COMPUTER SCIENCE TRIPOS Part II – 2026 – Paper 8

3 Bioinformatics (pl219)

(*a*) A class of proteins recognises patterns of rich CG dinucleotides in genomes.

Similarly to the task of gene identification and transmembrane structure, you can use a Hidden Markov Model (HMM) to detect CG rich regions (islands) by exploiting different nucleotide compositions. Consider a two-state HMM with states $S = \{+, -\}$ ($+$ = inside CG island, $-$ = outside) and observations $\Sigma = \{A, C, G, T\}$. The parameters are as follows.

Initial probabilities: $\pi_+ = 0.2$, $\pi_- = 0.8$.

Transition probabilities: $a_{++} = 0.8$, $a_{+-} = 0.2$, $a_{-+} = 0.1$, $a_{--} = 0.9$.

Emission probabilities: $e_+(A, C, G, T) = (0.15, 0.35, 0.35, 0.15)$,
$e_-(A, C, G, T) = (0.30, 0.20, 0.20, 0.30)$.

(*i*) Discuss how you would use the Viterbi algorithm to find the most likely state sequence for observation $O = \text{CG}$. [5 marks]

(*ii*) Give initialisation, recursion, and backtracking steps. [6 marks]

(*b*) Consider a multialignment of DNA sequences from four species. You assume to know the relationship between the species and apply the Small Parsimony Problem using Sankoff's algorithm to each column of the multialignment with a cost matrix where match = 0, transition (A $\leftrightarrow$ G, C $\leftrightarrow$ T) = 1, transversion (A $\leftrightarrow$ C, A $\leftrightarrow$ T; G $\leftrightarrow$ C, G $\leftrightarrow$ T) = 2. As input, use a rooted tree with leaves A: `T`, B: `G`, C: `C`, D: `G`, where X is parent of (A,B) and Y is parent of (C,D), and root R is parent of (X,Y).

(*i*) Find the optimal assignment and total parsimony score of the tree. [4 marks]

(*ii*) Discuss how to find alternative trees and then test their robustness. [5 marks]