Q1.)a.)

Average gene size was 990.4515103338633 nucleotides. I calculated this by differencing the end and start positions in the annotation file.

The average intergenic size is 1035. Intergenic size was found by taking the regions in-between gene areas.

Genic occurrence: ATT 0.030877040890044793, CTT 0.012710109329095024, GTT 0.016492478607779866, TAT 0.015676925299812166, AAA 0.035706143941948015, CAA 0.03317118592367834, AAC 0.0201319655155806, CAC 0.010610220103067957, GAA 0.03906307694777569, AAG 0.013532084316653021, GAC 0.014407038160830965, CAG 0.01834673858947808, GAG 0.024212942895214244, TGA 5.747403233315674E-4, TGC 0.0041162966173802756, TCA 0.010433623914334795, AAT 0.01890381929393633, TCC 0.00580680376952592, TGG 0.013069723386151647, CAT 0.012724558108173193, TCG 0.009136444637094832, GAT 0.03718794650740901, TTA 0.01925861709130023, TGT 0.00567355391802726, TTC 0.013673361267639551, TCT 0.011123954470291705, TTG 0.02260752299764003, TTT 0.02560323652651351, AGA 0.0025108767198060654, CGA 0.0051951387885501455, AGC 0.013941466390534445, CGC 0.01753921238099825, ACA 0.007519786800237602, CCA 0.012729374367865915, GGA 0.007407407407407408, ACC 0.020507633771612965, AGG 8.31607506943441E-4, CCC 0.00578432789095988, GGC 0.0251183997174461, CGG 0.0027821926824961067, GCA 0.019282698389763844, ACG 0.011234728443224325, GCC 0.02217887588498772, GGG 0.008568125993353562, CCG 0.010692096517844241, GCG 0.03044197209780218, AGT 0.011467514328372586, ATA 0.0035255020950729662, ATC 0.025158535214885452, CTA 0.008792884779013951, CGT 0.020210631090561735, ACT 0.01293326269485784, CTC 0.014495336255197546, ATG 0.025741302637704893, GTA 0.01099231003869062, GGT 0.02683459358795293, CCT 0.011188171266194672, TAA 0.0019505851755526658, GTC 0.014411854420523689, CTG 0.02906131098588836, TAC 0.014198333574146318, GCT 0.020942702563855577, GTG 0.02849299234214709, TAG 5.041018478383021E-4. Genic occurrence found by taking the codon at the start and then iterating through the gene. Gene truncated to a multiple of 3. If front is overlapped, took to where the overlap ends.

Intergenic Occurrence: A 0.2649275653933368, C 0.24376297574052377, T 0.2647649336701878, G 0.2265445251959517. Found by taking intergenic regions and averaging each nucleotide found.

Also found the start and stop codon emission from gene sequence.

Q1c.) File is named question 1c

Q1.d)

The fraction of annotated genes that:

- Perfectly match both ends of one of your predicted genes: 1127/2443
- Match the start but not the end of a predicted gene: 60/2443
- Match the end but not the start of a predicted gene: 0/2443
- Do not match neither the start not the end of a predicted gene: 1256/2443

The fraction of your predicted genes that:

- Perfectly match both ends of one of an annotated genes: 1127/2978
- Match the start but not the end of an annotated gene: 60/2978
- Match the end but not the start of an annotated gene: 0/2978
- Do not match neither the start not the end of an annotated gene: 1791/2978

1e.) If genes were not close to the average size of genes they were likely missed. The average length of complete matches was 986.6415261756877 which is close to the average length of all genes.

No genes were found that had the end point but no start point. Genes predicted but not present in annotation were close to the average gene size.

2.) Since moving into a new state has no knowledge of the past state, you must create a sequence of (*k-1)* long gene states. Connecting each gene state is a transition probability of 1 and each emit genic. Therefore, by entering the sequence from the start ensures that you stay in the gene state for the length of the sequence. The transition from intergenic to genic stays the same. But, the first genic node connects all the gene state sequences of length *k-1*. The transitions are the probability of going to a sequence of length *k-1* is set by Pr[length == *k*] (since the starting state is also genic). Therefore, when moving to the first genic region it chooses the length of the genic sequence by the probability of Pr[length == *k*].