UNIVERSITY OF CAMBRIDGE

M. Phil. in COMPUTATIONAL BIOLOGY

Friday, 11 May, 2012 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
None

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

1 Sequence Analysis

(i) [15%] If a fair dice is rolled n times, what is the probability of rolling a six only on the nth roll?

(ii) [20%] Define conditional independence. For events A, B, C and D, if A is conditionally independent of B, show that $P(A|B,C,D) = P(A|C,D)$.

(iii) [20%] What is a first-order discrete Markov Chain?

(iv) [25%] Sketch a HMM for global pairwise alignment of two sequences. Describe what its states and transition probabilities represent and how the transition matrix is parameterised.

(v) [20%] How many rooted and unrooted trees are there with n leaves? How many internal nodes does a tree with n leaf nodes have?

2 Network Biology

(i) [10%] Name two examples of experimental methods to perturb genes and three examples of phenotypes that are commonly observed.

(ii) In the lectures we have discussed different methods for network reconstruction from gene perturbation data, in particular Bayesian networks (BNs) and Nested Effect Models (NEMs).

a. [25%] Define BNs and NEMs. Draw a small schematic showing how the different parts of the definition come together to form the model.

b. [25%] Contrast BNs and NEMs using the criteria:

- Type of graph,
- Probabilities,
- Which underlying concepts are encoded in both models?
- How are single gene perturbations modelled?
- Area of application: When would you use BNs, when NEMs?

(iii) Explain how hill-climbing aims at finding an optimal model.

a. [10%] Describe step-by-step how hill-climbing works.

b. [10%] What key problem can hill-climbing run into? How would you address it? (A small graphic might help)

c. [20%] Why is hill-climbing in BNs simpler than in NEMs? Describe a heuristic method to infer NEMs.

3 Structural Biology

(i) With more than 65,000 protein x-ray structures deposited in the pdb, the structure of a particular protein is often represented by more than one entry, and for many, these entries belong to different crystal forms.

a. [16%] List four criteria that can be used to validate the quality of a crystal structure.

b. [16%] What does "different crystal form" refer to and what is the advantage over only one crystal form being available for a given protein.

(ii) Proteins often function as oligomers, rather than in the monomeric state.

a. [16%] Why is it common to find only the coordinates of one protein chain in the pdb file if the protein is known to exist as a specific oligomer in solution?

b. [16%] What types of interactions would you expect to find at the protein-protein interfaces within protein complexes?

(iii) More than half of all proteins interact with membranes. Of these proteins, integral membrane proteins (IMPs) are permanently bound to the core of the lipid bilayer, and carry out a range of important functions.

a. [18%] From the available high-resolution structures of IMPs, what are the known structural classes? Why must IMPs adopt one of these structural folds in a membrane environment?

b. [18%] There is a characteristic distribution of amino acid types on the surface of IMPs. What types of amino acids are found on the surface of IMPs, and with which regions of the lipid bilayer do they interact?

END OF PAPER