CS284A: Algorithms for Molecular Biology Assignment #2

In the last assignment, we worked on discovering motifs in the yeast species *Saccharomyces cerevisiae*. In the previous exercise, we have provided you a list of genes that share similar expression patterns. In this exercise, we will go to some of details to understand how this list of genes was derived. In particular, we will work on an algorithm to group genes into different clusters.

We will use the yeast cell cycle data gathered by Cho *et al.* (*Mol. Cell* 2:65-73, 1998), who used Affymetrix oligonucleotide microarrys to query the abundances of almost all yeast mRNA species in synchronized *Saccharomyces cerevisiae* batch cultures. The data provides us the measurement on the abundances of 6565 mRNA species with 15 time points, across two cell cycles.

1. Go to the course website:
   `http://www.ics.uci.edu/xhx/courses/CS284A/assignments/PS2/`
   Download the data: `cho_cell_cycle_ex90_100_data.tsv`

   Format of the file: each row represents one mRNA, starting with the name of the mRNA, followed by its abundances at 15 time points.

   Let X_{ij} denote the expression of the i^{th} mRNA at the time point j. Here $i = 1, \cdots, 6565$ and $j = 1, \cdots, 15$.

2. Data normalization. Write a program to normalize the expression of each mRNA across 15 time points such that the mean of its expression values across 15 points is 0 and the variance is 1. This can be done through the following three steps. For each mRNA, say the i^{th} mRNA,

   (a) Calculate the mean (μ_i) of the mRNA across 15 time points.
   (b) Calculate the standard deviation (σ_i) of the mRNA across 15 time points.
   (c) Normalize the data using the following formula:

   $$X'_{ij} = \frac{X_{ij} - \mu_i}{\sigma_i} \tag{1}$$

   Prove that the above procedure indeed leads to mean 0 and variance 1 for the normalized data X'.

3. Write a K-means algorithm to cluster 6565 mRNAs into $K = 20$ groups. Use Euclidean distance based on the normalized data to measure the distance between two mRNAs, that is, the distance between the m^{th} and the n^{th} mRNA is

   $$d(X_m, X_n) = \sum_{j=1}^{15} (X'_{mj} - X'_{nj})^2 \tag{2}$$

where $X_i = (X_{i1}, X_{i2}, \cdots, X_{i15})$ represents the i^{th} mRNA.

For this assignment, you will need to submit the following items:

(a) Your source code implementing the K-means algorithm.

(b) In the class, we show that the K-means algorithm minimizes the following error function:

$$V = \sum_{i=1}^{K} \sum_{X_j \in S_i} d(X_j, C_i)^2 \tag{3}$$

where S_i is the set of points in cluster i and C_i is the mean point of all the points $X_j \in S_i$.

Plot V as a function of the number of iterations in the K-means algorithm. The curve should monotonically decrease and converge to a fixed point.

(c) Plot each of the centers C_i for $i = 1, \cdots, K$ as a function of time (15 time points).