UNIVERSITY OF
CAMBRIDGE

M. PHIL. IN COMPUTATIONAL BIOLOGY

Friday, 13 May, 2011   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
PAM250 Matrix

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

1 (a) Sean Eddy once wrote "*Back in the good old days, so many things were easier to understand. ... The first sequence comparisons just assigned -1 per mismatch and -1 per insertion/deletion, and if you didn't like that, you could make up whatever scores you thought gave you better-looking alignments. Those days are gone ...*" Modern sequence alignment methods use alignment scores from a substitution matrix. Have a look at the accompanying PAM250 matrix for protein sequence alignments.

i. Describe the structure of the matrix and how it is being read.

ii. How are the scores derived (*not* used; rough principle is enough)?

iii. Give a biological interpretation of the scores.

iv. Why are cysteine (C) and the aromatic amino acids (F, Y, W) associated with large absolute numbers?

(b) Gene finding is a key task in the analysis of novel genome sequences.

i. Which characteristics can be used to detect protein-coding genes?

ii. Describe a computational identification strategy for each of two different gene classes (e.g. protein-coding genes and microRNA genes).

(c) Restriction mapping was often employed in conjunction with DNA sequencing before whole-genome shotgun strategies became feasible. Though following very simple principles, it can pose significant computational problems.

i. Describe a set of molecular biology experiments involving enzymes A and B (to cut exactly once at x_a, x_b) to infer the order of the three possible DNA fragments.

ii. Consider a situation in which the enzymes cut more than once. Why and how does this cause a problem?

2 (a) Consider representing a genome by a sequence of independent identically-distributed random variables, in which the base frequencies are $\pi_{\mathtt{A}} = 0.18$, $\pi_{\mathtt{C}} = 0.3$, $\pi_{\mathtt{G}} = 0.3$, $\pi_{\mathtt{T}} = 0.22$. If X_1, X_2 are two variables in the sequence, what are the probabilities of the following events:

- $(X_1 = \mathtt{T}, X_2 = \mathtt{C})$
- One of X_1 or X_2 (but not both) is G.
- $(X_1 = \mathtt{A} | X_1 \text{is a purine})$

(b) For three events A, B and C show that $P(A, B|C) = P(A|B, C)P(B|C)$.

(c) Consider a Markov chain with state space $S = \{s_1, \ldots, s_K\}$. Show that the distribution of waiting times in state s_i is Geometric. What does this imply for models built using Markov chains?

(d) How can we use a continuous-time Markov chain to model the evolution of a genome sequence? Explain how the rate matrix Q is defined, and how many parameters it has in its most general form. What simplifying assumptions can we make to reduce the number of parameters?

3 (a) Describe the architecture of a Hopfield network. You should describe how the weights w_{ij} between two nodes i and j are calculated, and how the activation v_i of node i is updated. [30 %]

(b) Draw a two-node network with $w_{12} \neq w_{21}$. Using an asynchronous update rule, demonstrate an example network state where the node activations do not converge to stable values. [10%]

(c) Draw a two-node network with $w_{12} = w_{21}$. Using a synchronous update rule, demonstrate an example network state where the node activations do not converge to stable values. [10%]

(d) Show that the following energy function for the Hopfield network (with symmetric weights and asynchronous update rule) always decreases a minimum. [20%]

$$E = -1/2 \sum_{i,j} w_{ij} v_i v_j$$

(e) Describe the key features of the 'dynamic clamp' technique, and how you would use it to study an individual neuron. How would you use it to connect two neurons into a virtual network, and what kinds of manipulations would be useful to perform with such a virtual network? [30%]

END OF PAPER