Comp 561 Assignment 3

Jonathan Pearce, 260672004

November 14, 2017

Problem 1a. Please refer to image attached at the end of the document

Problem 1b. Please refer to viterbi.py file

Problem 1c.1. The 104th protein (103 index in array) in the input file contained the longest hydrophobic region, it had length of 141 amino acids.

Problem 1c.2. 424 of the proteins had their entire sequence annotated as belonging to a Mixed region

Problem 1c.3. Please refer to graphs attached at the end of the document.
All 3 distributions do not match the properties listed in the question. The average length of the Hydrophobic, Hydrophilic and Mixed regions was about 38, 82 and 281 respectively which is much greater than the average outlined in the question for each type of region

Problem 1c.4. The order of the amino acids: 'A', 'V', 'I', 'L', 'M', 'F', 'Y', 'W', 'R', 'H', 'K', 'D', 'E', 'S', 'T', 'N', 'Q', 'C', 'G', 'P'

The frequency scores were as follows:

Hydrophobic: 0.106, 0.109, 0.09, 0.198, 0.047, 0.093, 0.048, 0.03, 0.017, 0.01, 0.017, 0.012, 0.012, 0.049, 0.038, 0.014, 0.015, 0.012, 0.04, 0.031

Hydrophilic: 0.045, 0.039, 0.022, 0.054, 0.012, 0.02, 0.014, 0.007, 0.067, 0.024, 0.057, 0.049, 0.097, 0.125, 0.072, 0.028, 0.05, 0.019, 0.085, 0.105

Mixed: 0.074, 0.068, 0.052, 0.108, 0.025, 0.042, 0.03, 0.015, 0.053, 0.024, 0.05, 0.046, 0.063, 0.07, 0.053, 0.035, 0.042, 0.018, 0.069, 0.052

The Hydrophobic region frequencies do a decent job of matching the properties in the question. The frequency of A,V,I,L and F are all higher than the non-hydrophobic amino acids which is accurate however M,Y,W are all particularly low, further most of the non-hydrophobic amino acids have relatively low frequency. The hydrophilic frequencies do a slightly worse job matching the properties in the question. The frequencies for E,S,T,G,P are all higher than the hydrophobic amino acid frequency which is good, however some of the

hydrophobic amino acid have too large of frequencies, particularly A,V and L. The Mixed region frequencies is not quite perfectly uniform but most of the values fall within a pretty tight range which is promising. In general with only using a small data set such as this one provided in the question, I feel as though the frequencies for each region match the properties in the question pretty well.

Problem 1d. Clearly from the analysis in part C. the average length of regions in our HMM is the biggest concern since it is so far from what we expected. Once the HMM moves into a state it is very difficult for the HMM to move out of that state to a new one and thus this produces incredibly long regions of the same state. In order to resolve this issue the transition probabilities would need to be changed, in particular the probability of staying in the current state would be need to be dropped significantly. This probability weight should be redistributed to the transitions to other states probabilities in proportion to their original ratios (i.e. if we removed 0.2 from the hydrophobic to hydrophobic transition probability then this 0.2 would be distributed in the 80/20 split outlined in the question for mixed and hydrophilic regions respectively, therefore the transition from hydrophobic to mixed would obtain 0.2*0.8 = 0.16 increase in weight and the hydrophobic to hydrophilic would obtain 0.2*0.2 = 0.04 increase in weight). A fair amount of tuning these probabilities shifts may be required to reach a new transition matrix that does a better job of following the statistical properties of each type of region.

Problem 2a. Consider the path through Intron1 in the image in the question. Suppose in the intermediate state where we generate one base pair we generate a 'T', then we cycle through the intron1 state for a while and then move back to the other intermediate state where we generate our second and third base pair that will complete our codon, it is entirely possible that we could generate an 'A' and then a 'G' making this codon a 'TAG' stop codon. In this case when the intron separating this Codon is removed and the exons are processed, there will now be a stop codon in the middle of our sequence. The same error can occur in the Intron2 path. For example 'T' and 'A' could be generated and then after the intron phase a 'G' could be generated creating a 'TAG' stop codon.

Problem 2b. Since the HMM only considers the previous state then we will not be able to keep track of the production of each codon in the Intron1 and Intron2 paths. I propose as a solution to simply partition these paths. For example in the Intron1 path, when we move to the intermediate state we will have 2 possible transistions, one will be a path where the base pair at the intermediate state can only be 'T' and the other path will be to an intermediate state where the base pair can be 'A','C' or 'G'.
The second path just described will behave almost exactly the same as the original Intron1 path in the question except for some minor probability adjustments to compensate for the first base pair not being allowed to be a 'T'.
The probability of moving to the other intermediate state described will simply be the probability of a base pair being 'T' and at that state it will be guaranteed to generate a 'T' base pair. From there the Intron1 state is the same. In the second intermediate state it will be impossible to generate 'AA', 'AG' or 'GA' thus ensuring a stop codon will not be produced. A picture has been attached below in the document depicting what I have just

described. A nearly identical strategy can be used for the Intron2 path, except we will need three different transitions from the Exon state. One to only allow for 'TA' generation, where after the Intron2 State an 'A' or 'G' base pair would be impossible. one to only allow 'TG' generation, where an 'A' base pair would be impossible after the Intron2 state and finally a third for all other cases that would be treated as normal.

Problem 3. Decomposing the Gene State into 1000 distinct Gene States is an approach that I believe would yield an HMM that would satisfy the target length distribution perfectly. The $non-gene$ state can only transition to the $gene_1$ state and the $gene_k$ state can transition to either the $non-gene$ state or the $gene_{k+1}$ state $\forall k \in 1, 2....999$ and finally the $gene_{1000}$ state can only transition to the the $non-gene$ state. A picture depicting this HMM structure is attached below.

Now we must consider the transition probabilities in order to satisfy the target length distribution. We can use conditional probability rules to calculate the probability of transition from state $gene_k$ to state $non-gene$.

$$\begin{aligned}
&P(\text{transistion from } gene_k \text{ to } non-gene) \\
&= P(\text{Stop at length K} \mid \text{We have reached length K}) \\
&= P\frac{(\text{Stop at length K, We have reached length K})}{P(\text{We have reached length K})} \\
&= \frac{P(\text{Stop at length K})}{P(\text{We have reached length K})} \\
&= \frac{p_k}{1-\sum_{i=1}^{k-1} p_i}
\end{aligned}$$

Using this we can get the probability of transition from state $gene_k$ to state $gene_{k+1}$

$$\begin{aligned}
&P(\text{transistion from } gene_k \text{ to } gene_{k+1}) \\
&= 1 - P(\text{transistion from } gene_k \text{ to } non-gene) \\
&= 1 - \frac{p_k}{1-\sum_{i=1}^{k-1} p_i}
\end{aligned}$$

Using this structure and these formulas to calculate the transition probabilities, this HMM will satisfy the target length distribution perfectly.

Problem 4a. I will use the scoring method we discussed in class. I will only examine the positive sequences for my analysis.

Given a consensus sequence C, let N(C) be the number of matches of C in $S_1, ...S_n$, let

E(C) be the number of expected matches for C in n random sequences. Define Z(C) as follows,

$$Z(C) = \frac{N(C) - E(C)}{\sqrt{E(C)}}$$

Whichever consensus sequence C produces the largest Z-value, this will be deemed the best consensus sequence

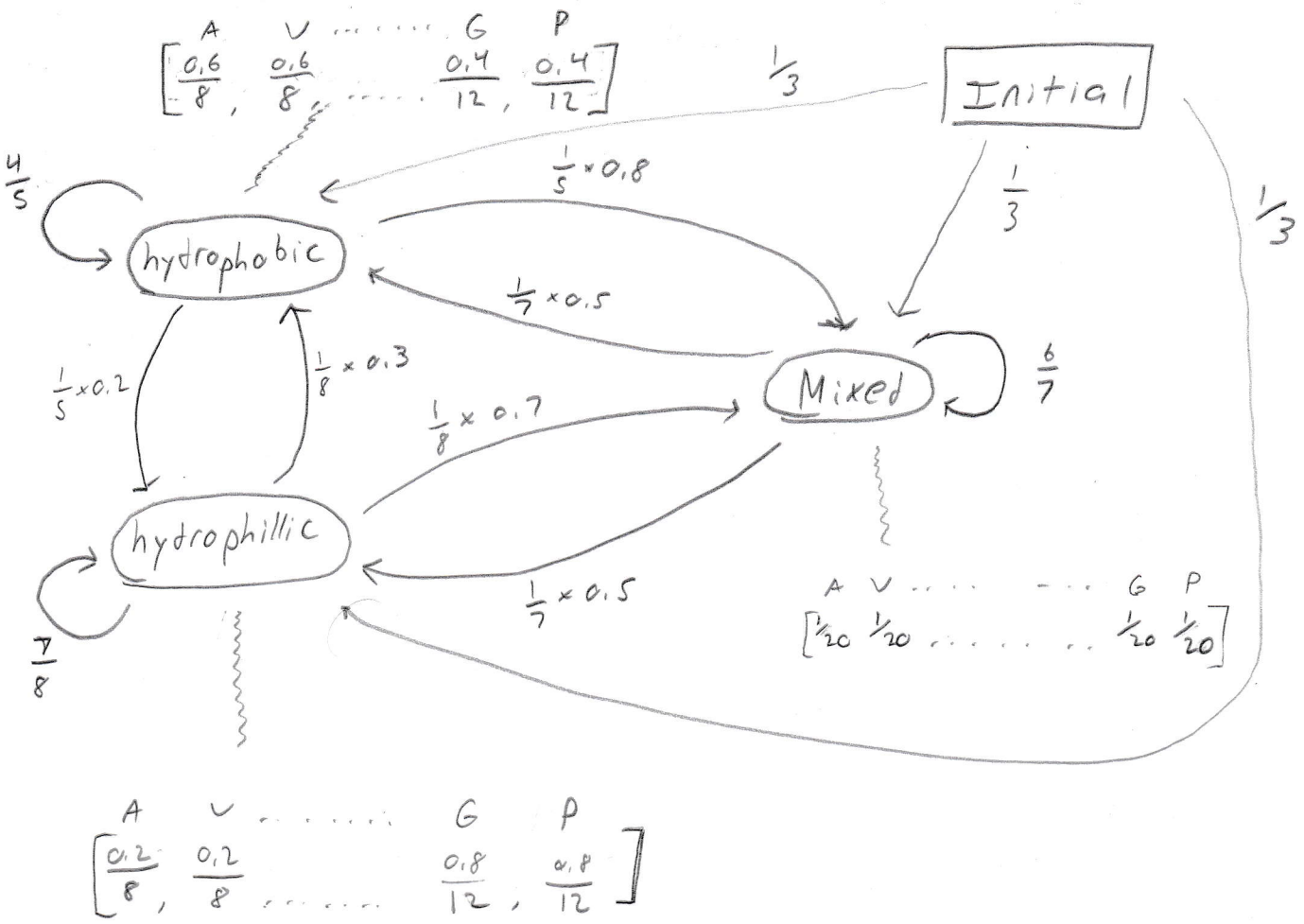
I do understand that the non-GATA2 file could have been used to improve the accuracy of the prediction of the consensus sequence

Problem 4b. Please refer to motif.py file

Problem 4c. Note: I finished my code too late on the day of the deadline and I was only able to test it with the first 12000 lines of input (instead of the 62000 provided).

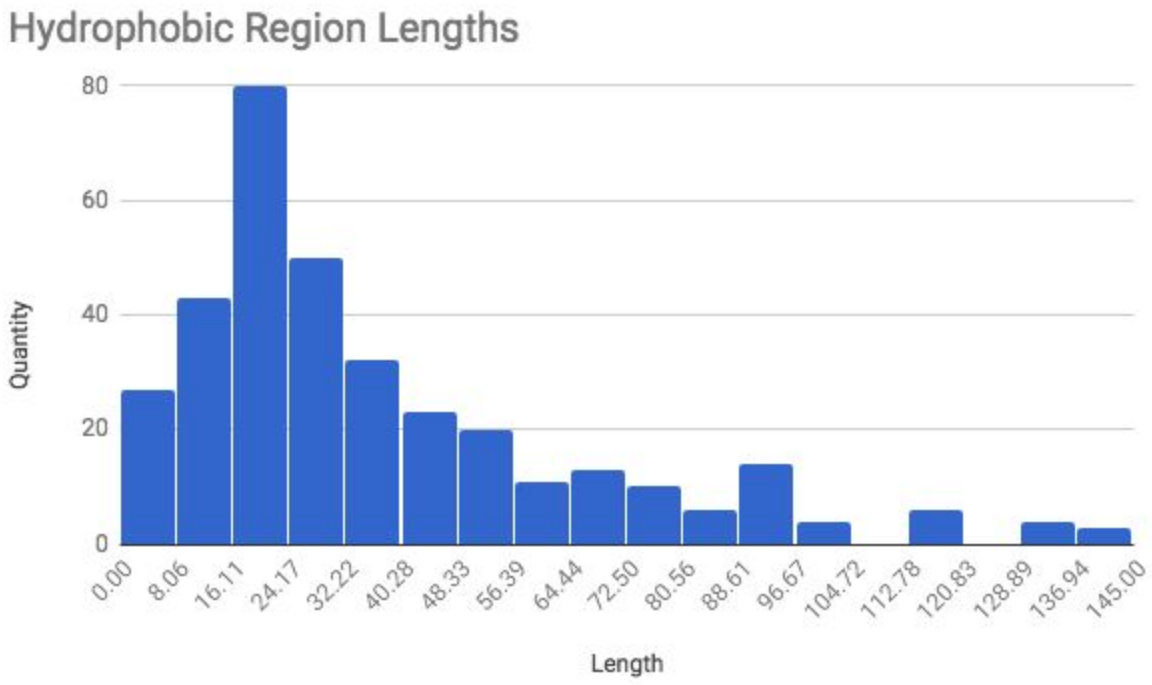
The best consensus sequence I found was: $(C \mid T)$, $(C \mid T)$, $(C \mid T)$, $(C \mid T)$, $(C \mid T)$, $(C \mid T)$. With a final score of 2847.9799

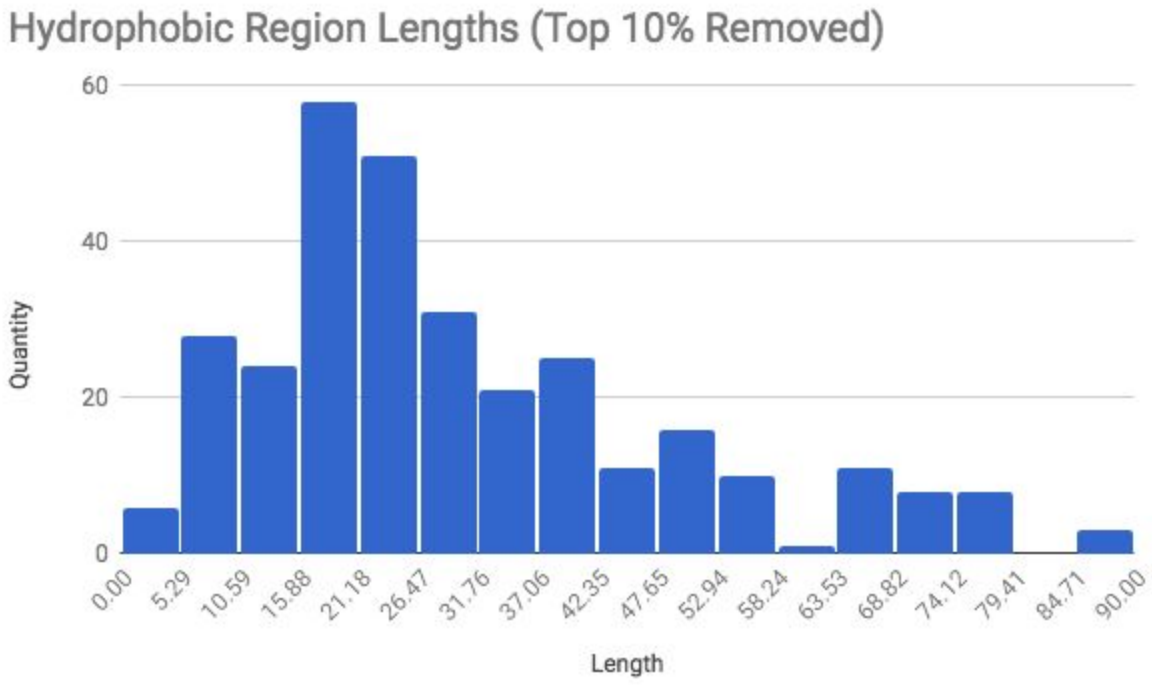
1) a)


A V ... G P
[0.6/8, 0.6/8, ..., 0.4/12, 0.4/12]
1/3
Initial
1/3
1/3
4/5
hydrophobic
1/5 × 0.8
1/7 × 0.5
1/5 × 0.2
1/8 × 0.3
1/8 × 0.7
Mixed
6/7
hydrophillic
1/7 × 0.5
7/8
A V ... G P
[1/20 1/20 ... 1/20 1/20]
A V ... G P
[0.2/8, 0.2/8, ..., 0.8/12, 0.8/12]

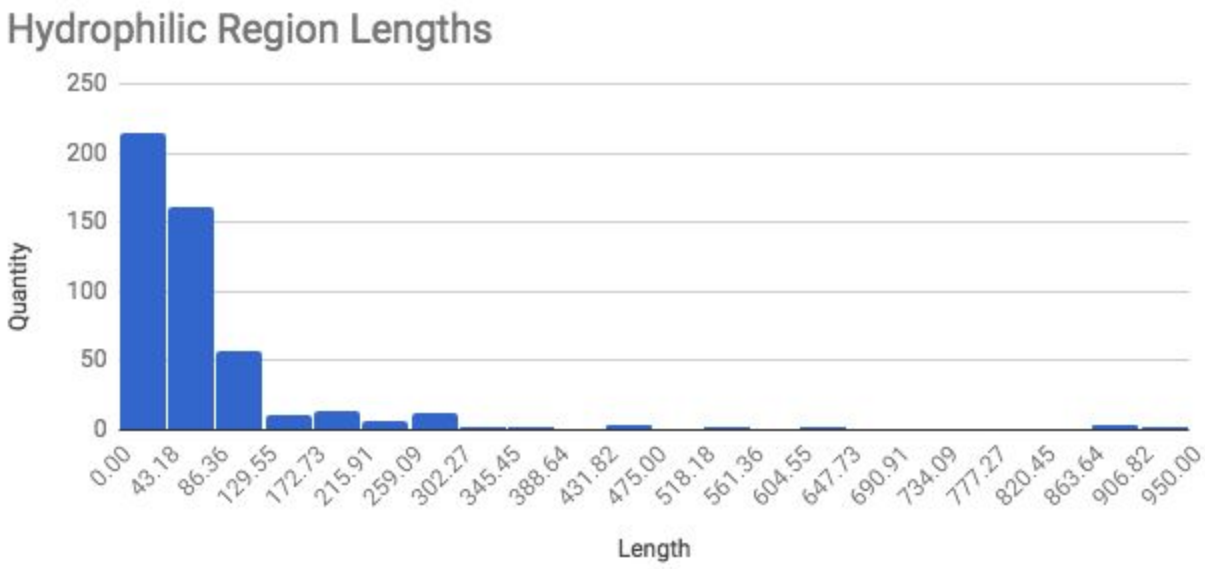
Comp 561 Assignment 3 Q1.c.3
Jonathan Pearce

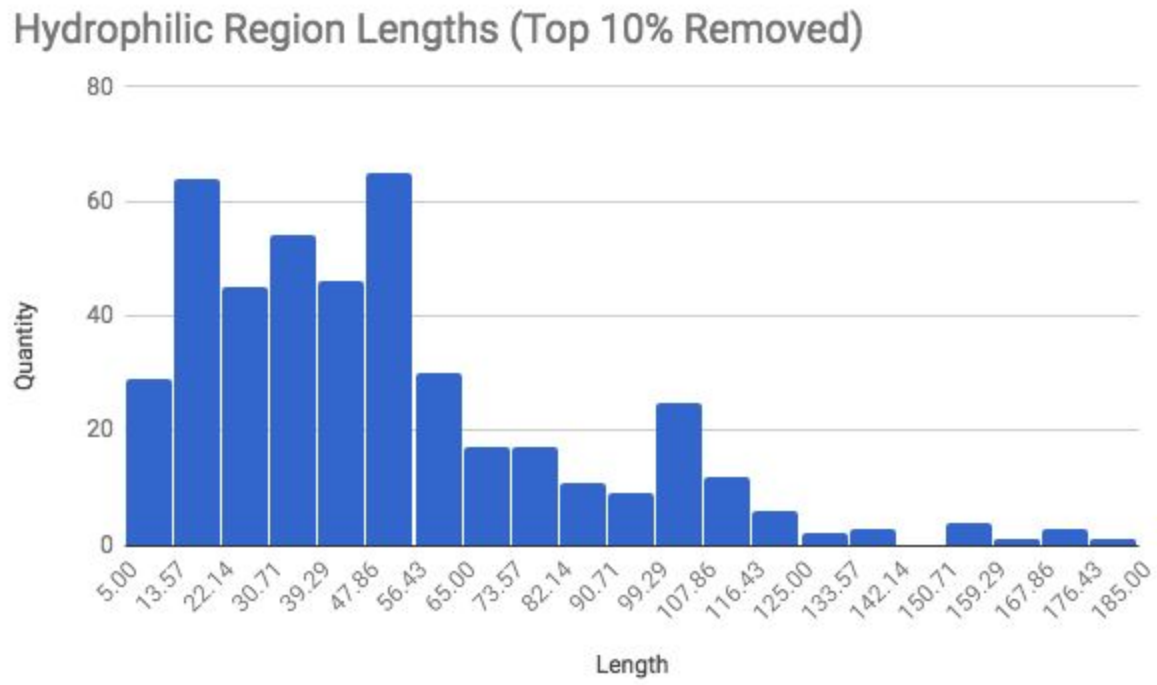
Hydrophobic Graphs



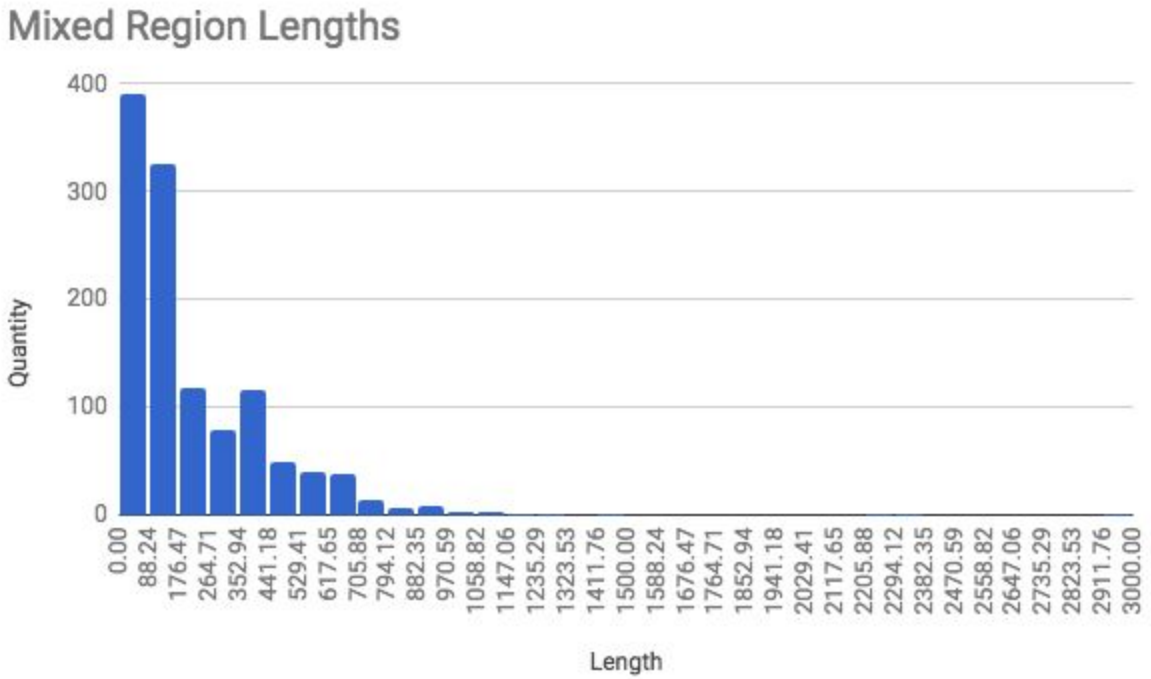


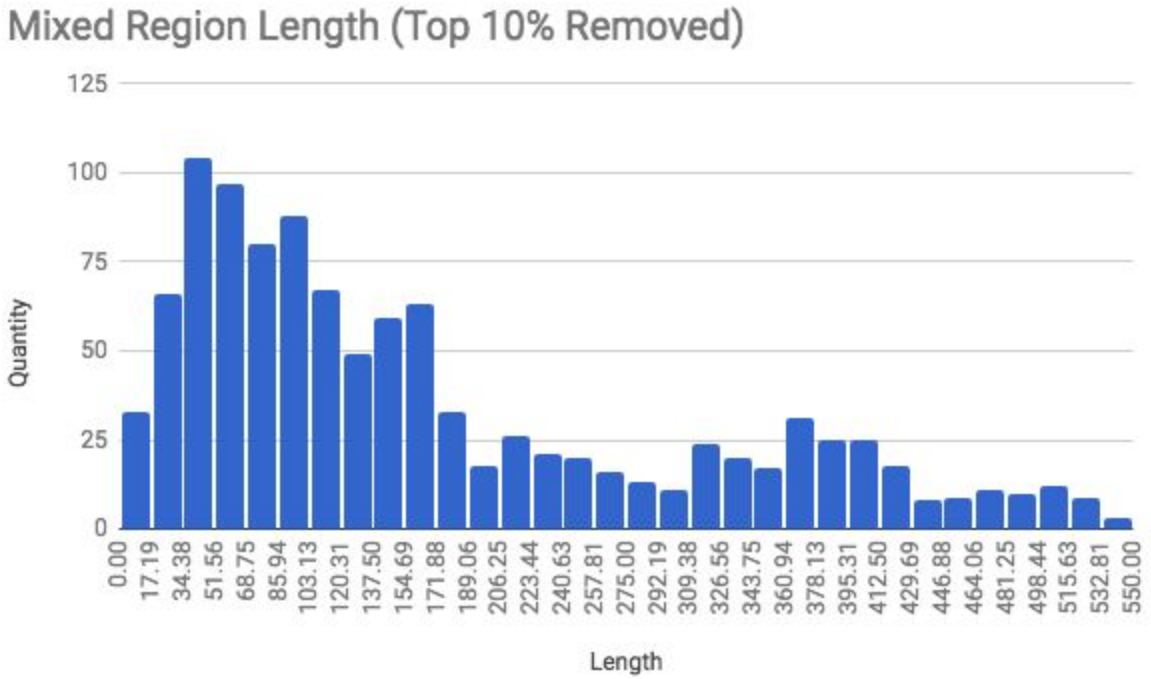
Hydrophilic Graphs



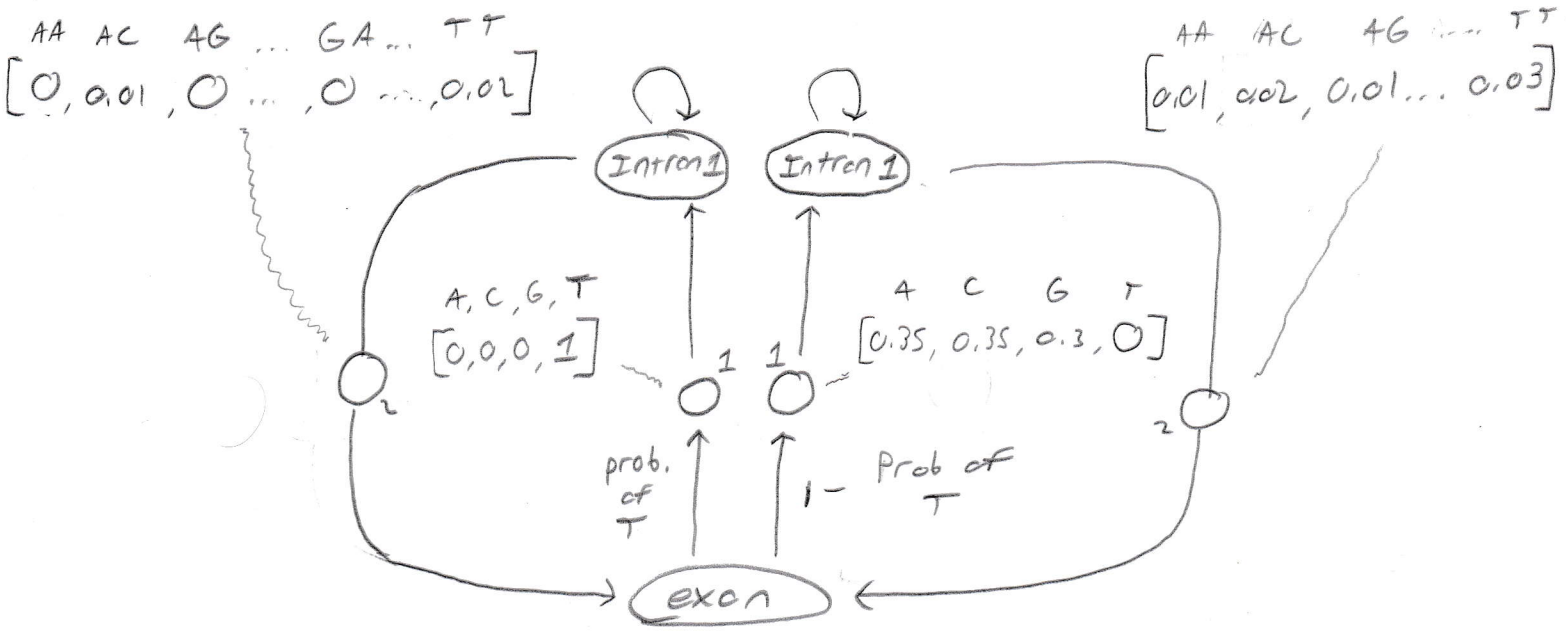


Mixed Graphs





2) b) Demonstrating for Intron 1, same idea would need to be applied to Intron 2 path.



3)