UNIVERSITY OF CAMBRIDGE

M. PHIL. IN COMPUTATIONAL BIOLOGY

Friday 16 May 2008   2.00 to 4.00

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 Statistical Genetics

(a) Conditional probabilities. A screening test has a probability 0.9 of being positive in true cases of a disease and a probability 0.995 of being negative in people without the disease. The prevalence of the disease is 0.001. What is the probability of a person having the disease given that they have a positive test? [34%]

(b) Hardy-Weinberg Equilibrium (HWE). At a given locus the counts of the three genotypes of a diallelic marker are (n_0, n_1, n_2). We assume that the population is in HWE at this locus, with frequencies $((1-p)^2, 2p(1-p), p^2)$.

1. Write the log-likelihood of the data (only consider the terms that involve p).
2. Take the derivative w.r.t. p and find the MLE $\hat{p}$.
3. Give a simple interpretation of the result. [33%]

(c) Bayes Theorem and GWAS. In a classic framework, the measure of significance in a GWAS is the p-value α. In a Bayesian situation, we rather use the posterior odds:

$$\begin{aligned}\frac{\mathbb{P}(\mathbb{H}_1|p \leqslant \alpha)}{\mathbb{P}(\mathbb{H}_0|p \leqslant \alpha)} &= \frac{\mathbb{P}(p \leqslant \alpha|\mathbb{H}_1)}{\mathbb{P}(p \leqslant \alpha|\mathbb{H}_0)}\frac{\mathbb{P}(\mathbb{H}_1)}{\mathbb{P}(\mathbb{H}_0)} \\ &= \frac{\text{power}}{\alpha} \times \text{prior odds}\end{aligned}$$

Comment on how the Bayesian formulation of this problem modifies the interpretation of significance. What does it mean in practice when one wants to set a threshold for the significance level in a GWAS? [33%]

2 Disease Dynamics

Here is part of an abstract from a research paper:

Bubonic plague is generally thought of as a historical disease; however, it is still responsible for many deaths each year worldwide. This paper develops a model for bubonic plague that encompasses the disease dynamics in rat, flea and human populations. Data consisting of daily human cases are used from three different outbreaks. Some key variables of the deterministic model, including the force of infection to humans, are shown to be robust to changes in the basic parameters, although variation in the flea searching efficiency, and the movement rates of rats and fleas will be considered throughout the paper. The stochastic behaviour of the corresponding metapopulation model is discussed. Short-lived local epidemics in rats govern the invasion of the disease and produce an irregular pattern of human cases similar to those observed. However, the endemic behaviour in a few rat subpopulations allows the disease to persist for many years. This spatial stochastic model is also used to identify the criteria for the spread to human populations in terms of the rat density.

(a) Explain what is meant by *stochastic* and *deterministic* in the context of disease dynamics. Explain why sometimes it is appropriate to choose a stochastic approach, and why sometimes a deterministic approach would be more suitable. [20%]

(b) The following terms appear in the abstract. Explain what is meant by each of them:

"Force of infection"

"Metapopulation model"

"Endemic behaviour" [30%]

(c) There are many models in the literature that have been developed for specific diseases. Explain some of the possible purposes of developing these models. Illustrate your answers with what you might expect to find in a paper based on the abstract above. [50%]

3 Methods and Models in Genomics

(a) What are the statistical and algorithmic differences between PatternHunter, BLAST, Smith-Waterman and Needleman-Wunsch algorithms? [30%]

(b) Discuss the properties and assumptions of Jukes-Cantor, Kimura 2 and Felsenstein models of DNA evolution. Your answer should include some discussion of the degrees of freedom in each model and how each model is derived. [40%]

(c) Describe an algorithm to reconstruct a genetic network from microarray perturbation data. [30%]

END OF PAPER