Homework 3 Comp 561 Xi Meng Huang 260608596

1.

a) see attached image 1

b) see attached program

c)

1. This contained the longest hydrophobic sequence, at length 167:

>tr|E9PGV9|E9PGV9\_HUMAN ATP-binding cassette sub-family G member 1 OS=Homo sapiens GN=ABCG1 PE=1 SV=1

2.Multiple proteins contained completely mixed regions. This was one of them and it contained the longest mixed region, with a length of 2345:

>sp|Q13085|ACACA\_HUMAN Acetyl-CoA carboxylase 1 OS=Homo sapiens GN=ACACA PE=1 SV=2:

3.

They match the properties described in the Hidden Markov Model to a degree. The hydrophobic is going to have shorter lengths(5) in comparison to the hydrophilic(8) and mixed(7), as its average length is shorter than the other two as described. However, the hydrophilic sections by that logic should also have greater lengths than the mixed, and I believe it is due the limited degree of distinguishability between the mixed regions and the hydrophobic and hydrophilic regions that lead to this result.

4. The hydrophobic amino acids are represented as ‘B’, hydrophilic-‘L’ and mixed –‘M’.

The hydrophobic regions are AVILMFYW, and hydrophilic regions are RHKDESTNQCGP. We will separate them into two graphs.

The hydrophilic amino acids demonstrate a large degree of hydrophilic presence, with the appearance of hydrophilic amino acids mostly showing up in hydrophilic regions, and thus in accordance with the model.

The hydrophobic graph separated into hydrophobic amino acids demonstrate that the frequency of appearance for the hydrophobic amino acids is indeed higher than it was in the hydrophilic graph.

d)Using the distributions from the data, i.e. the hydrophobic/hydrophilic amino acid distributions to remake the emission probability matrix and the acquired length distribution data to remake the transition probability matrix. Using the new probability matrices we re-construct an HMM and run the Viterbi algorithm again.

2.

a)The small possibility of the in-frame stop comes from

* when the state transitioning from exon to intron 1 emits a T and the state transitioning from intron1 to exon emits a ‘GA’, ‘AA’, or ‘AG’
* when the state transitioning from exon to intron 2 emits a ‘TA’, ‘TG’ and the state transitioning from intron2 to exon emits an ‘A’, or when it emits a ‘G’ when the previous was ‘TA’

This can result in an in-frame stop due to the fact that after splicing the codon will be a stop codon and stops transcription.

b)See attached image 2

3.See attached image 3

4.

a) The scoring method is the z-score method demonstrated in class. Z-score is calculated, for every candidate consensus. Z-score is found using the n-score and the e-score, where the n-score is the number of matches of the candidate in all the binding sequences. E-score is found in the following steps:

1. Using all the sequences in the non-binding random sequences, find sum, for all sequences I from 0 to n, all the values vi where vi= length of sequence I – k(in this case 6) + 1.

2. Multiply the sum with the probability of the consensus occurring, and the probability is found by going through every element in the consensus and adding up the probably of finding that element. For example, the probability of finding A is the number of times A is found in a congregated list of all the sequences, over the length of the congregated list. For candidate consensus with multiple possibilities, such as [A,G], the probabilities is the probability of finding A + probability of finding G.

3. The Z-score is found by subtracting the e-score from the n-score, and then dividing the whole thing by the square root of the e-score.

b)see attached program

c) [[‘A’,‘G’],[‘A’,‘G’],[‘A’,‘G’],[‘A’,‘G’],[‘A’,‘G’],[‘A’,‘G’]] with a z-score of 276.294050735