COMPUTER SCIENCE TRIPOS Part II – 2022 – Paper 8

2   Bioinformatics (pl219)

(a) What is the purpose of the boostrap approach in general and how can it be applied to phylogenetic trees? Using at least one numeric example, discuss how to interpret bootstrap values. [4 marks]

(b) What are the reasons for using progressive alignment in a multi-sequence alignment problem? Give the complexity of the various stages of the procedure and the overall complexity. [4 marks]

(c) Define the role of a scoring matrix in a matching algorithm and explain how it should be designed. [3 marks]

(d) Sketch the suffix tree for the genome GCTATA$. Give the time and space complexities of using a suffix tree for genome sequence assembly. Comment on finding repeated sequences. [5 marks]

(e) We often use Hidden Markov Models to predict genes, exons or introns. Outline how a Hidden Markov Model can be used as a binary classifier in such an application. What metrics can be used to evaluate its performance? [4 marks]