CAP5510 Fall 2019, HW#2, Due Oct 23rd, 2019 (hardcopy in class submission)

Total 50 pts

1. (10 pts) Differential gene expression test is to compare gene expression values under case and control conditions for a given gene. For example, let a gene's expression in three disease samples be <90,100,110> and its expression in three corresponding control samples be <100,150,350>, then t-test can be performed to provide a p-value represent how different this gene's expression values are between disease and control cases. Please describe in detail another approach used in the literature to evaluate difference between objects. Can this approach be used to evaluate the difference of gene expression under two grouped conditions? If yes, what would be its advantages and disadvantages?

2. (20 pts) Given the following 5 aligned motif instances that are corresponding to a motif x, please refer to our lecture notes on motif finding to answer the following questions:

1	**t**	**c**	**g**	**a**	**t**	**c**	**a**	**a**
2	**t**	**c**	**a**	**a**	**t**	**c**	**g**	**a**
3	**c**	**a**	**a**	**g**	**t**	**c**	**g**	**a**
4	**t**	**c**	**c**	**t**	**a**	**c**	**a**	**a**
5	**c**	**a**	**g**	**a**	**t**	**a**	**g**	**g**

(1) (5 pts) What is the profile representation for this motif?
(2) (5 pts) What is the consensus representation for this motif? What is the score and total distance for this consensus representation?
(3) (10 pts) We have talked about two exact algorithms for motif finding. One is the motif finding algorithm that try to maximize consensus score and the other is the Median String problem to minimize the total distance. Are these two algorithms computationally equivalent, i.e., do they output the same motifs given the same set of sequences? Why?

2. (20 pts) A motif is the common pattern of the DNA binding sites bound by a transcription factor. Given a group of genes that are potentially regulated by the same transcription factors, a routine approach to identify motifs of these transcription factors is to obtain the upstream 1000 base pair sequences of these genes and run computational software tools such as MEME (http://meme-suite.org/tools/meme) to identify overrepresented sequence patterns in these sequences. These overrepresented sequence patterns are considered as the motifs of the potential common transcription factors regulating these genes.
(1) (10 pts) Is the assumption of "overrepresented sequence patterns are potential motifs" reasonable? Why? Is there any limitation for motif finding with this assumption?
(2) (10 pts) Please download the sequences in the file ex2.FASTA from the "Files\HWs" folder at webcourses. These sequences are the upstream sequences of the target genes of a transcription factor. Please apply the software tool MEME to these sequences. What are the consensus sequences of the top 3 motifs predicted? What parameters did you use for this tool?