COMPUTER SCIENCE TRIPOS Part II – 2026 – Paper 9

3 Bioinformatics (pl219)

(*a*) To assess sequence similarity, we compute the alignment between two DNA sequences.

(*i*) Compute the local alignment of the sequences GGTTATA, TATAGG with the following rules: match score = +4, mismatch = −3, gap penalty = −4. Discuss how the alignment depends on the value of match scores, mismatch and gap penalty. [5 marks]

(*ii*) Explain how DNA stores information in its natural biological role and how it could be used as a potential medium for artificial data storage. [4 marks]

(*b*) You conduct a self-experiment with two phases to investigate the effect of diet on gut microbiome diversity. In Phase 1, you follow a strict ketogenic diet (high fat, very low carbohydrate) for 3 weeks. In Phase 2, you switch to eating exclusively at McDonald's (high carbohydrate, processed food, low fibre) for 3 weeks. At the end of each phase, you collect a fecal sample and send it to a sequencing company for 16S rRNA sequencing. The company returns a list of bacterial species identified in each sample.

(*i*) Explain how you would compare bacterial composition between the two dietary phases. [5 marks]

(*ii*) Discuss the main limitations of this experimental design. [1 mark]

(*c*) Lloyd's algorithm (k-means) requires the number of clusters k to be specified in advance. Describe how you would choose an appropriate value of k. [5 marks]