Homework 7

Due: April 22th at 11:59pm

1 Hidden Markov Models (100 pts)

1.1 CpG Islands

CpG-islands are regions of a genome that contain a high frequency of the dinucleotides CG and GC. The CpG notation is used to differentiate the linear sequence in the DNA from C-G pairing in the complimentary DNA strands. CpG islands are usually found near the start of a gene, in many cases within or near the regulatory regions, so they are useful for the identification of novel genes. In this problem you will use a Hidden Markov Model to identify CpG-islands in a segment consisting of 50,000 base pairs from the genomic sequence of Human Chromosome 1. Information about CpG islands in Human Chromosome 1 is available from this link. The sequence associated with this section of chromosome 1 has been uploaded to the website. The starts and ends of the blocks are:

```
1000k-1003k
1016k-1018k
1019k-1021k
1032k-1035k
1040k-1042k
1043k-1045k
1048k-1050k
```

The dinucleotide transition probabilities in CpG-islands are different from that in non CpG islands. Your HMM will have a total of 8 states - a group of 4 states A+, C+, G+, and T+ which emit A, C, G, and T respectively in CpG-islands, and another group of 4 states A-, C-, G-, and T- correspondingly to normal (non CpG) genomic regions. For transition probabilities within each group we will use the following transition probability tables:

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

-	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

1. (both COMP and BIOL students) Discuss how you will design the transition probabilities between the states across groups. How about initial state probabilities? Justify the choices you make.

2. (COMP students only) Implement this model. Find the most probable path, given the observed sequence, using the Viterbi algorithm. Use this path to annotate the given sequence as CpG island regions and non CpG island regions and include it in your answer to this question.

3. (COMP students only) Compare your results with the Genbank annotation available from the NCBI site mentioned above.

1.2 HMM Modeling

Please thoroughly explain your reasoning in each of the following questions. Include a description of the model and any assumptions made.

1. (both COMP and BIOL students) The secondary structure of a protein can be summarized by describing which amino acids lie within α-helices, β-sheets, or neither. As a simplifying assumption, assume we are only interested in α-helices, β-strands (the strand components of the sheets). Is it possible to infer the approximate secondary structure of proteins using HMMs? How about if we include β-sheets?

2. (BIOL students only) Is it possible to formulate regulatory binding site inference using HMMs?