Solution to HW4 – 2019

Q1.a.
All suffixes:

$	14
C$	13
GC$	12
AGC$	11
TAGC$	10
ATAGC$	9
GATAGC$	8
AGATAGC$	7
CAGATAGC$	6
ACAGATAGC$	5
GACAGATAGC$	4
AGACAGATAGC$	3
GAGACAGATAGC$	2
AGAGACAGATAGC$	1

Sorted in lexicographical order gives:

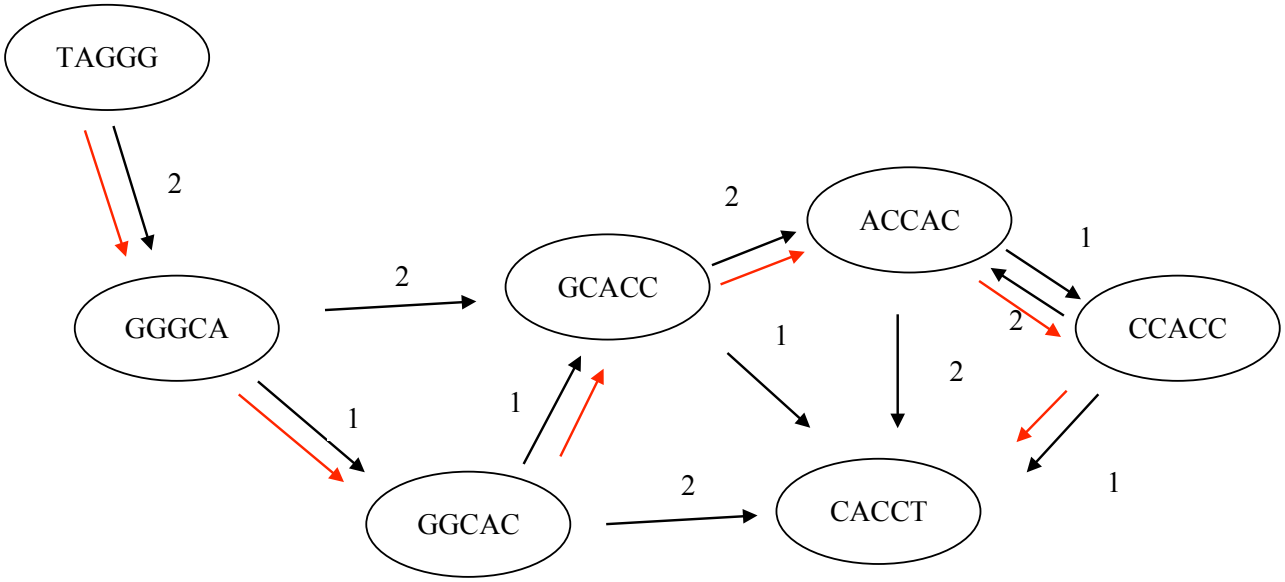
$	14
ACAGATAGC$	5
AGACAGATAGC$	3
AGAGACAGATAGC$	1
AGATAGC$	7
AGC$	11
ATAGC$	9
C$	13
CAGATAGC$	6
GACAGATAGC$	4
GAGACAGATAGC$	2
GATAGC$	8
GC$	12
TAGC$	10

b. Use binary search to locate the substring s in the suffix array S

```
Initialize l = 0, r = len(S)
while l < r:
        mid = (l + r) / 2
        if s > S[mid]:
                l = mid + 1
        elif s<S[mid]
                r = mid – 1
        else    return  mid
return None
```

Searching for AGAT, we will have:
l=0, r=13, mid=6

l=0, r=5, mid=2
l=3, r=5, mid = 4 ➔ Found AGAT
Q2.a.



b. TAGGGCACCACCT
The shortest Hamiltonian path is denoted with red arrows.

Question 3:

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)

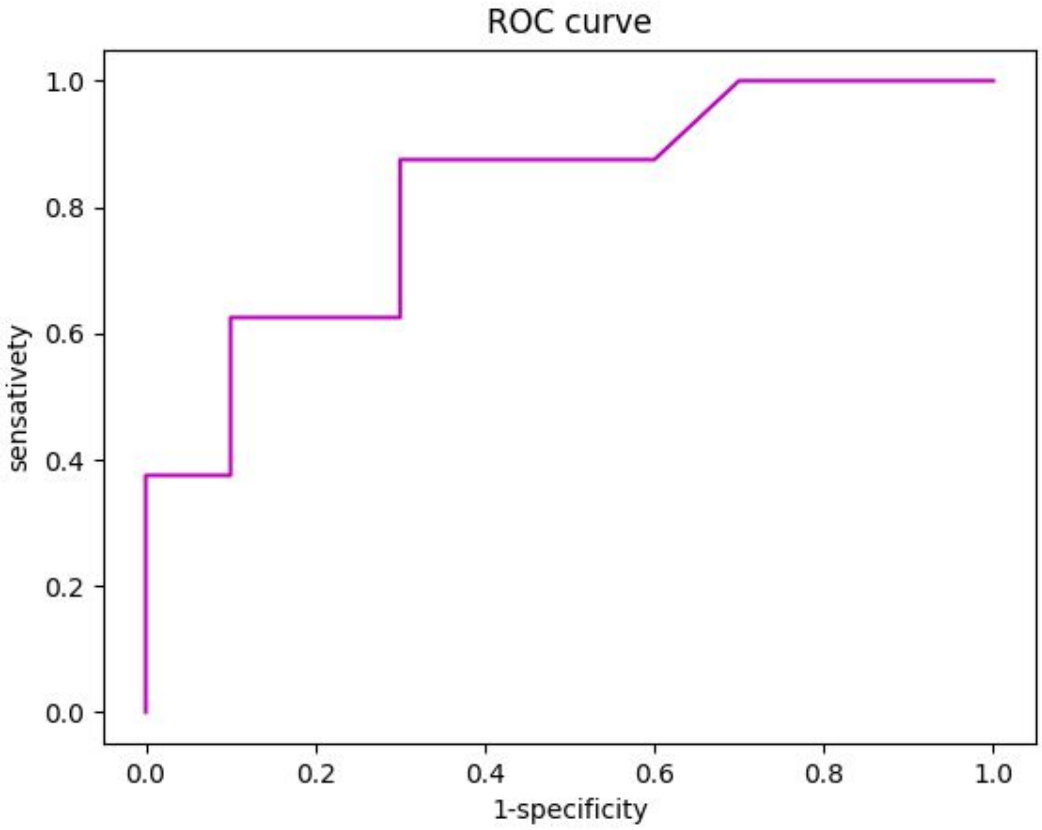
The main issue is that this will make the mapping of reads uncertain. Many reads will map equally well to both genes (these reads are called multi-mapping reads). We then have the choice to exclude multi-mapping reads (which would result in underestimating the expression of one or both of the genes), or counting them for both genes (which would result in over-estimating their expression), or assigning a count of 0.5 for each gene. The best approach would probably be to exclude multi-mapping reads, but then, when normalizing the gene expression using the FPKM approach, use the "mappable" gene length rather than the actual gene length, where "mappable" gene length corresponds to the length of the portion of the gene for which there is no multi-mapping issue. Still, if a gene have an exact copy elsewhere in the genome, we would not be able to know which copy is being expressed.

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?

See solution to past final exam.

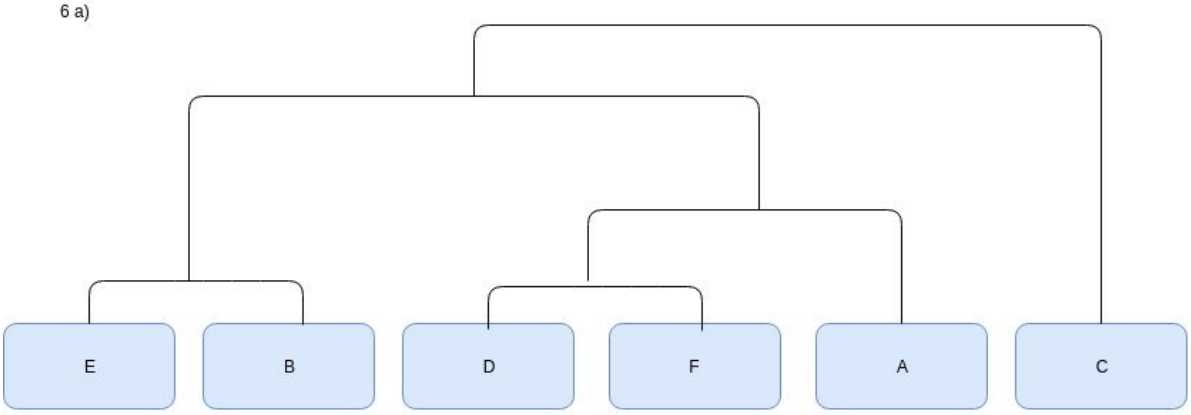
4 a) t-stat : 2.69, p-value : 0.0163
b) Consult the answer from the last year's assignment. The p-value of 0.017 with 100000 iterations using my code.
c) Consult the answer from the last year's assignment.

5 a) threshold 1.4, 2.1, 2.2 gives 4 prediction errors.
b) The ROC curve



6)

6 a)



6 b)