2 Bioinformatics (pl219)

(*a*) Compute the nearest neighbour phylogeny from the four species (B,M,H,O) distance matrix.

$$\begin{pmatrix} & B & M & H & O \\ B & 0 & 5 & 6 & 4 \\ M & 5 & 0 & 3 & 2 \\ H & 6 & 3 & 0 & 2 \\ O & 4 & 2 & 2 & 0 \end{pmatrix}$$

[6 marks]

(*b*) Can we always build a phylogenetic tree from a distance matrix? [2 marks]

(*c*) Derive the Burrows-Wheeler (BWT) transform of the string 'TAGTATA'. How can the transform be reversed? Comment on the use of BWT for a genome sequence that has many repeated substrings. [4 marks]

(*d*) Three analysis techinques for gene expression data (microarray) are hierarchical clustering, k-means and Markov clustering. Describe the structure of a set of experimental results that could be analysed by all three techniques and state what each form of analysis might identify and any additional inputs required. [4 marks]

(*e*) Discuss how a Hidden Markov Model can be used to identify different gene parts and how many sequences might be needed to compute reliable transition probabilities. [4 marks]