ENGINEERING TRIPOS PART IIB

Wednesday 11 May 2005 2.30 to 4

Module 4M8

BIOINFORMATICS

*Answer not more than **two** questions.*

All questions carry the same number of marks.

*The **approximate** percentage of marks allocated to each part of a question is indicated in the right margin.*

There are no attachments.

You may not start to read the questions printed on the subsequent pages of this question paper until instructed that you may do so by the Invigilator

(TURN OVER

1 You are given two nucleotide sequences:

seq1: CGCAT

seq2: ATCACT

with the linear gap penalty -3 and the transition scoring matrix

	A	C	G	T
A	4	−2	1	−2
C	−2	4	−2	1
G	1	−2	4	−2
T	−2	1	−2	4

(a) What is/are the global alignment(s) of these two sequences? [20%]

(b) What is/are the score(s)? [20%]

(c) Define the edit distance between two sequences. [20%]

(d) Describe the heuristic used by Clustal. [20%]

(e) Describe briefly the main differences between the following phylogenetic methods:

(i) neighbour joining;

(ii) maximum likelihood;

(iii) maximum parsimony.

[20%]

2 (a) Describe the main uses of unsupervised learning for microarray data. Give examples to support your argument. [25%]

(b) Compare the hierarchical and partitioning methods of unsupervised learning. What are their advantages and disadvantages? How would you decide when to use a hierarchical or partitioning method? [35%]

(c) Describe the K-means algorithm for clustering input samples. Include a description of how centroids are initialised and moved, and how to test for convergence of the algorithm. [25%]

(d) Describe a method for optimising the number of centroids, K, in the K-means algorithm. How does this method ensure that the number of centroids is neither too small nor too large? [15%]

3 Consider a system with stochastic reaction events $x \xrightarrow{\lambda} x+1$ and $x \xrightarrow{\beta x^3} x-3$.

(a) Write down the corresponding Markov process for the probability $p(x)$. [20%]

(b) Mathematically motivate the exact Gillespie algorithm for generating sample paths for the system. Write a short program in mock code (pseudocode) for the central part. [20%]

(c) Give the exact differential equation for the average, $\langle x \rangle$. Approximate the equation, expressing $\frac{d\langle x \rangle}{dt}$ in terms of $\langle x \rangle$, assuming that fluctuations are negligible. [20%]

(d) Using the approximation from (c) at steady state, calculate the average number of molecules $\langle x \rangle$, the elasticity H, the average lifetime τ, and the average chemical event size $\langle r \rangle$ (averaged over fluxes). Formulate the answers in terms of λ and β wherever possible. [20%]

(e) Use the normalized stationary Fluctuation Dissipation Theorem $M\eta + \eta M^T = D$, where $M = H/\tau$ and $D = \frac{2\langle r \rangle}{\tau \langle x \rangle}$, to calculate $\eta = \frac{\sigma^2}{\langle x \rangle}$, where σ^2 is the variance of x. Compare the answer with what you get by changing $x \xrightarrow{\beta x^3} x-3$ to $x \xrightarrow{\beta x} x-1$. Explain the result. [20%]

END OF PAPER