maximization The Expectation-Maximization (E-M) algorithm is a widely used technique in data science, applied to various tasks, including clustering. One of its simplest and most intuitive applications is in the k-means algorithm, which operates through an iterative process. The algorithm starts with an initial guess for the cluster centers and then follows a two-step procedure that is repeated until convergence:

- E-Step (Expectation Step): Assign each data point to the nearest cluster center based on current estimates.
- M-Step (Maximization Step): Recalculate the cluster centers by computing the mean of the data points in each cluster.

The E-step is focused on updating the expectation of which cluster each point belongs to, while the M-step maximizes the fitness of the cluster centers by adjusting their positions based on the mean. Although the E-M algorithm has been extensively studied in the literature, it is well established that each iteration typically improves the estimates of the cluster characteristics, leading to better convergence over time.

Once this is known, we could summarize the steps considered in an iteration during the execution of the algorithm as:

1. Calculate the distance from each point to every centroid.
2. Assign each point to the centroid with the shortest distance.
3. Recompute the centroids as the arithmetic mean of all the points belonging to the same cluster.

3 Cluster selection: Comparisons of Methods

3.1 Authors' Method for Selecting Clusters

In the paper, the authors implemented the K-means algorithm without using the elbow method or any

other systematic approach to determine the optimal number of clusters. Instead, they pre-defined the number of clusters based on prior biological knowledge. Specifically, the number of clusters, $K = 7$, was chosen following the study by Chu et al. (1998), which identified seven distinct temporal classes for gene expression during the sporulation process in budding yeast.

This approach, while grounded in biological reasoning, limits flexibility and adaptability. By fixing the number of clusters at 7, the authors assumed it was optimal without further validation. Although this worked in their specific study, it ignored the possibility that a different number of clusters could yield better results from the data. Thus, their clustering solution may not have been the best fit for the data structure.

3.2 Elbow Method for Cluster Selection

In contrast, our approach will utilize the **elbow method** to determine the optimal number of clusters. The elbow method is a data-driven technique that systematically tests a range of values for K and evaluates the fit of the clustering model. Instead of assuming that $K = 7$, we will run the K-means algorithm for a range of K values, typically from 1 to 10 or higher, and evaluate the clustering quality.

For each value of K, the *distortion* (i.e., sum of squared distances from each point to its assigned center) will be computed. The goal of K-means is to minimize this distortion. A plot of distortion versus K will be generated, where the x-axis represents the number of clusters, and the y-axis represents the distortion.

After running the K-means algorithm for a range of K values, we will select the optimal K based on the elbow point. This will ensure that our clustering solution is not only biologically meaningful but also validated by the structure of the data itself.

To summarize, the method implemented by the authors assumes the number of clusters to be 7 based on prior biological knowledge. In contrast, we apply the elbow method to determine the optimal number of clusters by running the K-means algorithm, analyzing the distortion, and identifying the best value for K.

Before applying this method and starting to work with the data, it is necessary to ensure that it is well prepared. On occasion, Clustering algorithms, such as k-means, can have problems when applied to high-dimensional data, for this reason we are going to apply **PCA** to the data. This technique **Principal Component Analys** is used to reduce the dimensionality of a data set, maintaining as much variability as possible. In this way it transforms the original data into a new coordinate system, where the axes are linear combinations of the original characteristics. The goal is for the first principal components to retain most of the information (variance) of the data, allowing the representation of high-dimensional data in fewer dimensions.

When applying this method we have considered a cluster range from 2 to 10, that is, the method will test from 2 to 10 clusters and return the best result. In Figure 1 we can see a point where the rate of decline begins to go down, the "elbow." This point tells us the optimal number of clusters is 4 due to the fit time given with it and the low distorsion score obtained.

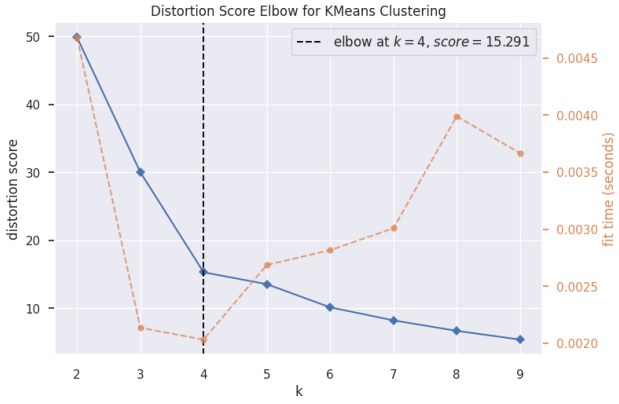

Fig. 3. Disortion Score Elbow for K-Means Clustering

3.3 Cluster Validation

Finally, we have performed the clustering with kmeans considering 4 clusters and to evaluate its performance we have used the silhouette metric, which will allow us to see how well defined the clusters are. A higher value (closer to 1) indicates that the points are well grouped into their respective clusters, while a value close to 0 suggests that the points might be incorrectly grouped.

In Figure 4 we can see a silhouette score graph to be able to visually observe the result.

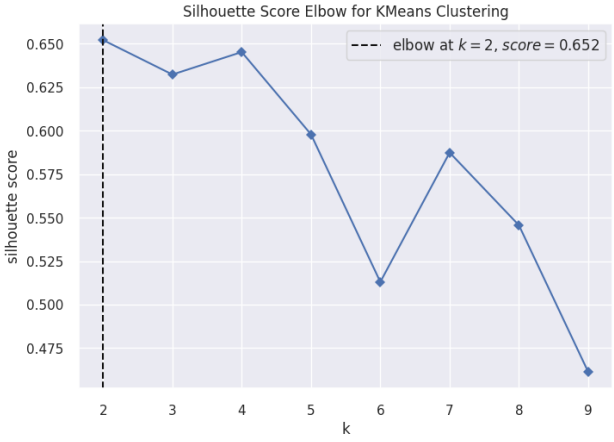

Fig. 4. Silhouette Score

In Figure 4 the elbow appears to be at k=2 with a score of 0.652. This suggests that 2 clusters is where there is a significant improvement in clustering, and then the improvement is less pronounced as you increase the number of clusters.

For k=4 (the number of clusters you are testing in the main model), the silhouette score appears to be slightly lower than the score of k=2, so according to this analysis k=4 would not be the optimal number of clusters, but it is a good option too.

3.4 K-Means implementation

Once the optimal value of k is obtained, we will run the K-Means algorithm to group the sporulation data. To visualize the results we have created the scatter plot in Figure 5.

In figure 5 we can see the 2 resulting clusters in two different colors,yellow and purple. Both clusters are densely grouped on the left and right respectively and they are clearly differentiated .

About the **centroids** , they have been calculated by the k-means algorithm and are represented by the 2 large red dots. At this point it is necessary to highlight that the centroids are the "geometric centers" of each cluster, and the objective of K-means is to

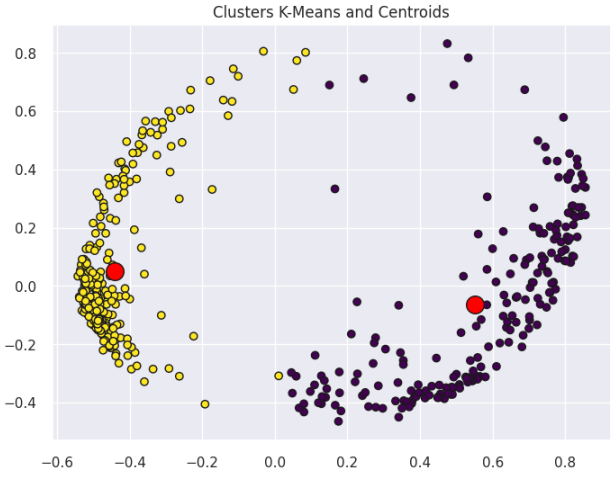

Fig. 5. Dispersion with k=2

minimize the distance of the points to their centroids.

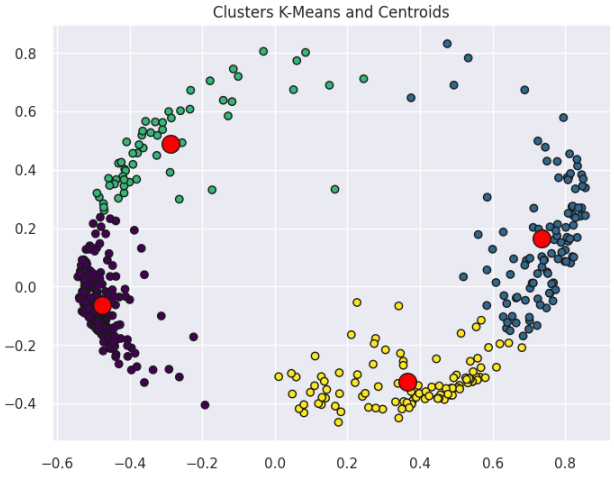

Fig. 6. Dispersion with k=4

The optimal value for k that we have obtained is k=2. Despite this, we have considered it interesting to represent the same scatter plot considering k=4 since it is the optimal value that we obtained in the beginning and the silhouette score is really close to the k=2. So, the Figure 6 is a scatter plot of the classification for 4 clusters.

This graph shows a more detailed division, but it could be artificially separating a single large group into subgroups, which does not completely follow the

natural structure of the data. The data seem to fit best for k=2, as expected, since the curve of the points suggests two natural groupings. Segmentation is more intuitive in this case.

4 Conclusion and comparison

The comparison between the seven initial clusters selected by the researchers in *Comparison and validation of statistical clustering techniques for microarray gene expression data* , based on prior biological knowledge, and the clusters derived from our implementation of the elbow method and validation using the silhouette score, reveals important insights into the robustness and coherence of the clustering outcomes.

The researchers' decision to predefine seven clusters grounded in biological principles suggests that these clusters may align with genuine biological patterns in gene expression during sporulation in budding yeast. However, our application of the elbow method indicated that the optimal number of clusters may not be restricted to seven. By identifying an inflection point in the explained variance, we suggest that a different number of clusters could yield greater separation and biological significance.

Furthermore, the silhouette score provided a measure of the quality of the formed clusters, allowing us to assess both the cohesion and separation of the generated groupings. While the initial seven clusters demonstrated an acceptable level of quality, our analysis indicates that more complex structures may exist within the data that were not captured by the initial selection. This raises the possibility that an alternative number of clusters could uncover biologically relevant subgroups that were overlooked in the original study.

This study was carried out more than 20 years ago and was based on the technologies of that time. Taking into account the great current technological advances, we consider that our study could be more reliable by having used different techniques to evaluate the optimal value of k before carrying out the implementation and by this way obtain better results.

In conclusion, although the predefined seven clusters are rooted in a solid biological context, our findings from the elbow method and silhouette score validation suggest that a more thorough exploration of the data may lead to a better understanding of gene expression patterns and the identification of biologically significant subgroups.

References

1. Datta, S., & Datta, S. (2003). Comparison and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19(4), 459-466.
2. Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., & Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science*, 282(5389), 699-705.
3. Knop, M. (2011). Yeast cell morphology and sexual reproduction – A short overview and some considerations. *Comptes Rendus Biologies*, 334(8), 599-606. Ten years of genomic exploration in eukaryotes: strategy and progress of Genolevures. https://doi.org/10.1016/j.crvi.2011.05.007