Genomic Signal Processing and Data Science EE 381V
Prof. Haris Vikalo HW# 5
April 11, 2022 Due: 04/25/22

Homework Set #5

1. **Problem 1 [Haplotype assembly (6 points).]** Assume that a genome of a diploid organism was sequenced using a high-throughput sequencing device. Sequencing was erroneous (with error rate $p_e = 0.005$), and performed at coverage $c = 20$. After SNP calling and genotyping, data is organized in a SNP fragment matrix R (which you can download from `http://www.ece.utexas.edu/~hvikalo/ee381v/SNP_Fragment_Matrix.txt`). The dimension of R is 80×194, implying that there are $n = 80$ reads and the length of the haplotypes is $m = 194$. Relying on matrix factorization ideas, reconstruct the haplotypes and find the corresponding minimum error correction (MEC) score. Please submit your code. *Remark:* To initialize the haplotype sequence in an alternating minimization algorithm, you may want to consider singular value decomposition of the SNP fragment matrix.

2. **Problem 2 [Gene clustering (6 points).]** Data set YeastCycle.xls (which you can download from `http://www.ece.utexas.edu/~hvikalo/ee381v/YeastCycle.xls`) contains expressions of 678 genes from the yeast cell data, measured in experiments performed on yeast cell samples which were acquired every 20 minutes.

   (a) Perform the singular value decomposition of the data, i.e., factorize $X = \Theta\Sigma V^T$. Plot the entries of the first two rows of $F = \Sigma V^T$ as a function of time (there are 12 time steps) on the same plot. Are there any patterns?

   (b) Use k-means clustering to cluster the genes for $k = 3, 4, 5, 6$. Is there evidence that one of these numbers of clusters is better than another?

   (c) For the clustering with $k = 4$ clusters, plot the cluster means obtained in (b) as a function of time on a single plot. Are there any patterns?

3. **Problem 3 [Hierarchical clustering (5 points).]** Perform a hierarchical clustering (i.e., give the cluster tree) of the genes (a, b, c, d, e), where their pairwise distances are given by

	a	b	c	d	e
a	0	3	8	7	8
b	3	0	4	8	8
c	8	4	0	5	6
d	7	8	5	0	6
e	8	8	6	6	0

   Complete this task using (a) single-link, and (b) complete-link hierarchical clustering method.

4. **Problem 4 [Graph representation of genes (3 points).]** The expression of one gene may have a regulatory effect on another gene. Gene regulation can be represented by a graph where the genes are vertices in a graph, and where a directed edge conects gene A to gene B if gene A directly regulates gene B (that is, binds to its promoter region or inhibits its translation).

   (a) If given a graph described above, how can we tell if a gene directly regulates itself? This type of regulation is known as autoregulation.

   (b) How can we detect if a gene indirectly regulates itself?

   (c) How can we augment this graph to represent the extent to which one gene directly regulates another? How do we represent a gene negatively regulating another one?