UNIVERSITY OF CAMBRIDGE

M. Phil. in COMPUTATIONAL BIOLOGY

Friday, 10 May, 2013 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
Calculator - students are permitted to bring an approved calculator.

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

1 Genome Bioinformatics

1. The sequencing of entire genomes is no longer just possible for large international consortia, but has become accessible even for smaller individual research groups.

   (a) What is shotgun sequencing?

   (b) Why does next-generation sequencing lend itself to this strategy especially when closely related genome sequences are already available?

   (c) What are the basic steps to producing a genome assembly?

   (d) Once the genome is assembled, what are the next important steps in genome annotation and what is to be discovered?

2. Transcription factors are proteins that bind to DNA. A naïve but often-communicated concept is that these proteins bind to a 'preferred DNA word'.

   (a) Explain why this concept is flawed, i.e. why it is not a 'word' that is being recognised? Try to use vocabulary you remember from the Structural Biology module wherever suitable.

   (b) Explain the difference between a position frequency matrix and a position weight matrix.

   (c) What is a scoring function in the context of position weight matrices?

2 Population Genetic Analyses of Genomic Data

Complete both parts A and B of the question.

A1

$$\frac{dq_i^1}{dt} = \sigma q_i^1(1 - q_i^1) + \mu(1 - 2q_i^1) + \chi_t(q_i^1) \tag{1}$$

Equation 1 describes the evolution of the allele frequency of a single-locus two-allele system (alleles labelled 0 and 1), the term q_i^1 denoting the fraction of allele 1 in the population. The noise term $\chi_t(q_i^1)$ has mean equal to zero, variance equal to $q_i^1(1 - q_i^1)/N$, and is uncorrelated in time.

What do σ, μ and N denote?

A2 Supposing that $\mu = 0$ and neglecting the noise term gives the simpler equation

$$\frac{dq_i^1}{dt} = \sigma q_i^1(1 - q_i^1). \tag{2}$$

Assume that $\sigma > 0$ and that the population starts at frequency $q_i^1(0) = 0.01$. Sketch how the system evolves over time. What happens if you increase σ?

B Consider next a system with two loci, i and j, each with two alleles, 1 and 0. The frequency of the allele a at locus k is denoted by q_k^a; for example, q_i^1 is the frequency of the allele 1 at locus i. The frequency of individuals with the allele a at locus i and allele b at locus j is denoted q_{ij}^{ab}, where a and b can be either 1 or 0. Throughout this part, assume that the population size is very large, and that there is no mutation.

B1 Suppose that, at a given time, $q_i^1 = 0.6$, $q_j^0 = 0.5$ and $q_{ij}^{00} = 0.3$. Calculate the values of q_{ij}^{10}, q_{ij}^{01}, and q_{ij}^{11}, and find the value of the linkage disequilibrium term D_{ij}.

B2 Suppose that the system undergoes one round of random mating, involving recombination. During mating, the probability of a recombination event occurring between i and j is equal to one half. Assuming that no more than one recombination event can happen in a single genome during mating, calculate the values of q_{ij}^{00}, q_{ij}^{10}, q_{ij}^{01}, q_{ij}^{11}, and D_{ij} following the mating event.

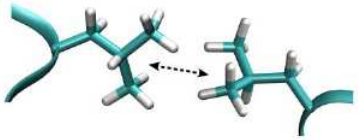
B3 Suppose that, following the random mating, there is a rapid change in the environment, creating positive selection for the allele 1 at locus i. Over time, the frequency of this allele increases to a final value of 1. Assuming that the fitness of an individual is dependent only upon the allele at locus i, and assuming that no further recombination has occurred, what is the final frequency of the allele 1 at locus j?

B4 Suppose that, in question 3, selection had acted with equal strength both for the allele 1 at locus i, and for the allele 0 at locus j. Assuming that the fitness of an individual is the sum of the fitness effects at loci i and j, and that no further recombination takes place, what is the initial rate and direction of change in the frequency q_{ij}^{11}? Describe, in terms of frequencies, the final state of the system.

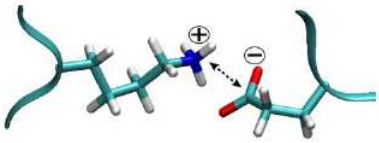
3 Biomolecular Simulations

(i) [30%] A molecular dynamics (MD) simulation force field is composed of terms that describe bonded and non-bonded interaction energies between pairs of atoms. Three examples of typical interactions in biomolecular systems are shown below. Which of them could most appropriately be represented by (a) a Coulombic potential; (b) a harmonic potential; and (c) the Lennard-Jones (6-12) potential?

A bonded hydrogen and oxygen atom within a water molecule.



A pair of hydrophobic isoleucine side chains.



An arginine-glutamate salt bridge.

(ii) [20%] An MD simulation may be used to investigate the functional motions of a solvated protein, on the basis of a static experimental structure derived from X-ray crystallography. Before running such a simulation, it is first necessary to perform energy minimization to remove any large forces in the system. Give two possible sources of such "large forces".

(iii) The principles of statistical mechanics make molecular simulations possible, by establishing a link between macroscopic properties and the behaviour of the individual atoms that make up the simulation system or "ensemble".

a. [30%] State three variables that may typically be controlled to achieve a particular simulation ensemble.

b. [10%] What combination of such variables would you choose to simulate a biological system and why?

c. [10%] It is normally necessary to carry out "equilibration" simulations to reach the desired ensemble, prior to production MD. Practically, how might this be achieved?

END OF PAPER