02-725 HW2

Anushka Sinha

1. EMISSION PROBABILITIES

The emission probabilities are as follows:
Encoding regions:

A	C	G	T
0.1000532	0.3999538	0.3999818	0.1000112

Non encoding regions:

A	C	G	T
0.2499583	0.2501792	0.249929	0.2499335

2. GENE FINDING

To assign a label of encoding regions vs. non-encoding region to every gene sequence in the unlabeled.txt, found the likelihood for every sequence given each state (encoding or non-encoding), compared both likelihoods and assigned the state to the sequence with higher likelihood. See code for reference.

3. BUILDING HIDDEN MARKOV MODELS

Used Viterbi decoding to find the encoding and non-encoding regions in the genome.txt file. See the jupyter notebook for code.

4. LEARNING HMMS

The transition and emission probabilities matrices started to converge after the third iteration. See code for reference. Here are the values they converged at:
Transition matrix $\begin{bmatrix} \text{9.99894692e-01} & \text{1.05308259e-04} \\ \text{3.27879289e-03} & \text{9.96721207e-01} \end{bmatrix}$

Emission probabilities for the encoding region:

A	C	G	T
0.111988,	0.384350	0.388680	0.114980

Emission probabilities for the non encoding region:

A	C	G	T
0.228922,	0.2708453	0.2711395	0.229092

5. HMM DURATION DISTRIBUTIONS

Assuming an exponential distribution, solved for $\Lambda_{\text{MLE}} = \frac{n}{\sum_{i=1}^{n} x_i}$

Obtained $\Lambda_{\text{MLE, encoding regions}} = 0.00388$
Obtained $\Lambda_{\text{MLE, non-encoding regions}} = 0.000142$