UNIVERSITY OF CAMBRIDGE

M. Phil. in COMPUTATIONAL BIOLOGY

Friday, 15 May, 2009 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

SPECIAL REQUIREMENTS
Genetic code table

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

1 Computational Neuroscience

(a) Given an input vector of length N with elements x_j^μ (the superscript denotes the pattern number and the subscript indicates the index of the vector), and a weight vector with elements w_j, a perceptron calculates its output using:

$$y^\mu = g(\sum_{j=1}^{N} w_j x_j^\mu)$$

where $g(\cdot)$ is an activation function. Define:

$$E = \frac{1}{2}(t^\mu - y^\mu)^2 + \beta \sum_{i=1}^{N} w_i^2$$

Explain briefly the roles of t^μ and β in the above expression and therefore what E measures.

(b) Two neurons are reciprocally connected by weights of strength w (assume $w < 0$) and receive external input i. The firing rates of the two neurons u_1, u_2 evolve according to:

$$\tau \frac{\mathrm{d}u_1}{\mathrm{d}t} = -u_1 + f(wu_2 + i)$$

$$\tau \frac{\mathrm{d}u_2}{\mathrm{d}t} = -u_2 + f(wu_1 + i)$$

$$\text{where} \quad f(x) = \begin{cases} x & x \geqslant 0 \\ 0 & x < 0 \end{cases}$$

When $w = -2, i = 1$, draw the phase space, along with the nullclines. Find the position of any steady-states, and draw flow lines to assess their stability. Interpret briefly the typical behaviour of the system.

(c) Assuming $f(x) = x$, rewrite the system in terms of a weight matrix $\mathbf{W}$, a state vector $\mathbf{u}$ and an input vector $\mathbf{i}$. Find the steady-state. For what range of values of w is the steady-state stable?

2 Genome Informatics

A genetic code table is provided as Supplementary Material.

2.1

The comparative sequencing of species has great utility in genome annotation.

(a) Give three (3) situations where a closely related genome may help to improve gene models.

(b) What are the caveats of considering too closely related and/or too few species in the annotation of regulatory sequences?

2.2

A specific single nucleotide polymorphism (SNP) can be detected in 2% of the population. Their polymorphism is G → A in the sequence context (the SNP is underlined):

5'-CAGCAT G GAATG-3'

(a) Describe the possible functional consequences of this SNP, depending on whether it is in a regulatory region or a protein coding context.

(b) Individuals that are homozygous for the SNP have never been observed, even in large samples. Suggest why this might be, and hence identify which of the possible consequences in Q2.2(a) is most likely.

2.3

Imagine yourself in the following scenario: You were given short sequence reads of plant RNA obtained from a next-generation sequencing machine (fragments of 20–30 nucleotides in length). You attempt to map them back to the genome, but a significant proportion of them do not align.

(a) Give three (3) obvious explanations why the alignment of short sequences can fail, apart from possible contamination or technical difficulties during the preparation of the RNA.

(b) There are some indications that the problematic sequences come from an uncharacterised plant RNA virus. What would you do next? What are the specific caveats with short sequence reads?

3 Hidden Markov Models

In a Hidden Markov Model with hidden variables X_0^N ($\equiv X_0, X_1, \ldots X_N$), hidden states $\{s_1, \ldots s_k\}$ and emission variables Y_0^N, the forward variable is defined as $\alpha_n(i) = P(Y_0^n, X_n = s_i)$, and satisfies a recursion relation which can be written as

$$\alpha_{n+1}(i) = \sum_{j=1}^{K} \alpha_n(j) b_i(v_{n+1}) A_{ji}$$

(i) What do A_{ji}, v_{n+1} and $b_i(v_{n+1})$ represent here?

(ii) If there are K hidden states and N elements in the observed sequence, roughly how many operations will be needed to calculate the likelihood $P(Y_0^N) = \sum_i \alpha_N(i)$ using this recursion?

(iii) If $c_n = P(Y_n \mid Y_0^{n-1})$, show that the scaled variable $\hat{\alpha}_n(i) = P(X_n = s_i \mid Y_0^n)$ satisfies

$$\hat{\alpha}_n(i) = \frac{\alpha_n(i)}{\prod_{m=0}^{n} c_m}$$

where $c_0 \equiv P(Y_0)$.

Hence derive a recursion relation for $\hat{\alpha}_n(i)$.

(iv) In terms of the probabilities they represent, explain why $\alpha_n(i)$ will tend to zero as n increases, while $\hat{\alpha}_n(i)$ should not. Why is this important computationally?

END OF PAPER