Gene Expression Analysis of Yeast Diauxic Shift

Name: Joshua Estrada || **Date**: 1/31/2025 || **Class**: Bioinformatics Algorithms (448-01)

Introduction

Saccharomyces cerevisiae, commonly known as baker's yeast, plays a crucial role in winemaking by converting the glucose found in fruit into ethanol through fermentation. When the supply of glucose is depleted, however, *S. cerevisiae* must adapt its metabolism to utilize the ethanol it has produced as a new food source. This metabolic inversion, known as the diauxic shift, is a complex process involving changes in the expression of many genes. The diauxic shift is not only essential for the survival of the yeast in environments lacking glucose, but it also has significant practical implications. For instance, if winemakers do not seal their barrels, the yeast in the barrel may begin metabolizing the ethanol it produced, thereby altering the quality of the wine.

In 1997, Joseph DeRisi conducted a landmark gene expression experiment by sampling an *S. cerevisiae* culture every two hours for six hours both before and after the diauxic shift. The resulting 6,400 × 7 gene expression matrix provides a detailed view of transcriptional changes during this critical metabolic transition.

Data and Preprocessing

The raw gene expression data for this experiment is provided in the file “diauxic_raw_ratios.txt”, which contains expression ratios for approximately 6,400 genes across seven time points (R1 through R7). Notably, the data is organized in a gene-by-time-point format, differing from the gene-by-cell format more commonly encountered.

Prior to analysis, the data format was verified. The file was found to contain raw expression ratios rather than log-transformed values. For consistency in scaling and to facilitate downstream statistical analyses, the raw ratios were converted to log2 space using the transformation:

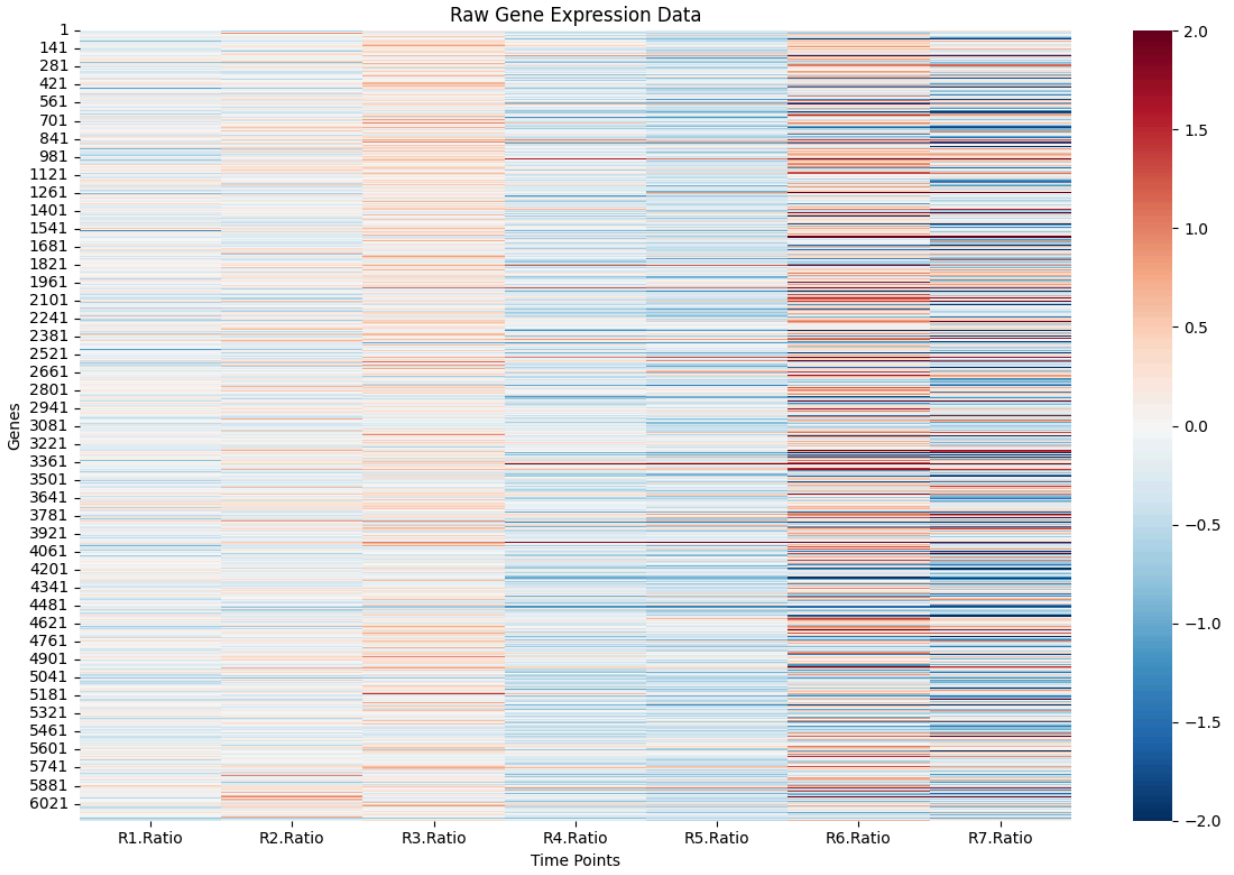
$$\text{log2\_value} = \log_2(\text{data.clip}(loIr = 1e - 10))$$

This conversion ensures that the data better meets the assumptions of many statistical methods used in expression analysis.

Heatmap Analysis

The first analytical step was to create a heatmap of the raw gene expression data. The heatmap provides a high-level visualization of the expression patterns across all genes and time points,

revealing both up-regulation (indicated by red) and down-regulation (indicated by blue) throughout the time course. This visualization serves as a preliminary check to detect any systematic patterns or anomalies in the data.



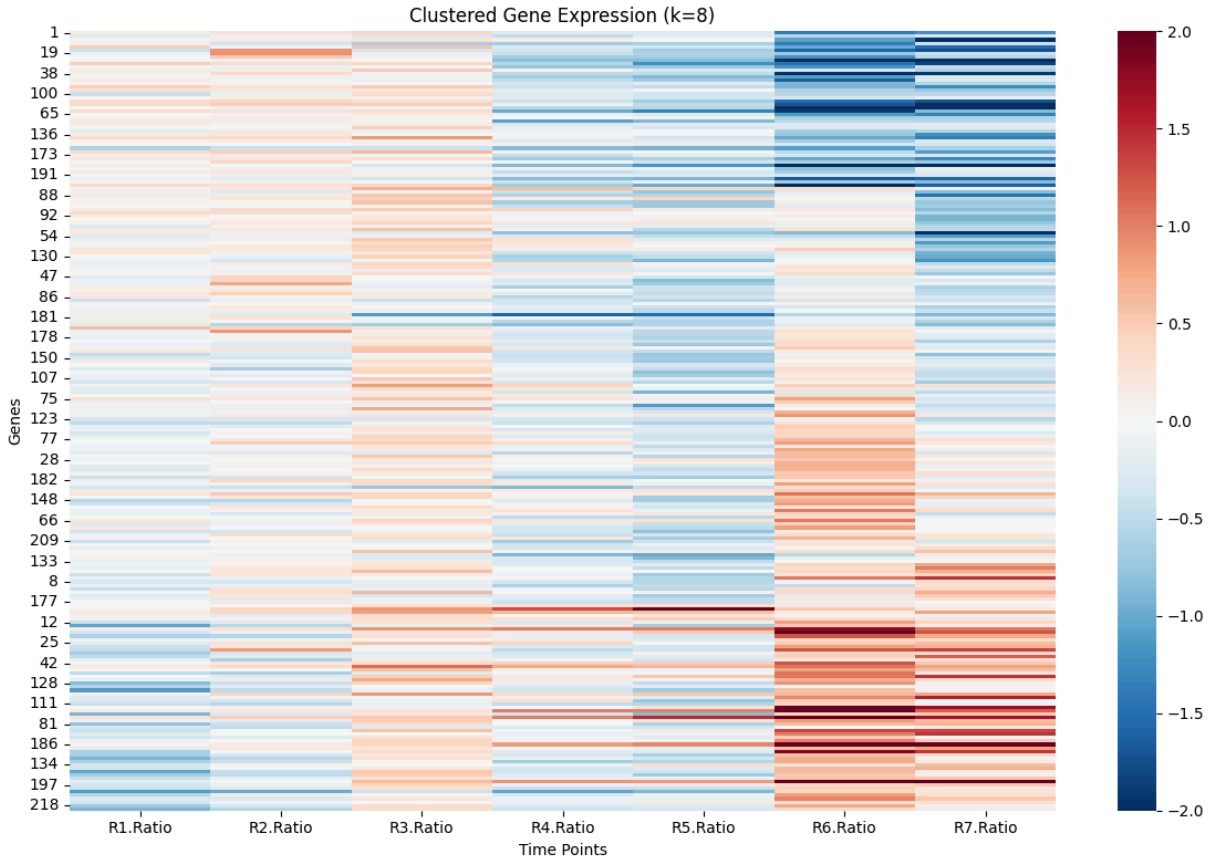
Clustering Analysis

Methodology

Hierarchical clustering was applied to the log-transformed gene expression data. The clustering employed the 'ward' linkage method with Pearson correlation as the distance metric. Several clustering solutions were explored, and eight clusters were ultimately selected based on a silhouette score of 0.193. The silhouette score, although relatively low, provides a quantitative measure of cluster separation, with a higher value suggesting more clearly defined clusters.

Results and Interpretation

The clustered heatmap reveals several interesting patterns. Groups of genes exhibit similar expression profiles, with some genes showing coordinated up-regulation or down-regulation around the diauxic shift (between time points R3 and R5). The modest silhouette score suggests that the clusters do not represent sharply delineated groups, indicating that the transition may be more continuous or gradual than discrete. This observation reflects the biological complexity of the diauxic shift, where multiple genes may change their expression levels gradually as the yeast adapts to the new metabolic state.



Identification of Variable Genes

Custom Metric Development

To concentrate on genes most relevant to the diauxic shift, a custom metric was devised to identify the 230 most variable genes. The metric was designed to capture both the magnitude of expression changes and the overall variability in gene expression across the time points. Specifically, the absolute log2 fold change between the mean pre-shift (R1–R3) and post-shift (R5–R7) expression levels for each gene was computed. This value was then multiplied by the square root of the gene's total variance across all time points:

$$\text{Metric}_i = \left| \log_2 \left(\frac{\mu_{\text{post-shift},i}}{\mu_{\text{pre-shift},i}} \right) \right| \times \sqrt{\text{Var}_i}$$

This metric is reasonable because it simultaneously considers the scale of change (via the fold change) and the variability (via the variance), ensuring that genes with both dramatic and consistent changes in expression are prioritized. One potential pitfall of this approach is that it may overweight genes with high variance due to noise rather than true biological signal; therefore, further filtering or biological validation may be necessary.

Comparison with the Authors' Selection

The list of 230 most variable genes was compared with the set provided by the original study (found in "230genes_log_expression.txt"). The overlap between the two lists was 11 genes, corresponding to a Jaccard coefficient of 0.024. This low overlap indicates that the custom metric and the authors' criteria are substantially different, suggesting that further refinement of the metric might be necessary to better capture the biologically relevant genes.

Clustering on the Reduced Gene Set

The clustering analysis was repeated using only the 230 genes selected by the original authors. This resulted in a slightly improved silhouette score of 0.229. Despite this marginal improvement, the resulting clusters did not clearly segregate into groups corresponding to distinct stages of the diauxic shift. This outcome suggests that while focusing on the most variable genes reduces noise, the underlying biological changes during the diauxic shift might still be gradual and overlapping.

Discussion and Future Improvements

Interpretation of Results

The analysis of gene expression during the diauxic shift in *S. cerevisiae* offers valuable insights into the transcriptional reprogramming that occurs during this metabolic transition. Although hierarchical clustering did not yield sharply separated groups of genes, the patterns observed do reflect the complex and continuous nature of the diauxic shift. The identification of variable genes using a custom metric provides a starting point for further analysis, even though the overlap with the authors' gene set is limited.

Areas for Improvement

1. **Methodological Clarity**:

    - The custom metric for assessing gene variability has been described in detail, but future work should explore alternative metrics or additional filtering methods to better align with known biological functions and reduce the impact of noise.
2. **Biological Relevance**:
    - Future analyses should explicitly investigate the biological roles of the identified genes. For example, correlating the expression profiles with known pathways involved in glucose and ethanol metabolism would help verify the functional significance of the results.
3. **Visualization Enhancements**:
    - Including well-annotated figures (heatmaps and clustered matrices) with clear captions in the PDF will strengthen the visual communication of the findings.
4. **Temporal Dynamics**:
    - Incorporating a more nuanced analysis of temporal expression patterns could identify subgroups of genes with distinct timing in up- or down-regulation, thereby offering clearer insights into the progression of the diauxic shift.

Future Directions

- **Algorithm Exploration**: Experiment with additional clustering algorithms and distance metrics (e.g., k-means, DBSCAN, or model-based clustering) to determine if more clearly defined clusters can be identified.
- **Gene Selection Refinement**: Investigate alternative variable gene selection methods that incorporate prior biological knowledge, such as pathway databases, to better capture genes relevant to the diauxic shift.
- **Functional Enrichment Analysis**: Perform functional enrichment analysis on the selected gene sets to determine if specific biological processes or pathways are overrepresented, providing further validation of the results.
- **Temporal Modeling**: Develop methods that explicitly incorporate the time-series nature of the data, such as dynamic time warping or state-space models, to better capture the progression of gene expression changes during the diauxic shift.

By integrating these improvements and further aligning the analysis with the underlying biological context, a more comprehensive understanding of the genes and expression dynamics driving the diauxic shift in *S. cerevisiae* can be achieved.