Clustering Methods to Analyze Gene Expression Data

Kyle Houfek, Nataša Jonoska, Masahico Saito

4202 E. Fowler Ave Tampa, FL 33620, USA

Introduction

RNA sequencing [1] is a method used to determine the quantity of RNA in a cell at a given point in time. By analyzing the changing quantities of RNA sequences in the cell over a period of time, we can estimate how gene expression rates change over time. Analysis of these expression rates, specifically those of mRNA, can help us determine protein function at various time points. Our data set comes from an ciliate species known as *Oxytricha trifallax*, with expression rates from more than 20,000 genes read across six different time points. This data allow us to cluster genes together based on their expression levels at each time point. By considering genes with high expression rates at a certain time point, we can understand the biological processes at that time point.

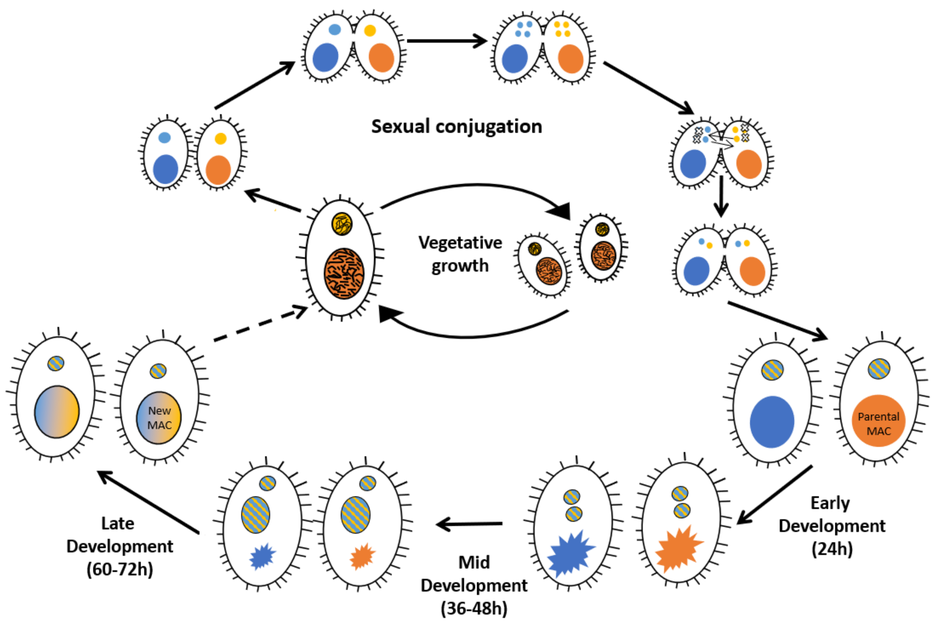


Figure 1: Lifecycle of *O. trifallax* during and after sexual reproduction. Ladweber Lab, Columbia University, 2019.

Clustering

We analyze the data by taking each gene as a vector, with the expression rate at each of six time points as an entry of said vector. These vectors are clustered together based on their expression rates at a given time. For example, those genes with high expression rates at 12 hours but low expression rates at all other times will be in a single cluster. Our primary clustering method used was k-means clustering [2], which first creates k random centers in the domain of our vector space, then assigns each vector to the cluster whose center is the closest. After all vectors have been assigned, a new center is calculated by taking the average of the expression rates at each time point for all vectors in the cluster. The process is repeated until it reaches a set number of iterations or no change in the centers is obtained.

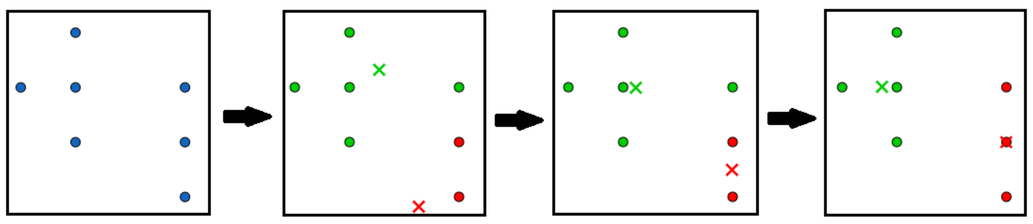
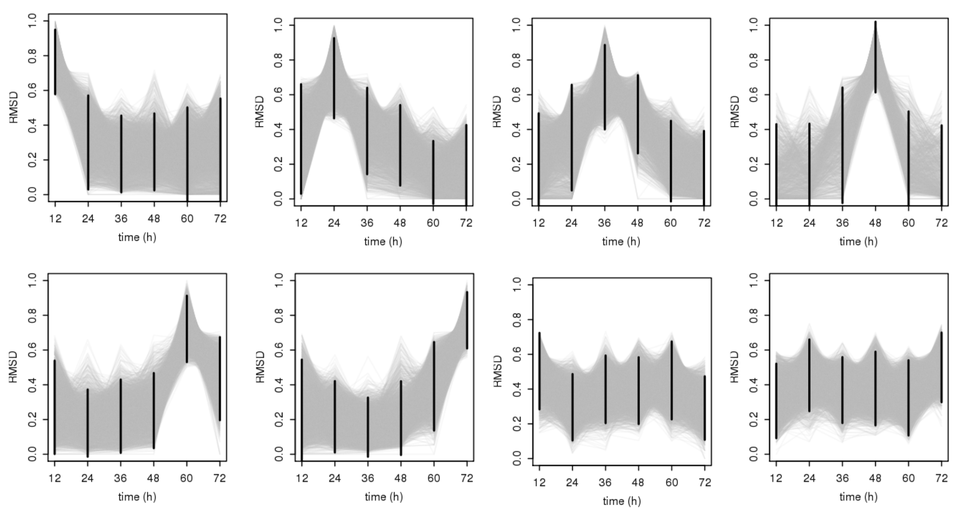


Figure 2: (1) Given the vectors, (2) random points in domain our selected as centers, and each vector is assigned to its closest center. (3) Centers are recalculated, and (4) vectors are reassigned and centers recalculated again.

We used k-means clustering on the data to form 8 clusters. Based on the expression patterns below, this gives 6 clusters which peak at one time point, and 2 clusters for genes with no peaks. We clustered vectors for the mean, variance, and index of dispersion (ratio of variance to mean) values using three replicates in our data. Each value has distinct biological significance: for example, mean measures how many resources are allocated to a gene at a given time, while variance measures how varied that gene's production is allowed to be at that time.



Gene Set Enrichment Analysis

Gene Set Enrichment Analysis [3] provides a method to test whether or not the clusters are statistically significant. Consider the example set of data below:

	t1	t2
Gene A	7	2
Gene B	1	5
Gene C	3	5
Gene D	4	1

→

	t1	t2
Gene A	7	2
Gene D	4	1

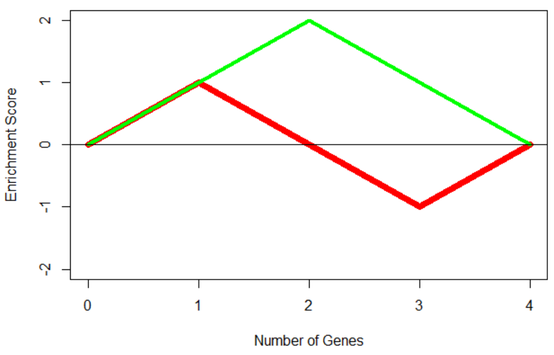
	t1	t2
Gene B	1	5
Gene C	3	5

For each cluster, determine where the expression rates peak and where they do not. For example, the expression rates for the genes in the top cluster peak at t1, but not at t2. A partial order is then created based on the ratio $m = t1/t2$, and we say for our genes that $G_1 \leq G_2$ if $m_1 \leq m_2$, where G_1, G_2 are two genes with m_1, m_2 their respective ratio. The list is sorted below based on this ordering:

	t1	t2
Gene A	7	2
Gene B	1	5
Gene C	3	5
Gene D	4	1

→

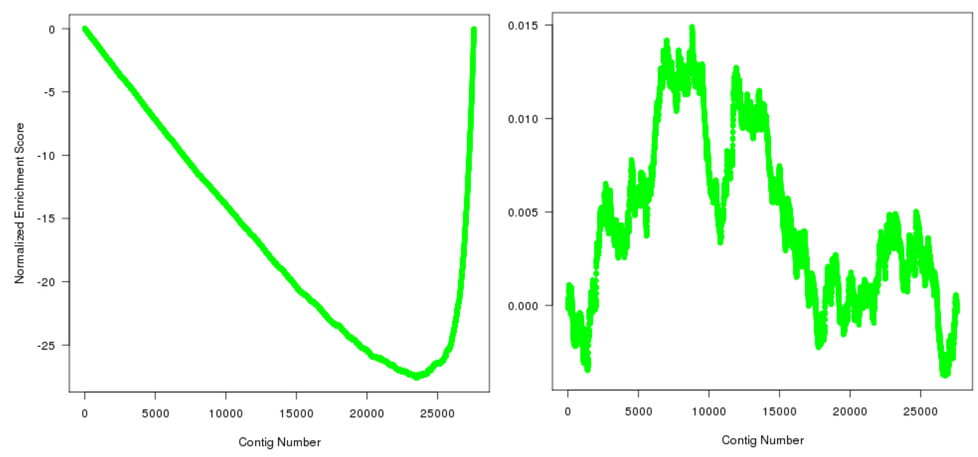
	t1	t2	m
Gene D	4	1	4
Gene A	7	2	3.5
Gene C	3	5	0.6
Gene B	1	5	0.2



It is expected that this method will sort the genes in the cluster to either the top or bottom of the list. Next, on the n genes in the data we define a recursive function $f_C : [0, n] \to \mathbb{Z}$ for the cluster C as follows: set $f_C(0) = 0$, and say

$$f_C(i+1) = \begin{cases} f_C(i) + 1 & \text{if } x_i \in C \\ f_C(i) - 1 & \text{if } x_i \notin C \end{cases}$$

We plot this function for each cluster as above (green), as well as for a random cluster (red). Using results from the actual data, it is clear that a real cluster and a random cluster have very different graphs.



The largest deviation from 0 on the graphs above is the enrichment score (ES), and is the primary measurement used to determine statistical significance of a cluster. Using GSEA on our eight clusters to the left, we could determine that these clusters could not have been randomly generated, and are thus viable to use for future analysis.

Future Work

Our ultimate goal of this project is to extract useful information, such as protein function, from a given RNA sequence data set. Using the above clustering methods, we can find protein families performing similar biological functions at certain time points, and from this infer the actual biological processes.

References

[1] Swart, E. C. et al. (2013) *The Oxytricha trifallax Macronuclear Genome: A Complex Eukaryotic Genome with 16,000Tiny Chromosomes*, PLoS Biol (11)1.

[2] Lloyd, S. P. (1982) "Least square quantization in PCM". *Bell Telephone Laboratories Paper.*

[3] Lander, ES, et al. (2005) *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles*. PNAS 102(43):15545-15550.