Q1.

a.

i. Use the difference of start of next CDS- stop of last CDS, I obtained the intergenetic regions. Take average of the data, I got the estimated length of intergenetic regions is 793.141494. Since the length should be integer, then I use 793 bps for intergenetic length.

ii. I modify the gff3 file into text file, and only keep start and stop column for CDS. Then, I calculated the difference of stop- start, and sum up. Here I got the sum of difference is 1868982, and 1887 positive strands. Average length of gene length= 1868982/1887=990.45151. Since the length of gene should be integer, so I take 990 bps for average of gene length.

```
Sum of 1rd field: 1868982. Total number of lines: 1887
```

iii.

A	T	C	G
0.266	0.265	0.243	0.226

iv.

TTT	0.0255	TCT	0.0111	TAT	0.0156	TGT	0.0057
TTC	0.0137	TCC	0.0059	TAC	0.0142	TGC	0.0041
TTA	0.0192	TCA	0.0104	TAA	0.0020	TGA	0.0006
TTG	0.0226	TCG	0.0092	TAG	0.0006	TGG	0.0131
CTT	0.0127	CCT	0.0112	CAT	0.0127	CGT	0.0202
CTC	0.0145	CCC	0.0058	CAC	0.0106	CGC	0.0175
CTA	0.0088	CCA	0.0127	CAA	0.0331	CGA	0.0052
CTG	0.0290	CCG	0.0107	CAG	0.0184	CGG	0.0028
ATT	0.0308	ACT	0.0129	AAT	0.0189	AGT	0.0115
ATC	0.0251	ACC	0.0205	AAC	0.0201	AGC	0.0140
ATA	0.0036	ACA	0.0075	AAA	0.0357	AGA	0.0026
ATG	0.0260	ACG	0.0112	AAG	0.0136	AGG	0.0009
GTT	0.0165	GCT	0.0209	GAT	0.0371	GGT	0.0268
GTC	0.0144	GCC	0.0221	GAC	0.0144	GGC	0.251
GTA	0.0110	GCA	0.0193	GAA	0.0390	GGA	0.0074
GTG	0.0285	GCG	0.0304	GAG	0.0242	GGG	0.0086

b. The code for b is in the Viterbi.py. To probably run it, you need to put the path of test fasta file in the line of fasta_file=" ". For the configuration file, I put it in the viterbi.py which like dictionary. The transmission and emission probability calculated by hand using the information in a. For example, the T(start|intergene)=1/length(intergene), T(intergene|intergene)=1-1/length(intergene). For emission probability, I am trying to convert it into probability with single nucleotide.

c. See Q1c.gff3

d. My program failed to detect the end position for all strand (where I put -1). Also, the if of output of my program not in proper order. Some of the start position (the numbers) match with the annotated gene file, but I think it's more likely to be random match. I think there are several reasons that my program fails in gene region that emit a codon. I was not able to implement these cases with proper code. Secondly, I ignore the score and phase of the gff3 file at all, so all the strands in my predation have same score and phase.

e.

As I mentioned, I completely ignored the difference in score and phase in my program. Annotated gene have more varies of gene with different score and phase which might be completely missed by my predication. Secondly, the annotated gene contains both positive and negative strands which increase the difference. With both direction of strands, more gene coding region will be contained.

The number of CDS regions in annotated gene

For my predication, I mess up the part with emit codon in gene region which much lower the accuracy. Besides, the annotated gene dealt with the overlapping region, and I basically ignored it at all. And I failed to deal with end position, which might relate to failing to detect the codon, which make my result have too much gene regions without proper cut.

Number of CDS regions in my predication

Q2.

We would like to have the duration of stay on the gene region is exactly to the desired length distribution, then we need our best path found very close to the desired length. I would like to use algorithm like Viterbi training. However, the difference is that Viterbi is looking for the optimal path no matter if it's gene or intergenetic region for next. In my modification, I would like to find the optimal path for stay in gene region longer. Firstly, I would like to estimate reasonable parameters (transition probability) with desired length probability pk. I assumed that T(gene|non-gene)=1/l where l is the length of non-gene region, and T is the transmission probability. The length of gene would depend on how long the path state will stay in the gene region, then I will splice the gene state into three sub-states that are start, coding and stop reions. T (gene | start codon) = 1, T (gene| gene) = 1-(1/(k/3)) =1-3/k since a group of three nucleotides are coding into one amino acid, and T (stop| gene) =1/(k/3) =3/k since only one stop codon will stop the coding region. T (non-gene| stop) =1 since gene coding stop after stop codon. With the parameters, I assume M(si,x1)=π(si)E(x1|si), $\forall$si where initial probability π and emission E are known from 1st HMM model. Then, I would like to change the transmission probability to make the path prefer gene->gene, and less preferable to gene->stop, and the coding of gene will not stop until coding length=k. Therefore, $^{k}{}_{0}|M(si,xj)_{gene}=\,^{k}{}_{1}|\max\{M(si,x_{j-1})T(gene|gene)E(xj|gene), \forall si,j>1\}$ which means that $M(si,xj)_{gene}$ have to repeat for k times, and count decrease after every run until It reach 1. Then it goes to $M(si,xj)=\max\{M(si,x_{j-1})T(stop|gene)E(xj|gene), \forall si,j>1\}$. The rest of part of the HMM will be same as Viterbi algorithm. In my modification, I would like to have the gene region to extend as long as k, hence the final duration of stay on the gene region is more likely to have same distribution with desire length.