TÉCNICO LISBOA

COMPUTATIONAL BIOLOGY

Instituto Superior Técnico

2017/2018

BC_LAB# 4 - PROBABILISTIC MODELS

Group I (10 points)

CpG islands are regions of DNA characterized by a large number of adjacent cytosine and guanine nucletoides 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, retrieved Feb 2007).

1. Consider the DNA sequence in file *genome.tx*t. By using tools from the "Sequence Manipulation Suite (http://www.bioinformatics.org/sms2/)" and CpGPlot available at (http://www.ebi.ac.uk/Tools/emboss/cpgplot/index.html?) the EMBL-EBI web site, characterize your genomic sequence and detect regions that are rich in the CpG pattern.

   a. (6 points) Present and comment results from the following tools: CpGPlot and CpG Islands; DNA stats and ORF Finder.

   b. (2 points) Compare the genomic sequence in file *genome.tx*t with the ones available at the Genbank Nucleotide database.

   c. (2 points) Compare the CpG islands identification results with the annotation of the most homologous sequence retrieved from the Genbank.

Group II (10 points)

2. Formally, an HMM M is defined by: an alphabet of emitted symbols; a set of (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 **Intro** region. Initial probabilities for all the three states are equal and transitions from all the states to an end state are also equal.

   a) (1 point) Using a graphical representation, detail this model considering the transition probabilities: a11 = 0,6; a12 = 0,4; a22=0.5; a21=0.25; a23 = 0.25; a33=0.5; a31=0.25; a32=0.25 and the emission probabilities: eA = 0.4; eT = eG =0.3 and eC = 0 for state 1; eA = eT = 0,1 e eC = eG = 0,4 for state 2; and eA = 0.4; eT = 0,3; eC = 0.3; eG = 0 for state 3;

   b) (5 points) Using the previous HMM model and considering the DNA sequence X = CATGCGGGTTATAAC, build a program to compute the most probable sequence of states that generates sequence X. Use the programming language of your choice.

   Input:

   The sequence X generated by the HMM described in 2.

   Output:

   A path that maximizes P(X|π) over all possible paths π

   c) (1 point) Which algorithm should be used if we need to compute the probability of this sequence being generated by this model? Justify.

   d) (3 points) Using the previous HMM model and considering the previous DNA sequence X, compute the probability P(X). Adapt the algorithm developed in b) to compute this probability.