ENGINEERING TRIPOS PART IIB

Thursday 3 May 2007 2.30 to 4

Module 4G1

COMPUTATIONAL AND SYSTEMS BIOLOGY

*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.

STATIONERY REQUIREMENTS
Single-sided script paper

SPECIAL REQUIREMENTS
Engineering Data Book
CUED approved calculator allowed

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

1 (a) Discuss, with examples, the difference between the edit distance and the Hamming distance. [20%]

(b) Do repeats in vertebrate genomes make the alignment task easier or more difficult? [10%]

(c) Describe the local alignment algorithm. Illustrate your answer with an example consisting of 2 sequences with match, mismatch and gap penalties. [30%]

(d) Discuss the difference between the Fitch and Sankoff parsimony algorithms. Is evolution necessarily parsimonious? [20%]

(e) How can a distance matrix be used to generate a phylogenetic tree? [20%]

2 (a) Explain how a microarray measures gene expression levels. In your answer, include brief definitions of the following terms:

(i) reverse transcription;

(ii) probe;

(iii) target;

(iv) hybridisation;

(v) control spots;

(vi) two-colour vs. single-channel arrays.

[40%]

(b) Describe, with a simple example, the rank products technique for detecting differentially expressed genes. How is significance of the ranking assessed? [30%]

(c) What is hierarchical clustering? Draw a sketch to illustrate your answer. [10%]

(d) Describe, with examples, two ways in which hierarchical clustering can be applied to the study of gene expression data, depending on how samples are taken from the gene expression matrix. [20%]

3 (a) Describe, with the aid of diagrams, the structure, function and regulation of the *lac* operon. Identify as many positive and negative feedback loops as you can. [45%]

(b) Assume that, in the absence of glucose, expression of *lac* is governed by the reactions

$$O + R \underset{k_{1-}}{\overset{k_{1+}}{\rightleftharpoons}} (OR)$$

and

$$2I + R \underset{k_{2-}}{\overset{k_{2+}}{\rightleftharpoons}} (RI_2)$$

where O is the operator, R the repressor and I the inducer. Suggest a suitable value for k_{1-}, and show that

$$\mathrm{P(ON)} \approx \frac{1 + k_2[I]^2}{1 + k_1 R_t + k_2[I]^2}$$

where $k_1 = \frac{k_{1+}}{k_{1-}}$, $k_2 = \frac{k_{2+}}{k_{2-}}$, R_t is the total concentration of repressor molecules and P(ON) is the probability that the the operator is not bound. [40%]

(c) If the rate of transcription of the *lac* operon, for constant inducer concentration, was assumed to be a constant proportional to the P(ON) derived in part (b), then what would be the relationship between the mean and the variance of the number of mRNA molecules? How realistic is this assumption? [15%]

END OF PAPER