LAB#3 – Probabilistic Models

The goal of this Lab is to analyze biological sequences using probabilistic models, such as Hidden Markov Models (HMM). Group I addresses CpG islands characteristics; the goal of Group II is to practice probability and statistics concepts; and Group III includes the implementation of algorithms to infer the most likely paths in HMM.

Group I (3.5 points)

CpG islands are regions of DNA characterized by a large number of adjacent cytosine and guanine nucleotides linked by phosphodiester bonds. Additionally, CpG islands appear in some 70% of promoters of human genes (40% of mammalian genes). Unlike CpG sites in the coding region of a gene, in most instances the CpG sites in CpG islands are unmethylated if genes are expressed. This observation led to the speculation that methylation of CpG sites in the promoter of a gene may inhibit the expression of a gene [Wikipedia].

Consider the DNA sequence in file `genome.txt`. By using tools from the **Sequence Manipulation Suite** (http://www.bioinformatics.org/sms2/) and **CpGPlot** available at the EMBL-EBI web site (https://www.ebi.ac.uk/Tools/seqstats/emboss_cpgplot/), characterize your genomic sequence and detect regions that are rich in the CpG pattern.

- Present and comment the results obtained from the following tools: CpGPlot and CpG Islands, DNA stats, and ORF Finder, in particular the influence on the parameters chosen in the overall results obtained.
- Compare the genomic sequence in file `genome.txt` with the ones available at the GenBank Nucleotide database (Hint: use NCBI BLAST). Compare the CpG islands (identified before) with the annotation of the most homologous sequence retrieved from the GenBank, in particular to which regions of the genome they correspond to. You can complement the analysis and interpretation by including references to papers of your choice.

Group II (4.5 points)

In this group, justify all the deductions made using the main properties of probabilities and relevant theory.

a) Consider a first order Markov Chain model with transition probabilities a_{st}. The sum of the probabilities of all possible sequences of states of length L can be written as follows:

$$\sum_{x} P(X) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_L} P(x_1) \prod_{i=2}^{L} a_{x_{i-1} x_i}$$

Show that this sum is equal to 1.

b) In a Hidden Markov Model (HMM), let $P(x,\pi)$ be the joint probability of an observed sequence x and a state sequence π. We define the most probable path π^* as:

$$\pi^* = \arg\max_{\pi} P(x, \pi)$$

Show that this definition is equivalent to:

$$\pi^* = \arg\max_{\pi} P(\pi|x)$$

c) Consider a first order Markov Chain Model with transition probabilities:
$a_{st} = P(X_{i+1} = t \mid X_i = s)$, where s and t belong to a given alphabet A. Show that the probability of a 2-step transition $a_{su} = P(X_{i+2} = u \mid X_i = s)$ is given by $a_{su} = \Sigma_{t \in A}\, a_{st}\, a_{tu}$.

Group III (12 points)

Formally, an HMM is defined by: an alphabet of emitted symbols; a set of k (hidden) states; a matrix of state transition probabilities and a matrix of emission probabilities. Consider an HMM model with three states, to identify DNA coding regions: **state 1** corresponds to the *Start Site signal*, **state 2** corresponds to an *Exon* region and **state 3** corresponds to an *Intron* region. Initial probabilities for all the three states are equal and transitions from all the states to an end state are also equal.

a) Using a graphical representation, detail this model considering the **transition probabilities**: $a_{11} = 0.6$; $a_{12} = 0.4$; $a_{22}=0.5$; $a_{21}=0.25$; $a_{23} = 0.25$; $a_{33}=0.5$; $a_{31}=0.25$; $a_{32}=0.25$ and the **emission probabilities**: $e_A = 0.4$; $e_T = e_G = 0.3$ and $e_C = 0$ for state 1; $e_A = e_T = 0.1$ and $e_C = e_G = 0.4$ for state 2; and $e_A = 0.4$; $e_T = 0.3$; $e_C = 0.3$; $e_G = 0$ for state 3.

b) Using the previous HMM and considering the DNA sequence ***S* = CATGCGGGTTATAAC**, build a program to compute the most probable sequence of states π that generate sequence S. Use the programming language of your choice.

Take into account that the program should also run adequately in long sequences (i.e., the implementation is robust to longer inputs).

Input:

A DNA sequence X

Output:

A path π^* that maximizes $P(X,\pi)$ over all possible paths π, for the given HMM model

What is π^* for sequence S? What does that mean?

c) Using the given HMM model and considering the previous DNA sequence S, compute the probability $P(S)$. Which algorithm should be used? Justify. Adapt the algorithm implemented in b) to compute this probability.

d) What are the posterior probabilities $P(\pi_2=k|x)$ and $P(\pi_9=k|x)$? Compare with π^* and interpret these results.