COMP 462 – Assignment #4
Due on December 3rd 2019, 23:59

1. Suffix-array (20 points)

a) (10 points) What is the suffix array for sequence AGAGACAGATAGC ?

b) (10 points) Illustrate schematically how the search for substring AGAT would proceed.

2. Genome assembly (20 Points)

Consider the following set of short reads obtained from a (toy) genome sequencing experiment.

ACCAC, CACCT, CCACC, GCACC, GGCAC, GGGCA, TAGGG,

a) (10 points) Suppose we want to assemble the genome from which these reads were obtained. Draw the overlap graph of this set of reads, and specify the weight of the edges. Only consider overlaps of 3 bases or more.

b) (10 points) What is the sequence of the mini-genome that are the most likely to have generated this set of reads?

3. RNA-seq technologies (10 points)

a) (5 points) In the genome, some genes have sequences that are very very similar to others (e.g. there might be two genes whose sequences are 99% identical). What difficulties would that cause to the analysis of RNA-seq data? (2-3 lines for each)

b) (5 points) Give two reasons why the expression measurements obtained from RNA-seq experiments may not necessarily reflecting the abundance of proteins in the sample?

4. Class distinction (20 points)

Consider the following set of expression measurements for the estrogen receptor gene in healthy and disease patients:

Healthy (m=8) :	1.2	3.3	1.6	2.3	2.5	2.7	2.3	1.5	
Disease (n=10):	1.1	0.5	1.8	2.0	1.3	1.2	1.3	0.3	1.3	2.4

a) (5 points) Compute the t-statistic and corresponding p-value for the differential expression of the two classes of patients, using a Student t-test.

b) (10 points) A permutation test is an alternate approach to measure the significance of a statistic t_{obs}. In a nutshell, it randomly permutes the healthy/disease labels of the samples is sees how often the resulting permuted data set has a t_{rand} value greater than t_{obs}.

nSuccess = 0
for i = 1 to nRep do
    Randomly choose m of the m+n samples (without repetition) to form the randomized healthy group, and assign the rest to the randomized disease group
    t_{rand} = t-statistic (randomizedHealthy, randomizedDisease)
    if ($|t_{obs}| < |t_{rand}|$) nSuccess = nSuccess + 1

return p-value = nSuccess / rRep

Use a permutation test to estimate the p-value of the t statistic observed in your data. How many repetitions are needed to estimate the p-value to within an accuracy of plus or minus 0.01 ?

c) (5 points) Why are the p-values obtained in (a) and (b) different?

5. Class prediction (20 points)

Now, suppose that we are interested in using the same set of estrogen receptor expression measurements used in 2(a) to train a classifier that will predict whether a new sample is healthy or not, based on the expression of that estrogen receptor (here, we are restricting our attention to a classifier based on a single gene for simplicity; in general more than one gene would be considered at the same time). Our classifier would then have the following form:

If expression < t:
    predict Disease
else:
    predict Healthy

where t is the threshold that needs to be learned to best separate healthy from disease examples.

a) (10 points) What is the threshold t that leads to the minimal number of classification errors on the complete data set.

b) (10 points) Draw the Receiving-operating curve for this classifier on the complete data set.

6. Class discovery (10 points)

a) (5 points) Assume that the following similarity matrix has been computed for six genes:

S =	A	B	C	D	E	F
A	1	0.15	0.2	0.6	0.2	0.6
B		1	0.4	0.3	0.8	0.5
C			1	0.12	0.22	0.24
D				1	0.1	0.8
E					1	0.14
F						1

Draw the result of applying single-linkage clustering to the data. Your answer should be in the form of a dendrogram (a.k.a a tree).

b) (5 points) Draw the result of applying complete-linkage clustering to the data. Your answer should be in the form of a dendrogram (a.k.a a tree).