Lab 1: Clustering

Alejandro Silva Rodríguez

Marta Cuevas Rodríguez

Aprendizaje Computacional
Universidad de Málaga

Septiembre 2024

Índice

1. Introducción

The process of **sporulation in yeast** is a well-established model for studying cellular differentiation and gene regulation. Sporulation involves a series of highly regulated biological stages during which the yeast cell transitions into a spore, primarily in response to nutrient deprivation. Gene expression in yeast during sporulation is characterized by distinct temporal patterns, making it an ideal candidate for clustering analysis. Through clustering, genes with similar expression profiles can be grouped, aiding in the identification of genes that may participate in similar biological functions or regulatory pathways.

With the advent of **microarray technology**, the ability to measure the expression levels of thousands of genes simultaneously across different time points has significantly advanced. This vast amount of data requires effective computational tools for analysis. One such tool is **clustering**, which groups genes based on the similarity of their expression profiles. In this context, **k-Means clustering** has emerged as a widely used technique due to its simplicity and effectiveness. The algorithm attempts to partition genes into k clusters by minimizing the variance within each cluster, leading to groups of genes that exhibit similar temporal expression patterns during sporulation.

In this project, we aim to evaluate the performance of **k-Means clustering on the sporulation dataset** of budding yeast, comparing the results to those presented in [2] by Datta and Datta (2003), which explores various clustering methods including hierarchical clustering and Diana. The primary objective is to assess the effectiveness of k-Means in clustering genes during sporulation, and to analyze how it compares to more complex methods discussed in the literature.

2. Objectives

The main objective of this project is to **evaluate the performance of the k-Means clustering algorithm** when applied to the sporulation dataset of yeast, which contains gene expression profiles measured across multiple time points. By clustering these genes, the aim is to **identify groups of genes** that exhibit similar expression patterns throughout the sporulation process. To assess the effectiveness of k-Means, the results will be compared to those obtained in Datta and Datta's (2003) study [2], which evaluated various clustering techniques, including hierarchical clustering and divisive clustering (Diana). Additionally, metrics such as the silhouette score will be used to quantify the quality of the clustering results. Through this comparison, the project seeks to determine the strengths and limitations of k-Means in clustering biological data and to explore its applicability in gene expression analysis during yeast sporulation.

3. Acceso al Repositorio

Toda la información adicional, incluyendo el código fuente y la documentación completa de este proyecto, está disponible en el repositorio de GitHub [1].

Referencias

[1] Alex Silva. Practica1_almacenes_de_datos. `https://github.com/AlexSilvaa9/Practica1\_almacenes\_de\_datos`, 2024. Último acceso: 1 octubre 2024.

[2] Susmita Datta and Somnath Datta. Comparison and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19(4):459–466, 2003. `http://bioinformatics.oxfordjournals.org/content/19/4/459`, 2003.