CBIO Exercise 3

Noam Blum & Alon Ben Zion

# Question 1

We used the model suggested in the exercise description, that is for each base we have one hidden state for it being in an island and one where it isn’t, denoted A+, A-, etc. The initial parameters are as following:

Table

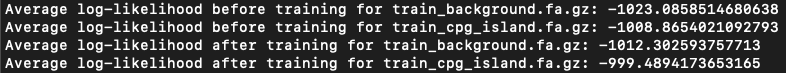
Description automatically generated

The logic behind these choices is as following:

1. For the transition matrix, we first took the matrices shown in lecture 5 of this course, showing the probability of any base coming after another within and outside of CpG islands. They correspond to the top-left and bottom-right quarters of this matrix, respectively.  
   Then we denoted the probability to stay in an island and the probability to stay outside an island and respectively. Due to the Markovian Independence assumption, we could look at exiting an island as a geometric process, that is denoting as a random island’s length we get . Since we know CpG island length averages 200bp, we could deduce and easily calculate . We assumed would be higher and chose .  
   We could then use these values to unify the two matrices from the lecture as following: we multiplied each matrix by its correspondent probability, and assumed that for all bases, exiting or entering an island would be given equal probability for all pairs, i.e. when you enter or exit an island, you forget which base the previous section ended with. Therefore, we gave the top-right and bottom left sections the values , respectively.
2. For the starting probability, we took the amount of CpG islands in the human genome, which is 45,000, multiplied it by the average length of a CpG island, 200bp, and divided the result by the length of the human genome, bp, which got up the percentage of CpG islands in the human genome. We then gave within CpG islands a 15% probability for A/T and 35% for C/G, and outside islands 25% each. Each base probability was multiplied by its respective CpG/Non-CpG probability to produce that vector
3. The emission matrix is obvious, as each hidden states represents a base

When we trained the model, we didn’t let it change its emissions matrix, for the reason specified above, so it learned only the transition and start probability matrix, with the result parameters shown below:Table

Description automatically generated

The difference in average log-likelihoods for the training data before and after training the model is shown below  


# Question 2

The average likelihood per base for the two test files are shown below:  


For the non-island test file, it is slightly higher than we would expect, but not by much, since we would expect all bases to be as likely, so that would average around 25%.

For the island test file it is lower than we would expect, since we could expect that having a high presence of C and G, with matching high probabilities in the model, would increase the likelihood.

# Question 3