UNIVERSITY OF CAMBRIDGE

M. Phil. in COMPUTATIONAL BIOLOGY

Friday, 14 May, 2010 2:00 pm to 4:00 pm

COMPUTATIONAL BIOLOGY

*Attempt **ALL** questions.*

*There are **THREE** questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1 Disease Dynamics

(a) Write down the standard SIR model for the spread of an infectious disease, explaining what all of your variables and parameters represent. Host demographics need not be included. Sketch an output from your model, showing a typical epidemic. [30%]

(b) Explain what R_0 means biologically, and give it in terms of your parameters. [10%]

(c) The "Force of Infection" is the rate at which one susceptible individual becomes infected. What is it in terms of your parameters and variables? [10%]

(d) The basic SIR model is for homogeneous mixing, but in practice we often want to consider the spatial structure of populations. Several different ways to extend the SIR model spatially were given in lectures. Choose one of these, and either give new equations for the spatial SIR model, or describe an algorithm to implement the system computationally. [50%]

2 Functional Genomics

A scientist is interested in the changes on gene expression in prostate tumours in old males introduced by the consumption of coffee and red wine. A clinical trial was performed and RNA samples were extracted from biopsies of prostate tumours of: patients who had drunk no coffee nor red wine for a month (Control), patients who had drunk two *espressos* a day but no red wine for a month (Coffee), and patients who had drunk two glasses of red wine a day but no coffee for a month (Wine). The samples were hybridised against a total of six two-colour microarrays. The following table summarises the experimental design, specifying the sample types hybridised against each array and the dyes used in labelling each sample:

Array #	Cy3 (green)	Cy5 (red)
1	Control	Coffee
2	Coffee	Control
3	Wine	Coffee
4	Coffee	Wine
5	Wine	Coffee
6	Control	Wine

(a) Draw a diagram of the layout of this experimental design, using conventional arrows to represent the two-colour arrays. [15%]

(b) Determine the design matrix, taking the coffee effect and the wine effect as your independent parameters. [20%]

(c) Determine which effect this design would allow you to estimate with higher precision. Explain the caveats of these precision estimates by discussing the concept of effective replication. [35%]

Hint: The inverse of a 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is $A^{-1} = \dfrac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.

(d) Explain the difficulties of judging statistical significance and the merits of a moderated t-statistic, in particular the empirical Bayesian log odds, for assessing differential expression in microarray experiments. [30%]

3 Hidden Markov Models

Consider a HMM where $\{X_0, \ldots, X_N\}$ is a homogeneous Markov chain on the space of hidden states $S = \{s_1, \ldots, s_K\}$, and $Y_0^N = \{Y_0, \ldots, Y_N\}$ is the sequence of observed variables drawn from the states $V = \{v_1, \ldots, v_J\}$.

(a) Given that

$$\alpha_n(i) \;=\; P(Y_0^n, X_n = s_i)$$

$$\beta_n(i) \;=\; P(Y_{n+1}^N | X_n = s_i)$$

show that

$$P(X_n = s_i | Y) \;=\; \frac{\alpha_n(i)\beta_n(i)}{\sum_j \alpha_n(j)\beta_n(j)}\,.$$

How can this result be used in the context of posterior max decoding of the sequence of hidden states corresponding to some observed sequence Y_0^N? [40 %]

(b) The Viterbi algorithm involves storing values of the following variable:

$$\psi_n(i) \;=\; \underset{i_{n-1}}{\text{argmax}}\; A_{i_{n-1}i}\,\delta_{n-1}(i_{n-1})$$

How many such values are there? How is $\delta_{n-1}(i_{n-1})$ defined?

If each value $\psi_n(i)$ can be stored in one byte of memory, and $K = 10$, what is the longest Viterbi sequence we can calculate on a machine with 2 Gb of memory (assuming the whole calculation is done in memory)? [40%]

(c) Outline how the Viterbi algorithm can be used to infer model parameters given a sequence of observed variables. [20%]

END OF PAPER