UNIVERSITY OF LONDON
IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2003

BEng Honours Degree in Computing Part III
MSc in Computing Science
BEng Honours Degree in Information Systems Engineering Part III
MEng Honours Degree in Information Systems Engineering Part III
BSc Honours Degree in Mathematics and Computer Science Part III
MSci Honours Degree in Mathematics and Computer Science Part III
for Internal Students of the Imperial College of Science, Technology and Medicine

This paper is also taken for the relevant examinations for the Associateship of the City and Guilds of London Institute
This paper is also taken for the relevant examinations for the Associateship of the Royal College of Science

PAPER C341=I3.36

INTRODUCTION TO BIOINFORMATICS

Tuesday 29 April 2003, 14:30
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators required

Partial BLOSUM-62 Matrix:

	N	D	E	Q	M	I	L	V
N	6	1	0	0	-2	-3	-3	-3
D	1	6	2	0	-3	-3	-4	-3
E	0	2	5	2	-2	-3	-3	-2
Q	0	0	2	5	0	-3	-2	-2
M	-2	-3	-2	0	5	1	2	1
I	-3	-3	-3	-3	1	4	2	3
L	-3	-4	-3	-2	2	2	4	1
V	-3	-3	-2	-2	1	3	1	4

1a i) Roughly how many base letters are there in a gene? Explain why the number of letters in the residue sequence of a gene is a third of the number in the base sequence.

ii) Name the three ways in which random mutations can occur when a gene evolves. Suggest a way in which base sequence 2 might have evolved via random mutation from base sequence 1 below.

Sequence 1: `GATCATTGA`

Sequence 2: `GATTATGA`

iii) Explain why such random mutations are nearly always deleterious (harmful) to the organism inheriting the mutated gene.

b i) In the dynamic programming diagram below, what do the numbers in the brackets signify? Carefully copy out the diagram and use the partial BLOSUM-62 matrix given above and the Needleman-Wunsch algorithm to fill in the rest of the diagram.

	Gap	N	D	E	V
Gap	0	-6	-12	-18	-24
L	-6	-3 (-3)	-9 (-4)	-15 (-3)	-17 (1)
N	-12	0 (6)	-2 (1)	-8 (0)	
D	-18	-6 (1)	6 (6)		
V	-24	-12 (-3)			

Use the completed diagram to write down the best global alignment(s) for the sequences `NDEV` and `LNDV`. What is the BLOSUM-62 score for the alignment(s)?

ii) How is the Needleman-Wunsch algorithm altered to give the best <u>local</u> alignment of two sequences?

c The compositional complexity, K, of a sequence is given by this formula:

$$K = \frac{1}{L} log_{20} \left(\frac{L!}{\prod_{i=1}^{20} n_i!} \right)$$

Where L is the length of the sequence and n_i is the number of times residue number i is seen in the sequence.

i) Use this formula to calculate the compositional complexity of this residue sequence:

NNNDQMQMQI

[Remember that $\log_x(y) = \ln(y)/\ln(x)$].

ii) Generate a high scoring pair by extending the alignment in the high scoring triple below to the left and to the right. Use the BLOSUM-62 scoring scheme for this (matrix given above) and report the score for the resulting high-scoring pair.

```
N N N D Q M Q M Q I
M D N D M I Q Q I I
```

iii) The calculation of compositional complexity and the extension of alignment of regions form part of which algorithm? For what is this algorithm used in practice? Give an overview of the separate parts of this algorithm.

The three parts carry, respectively, 25%, 40% and 35% of the marks.

2a i) What does a phylogenetic tree indicate?

ii) Calculate the matrix of genetic distances for the four sequences given in the multiple sequence alignment below.

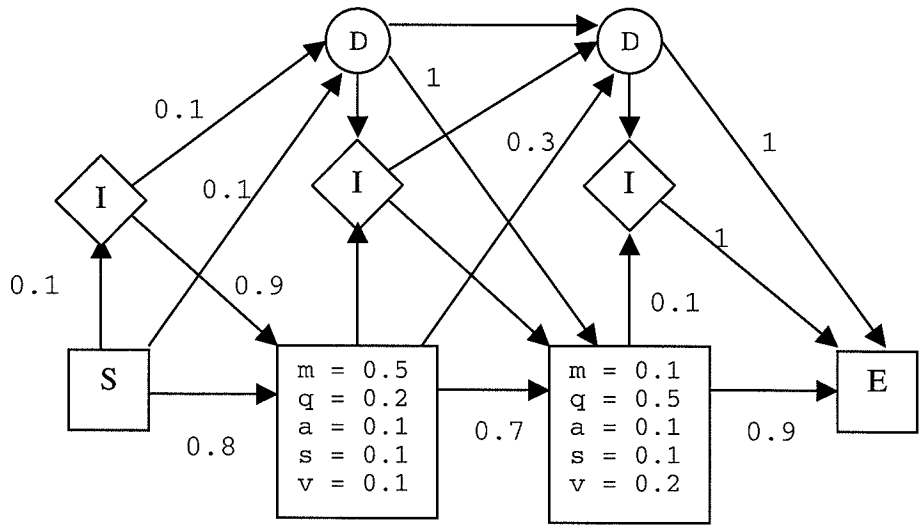
```
gene1: QQMINDVVVQ
gene2: Q-MI-NEEQM
gene3: Q-MI-NQEQM
gene4: MIMQ-N-VEN
```

iii) Use this to infer (by eye) and draw a phylogenetic tree for the genes.

iv) If used as a guide tree in the CLUSTAL algorithm, what would your phylogenetic tree guide the algorithm to do?

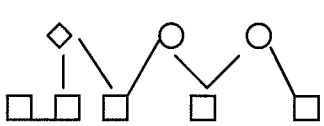
b i) Why are multiple sequence alignments (MSAs) more useful to geneticists than just sets of aligned pairs?

ii) Suppose this Hidden Markov Model (HMM) has been learned for a MSA, where any missing probabilities are assumed to be zero:

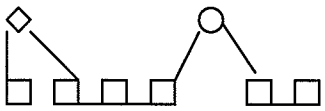


Draw two possible paths with non-zero probabilities through the HMM for this sequence: `MQA`. What are the probabilities of the sequence being generated by each path?

iii) Suppose a HMM has been learned and the best paths through the HMM for four sequences are as follows:

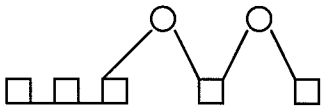


Sequences: NDNE



ENDNQ

Sequences: NNDEQ



NDE

Use these paths to produce the MSA for the four sequences indicated by the HMM.

c i) Explain in English what the Viterbi and Forward algorithm calculate. Explain how the Forward algorithm can be used to calculate the probability of a sequence being generated by a HMM.

ii) Give an overview of how a Hidden Markov Model would be trained in order to represent a given MSA for a set of residue sequences.

iii) How and why are pseudocounts used in HMM parameter learning?

The three parts carry, respectively, 25%, 40% and 35% of the marks.

3a Describe the basic procedure of acquisition and warehousing of gene expression data.

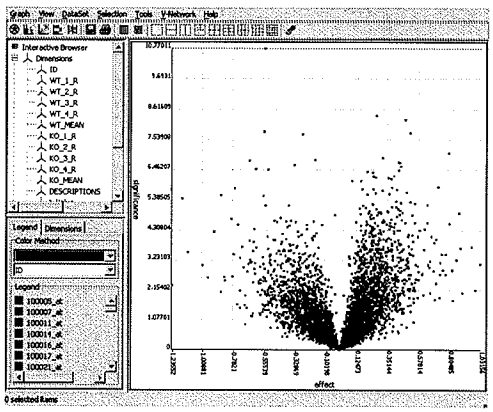
b List two of the commonly used methods in the analysis of gene expression data and describe their main functions.

c i) Explain how hierarchical clustering algorithms work. Make sure your answer describes what is meant by a linkage method and how it is used.

ii) Based on a Euclidean similarity measure, calculate the similarity matrix between the following observations of gene expressions at the two time points (T1, T2):

Gene ID	T1	T2
G0001	1	1
G0002	1	4
G0003	5	1
G0004	5	4

d Given a gene expression profile of two populations (gene expressions in untreated tissue samples vs. gene expressions in its corresponding treated tissue samples), the volcano plot provides a visualisation of significance of changes in expression values between two populations using the hypothesis test method.



Draw an analysis workflow diagram for the required data analysis.

Explain the steps in each of the stages in your workflow.

The four parts carry, respectively, 15%, 15%, 40% and 30%.

4a i) Define the term "Machine Learning".

ii) Provide a list of five different classes of machine learning task within bioinformatics, with an example of each class.

iii) Describe the advantages of probabilistic and logical representations for Machine Learning.

b You are given 10 positive examples and 10 negative examples of neuropeptide precursor proteins. How do you:

i) Machine learn rules for recognising neuropeptide proteins?

ii) Test the performance of the learned rules?
(Include a confusion matrix in your answer).

iii) Estimate the error in predictive accuracy?

c Suppose that 62 out of 80 test examples are correctly classified by a set of rules. What is:

i) The sample estimation of true error?

ii) The standard deviation of the error of this estimate?

iii) The 90% confidence interval associated with this estimate?

The three parts carry, respectively, 30%, 35% and 35% of the marks.