2008 Paper 9 Question 1

Bioinformatics

(a) A long DNA sequence is used as a training set for parameter estimation of the DNA statistical model. The observed counts of sixteen dinucleotides N_{XY} are as follows:

$$\begin{pmatrix} & T & C & A & G \\ T & 306 & 228 & 126 & 114 \\ C & 144 & 102 & 216 & 138 \\ A & 222 & 120 & 132 & 126 \\ G & 114 & 102 & 132 & 132 \end{pmatrix}$$

Calculate:

(i) the transition probabilities P_{TT} and P_{AG} of the first-order Markov model of the DNA sequence; [3 marks]

(ii) the transition probabilities P_{TT} and P_{AG} of the first-order Markov model of the DNA sequence complementary to the given sequence. [3 marks]

(b) Build the tree from the following distance matrix between species A, B, C, D using the UPGMA (Unweighted Pair Group Method using arithmetic Averages) method. [7 marks]

	A	B	C	D
A		0.26	0.34	0.29
B			0.42	0.44
C				0.44
D				

(c) Describe how you would build a hidden Markov model (HMM) to predict protein secondary structure. [7 marks]