**Homework #4**

**This homework will be due Oct 20th. When you submit your answers, please include your outputs in the write-up.**

1. The promoter of a gene (the regulatory DNA sequence upstream of the transcribed region) is centrally important in determining if and when transcription will be initiated. Within the promoter, transcription factor (TF) binding motifs (TFBM) are of great importance for TF binding and their length is usually around 6bp. The sequence of a TFBM could be variable, some positions have the fixed bases, while there could be a number of bases at some other positions in the motif. Now imagine that we are studying a TF binding motif, and as shown in the table below, we have a different probability of bases at each sequence position and the total probability of the string is a product of these numbers. Please finish all the following questions.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | A | T | C | G |
| Position1 | 0.3 | 0.0 | 0.0 | 0.7 |
| Position2 | 0.0 | 0.4 | 0.4 | 0.2 |
| Position3 | 0.9 | 0.0 | 0.0 | 0.1 |
| Position4 | 0.0 | 0.3 | 0.3 | 0.4 |
| Position5 | 0.5 | 0.1 | 0.1 | 0.3 |
| Position6 | 0.2 | 0.2 | 0.2 | 0.4 |

1. Calculate the probability of “ATACAG”. Also, among all possible 6-mers, which 6-mer has the highest probability?
2. Write a program to detect the motif that is modeled by the table above. Assume that you have an input string and the outputs are all the start position of all substrings that fit the motif model with probability higher than 0.005. (The promoter sequence motif.fasta is provided. Also, for your convenience, you can save the above sequence profile table as .csv, and import it into your code)

**For question 2 & 3, pick one to complete:**

2. Assume that the probability for the first starting letter of a motif in the promoter is as follows:

P(A) = 0.4, P(T) = 0.1, P(C) = 0.1, P(G) = 0.4.

Then the next letters are generated by a Markov process with the following transition probability:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | A | T | C | G |
| A | 0 | 0.4 | 0.4 | 0.2 |
| T | 0.8 | 0.05 | 0.05 | 0.1 |
| C | 0.6 | 0.1 | 0.2 | 0.1 |
| G | 0.1 | 0.3 | 0.3 | 0.3 |

1. Write a program which is able to process any strings and compute the its probability according the above model. (The string’s length is 6bp)
2. Using your Markov Model program to calculate the probabilities of the following strings:

ACCTGC

AGGTCG

3. Suppose a string has two states: conserved promoter motif (M) and random background (BG). In the M state the string has the following letter output probabilities:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | A | T | C | G |
| Position1 | 0.3 | 0.0 | 0.0 | 0.7 |
| Position2 | 0.0 | 0.4 | 0.4 | 0.2 |
| Position3 | 0.9 | 0.0 | 0.0 | 0.1 |
| Position4 | 0.0 | 0.3 | 0.3 | 0.4 |
| Position5 | 0.5 | 0.1 | 0.1 | 0.3 |
| Position6 | 0.2 | 0.2 | 0.2 | 0.4 |

In the BG state the letters occur in a completely random chance, which means the P(A), P(T), P(C), p(G) are always 0.25.

The probability of the M state is 0.6, and the probability of the BG state is 0.4. Transition probability between M and BG is given by the following table:

|  |  |  |
| --- | --- | --- |
|  | M state | BG state |
| M state | 0.8 | 0.2 |
| BG state | 0.3 | 0.7 |

Write a program for the forward algorithm of Hidden Markov Models (HMM) to calculate the probability of observing the following sequences (You should only consider the case that each string 6-mers followed by 6-mers) :

Seq = “AGTGGAATTAAACATAATTAAAGACTTTAAATATAAGTTTGAATTAAA”