UNIVERSITY OF LONDON
IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2004

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
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

Friday 7 May 2004, 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) What is a chromosome? What is a gene? What do we mean if we say that two genes are homologous?

ii) In the multiple sequence alignment (MSA) below, where are the insertions and deletions? What does this MSA suggest about the evolution of these genes?

```
S1   MEQMMILN
S2   M--MMILD
S3   --EMMILV
S4   M-QMMI-V
```

iii) Calculate a matrix of genetic distances for the sequences as they appear in this alignment.

b i) Below is a Needleman-Wunsch diagram to align the sequences QMNN and EMN. What kind of algorithm is Needleman-Wunsch? What do the numbers 0, -4, -8, -12, -16 represent? These numbers decrease by a fixed amount. Give an example of an alternative scheme for these numbers.

		Q	M	N	N
	0	-4	-8	-12	-16
E	-4	2 (2)	-2 (-2)		
M	-8	-2 (0)			
N	-12				

ii) Carefully copy out and complete the above diagram using the Needleman-Wunsch algorithm. [You will need to use the partial BLOSUM-62 matrix given above].

iii) Use the completed diagram to draw dotplots for the two best alignments of these sequences and determine the alignments the dotplots dictate. What is the BLOSUM score for the two alignments?

c i) In the Position Specific Scoring Matrix (PSSM) below, what do the numbers 1/15 and 3/14 indicate?

	1	2	3	4	5	6	7	8	9
A	0	0	2/15	2/15	0	0	0	0	0
C	3/12	15/15	10/15	0	0	0	2/14	0	0
D	0	0	1/15	3/15	0	0	0	0	0
E	0	0	0	10/15	4/15	4/15	3/14	0	0
G	0	0	0	0	2/15	4/15	8/14	0	0
H	0	0	0	0	0	2/15	0	0	0
I	0	0	0	0	8/15	0	0	0	1/12

ii) Suppose the above matrix has been derived by the PSI-BLAST algorithm. Use it to determine whether the following sequence fragment matches with the current Multiple Sequence Alignment better in position 1 or position 2:

DDDEHGC

iii) Give an overview of the PSI-BLAST cycle and explain why we should be cautious about the results it provides.

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

2a After a search for matches to a sequence you estimate that, of the 250 sequences (hits) returned by BLAST, 55 are not related to your sequence, and that BLAST has missed 75 sequences you know are related to your sequence.

i) Calculate the selectivity and sensitivity of this particular BLAST search, and explain what these terms mean.

ii) Explain how the BLAST algorithm generates high scoring triples and how these triples are used in the algorithm.

iii) What high scoring pair (HSP) would be generated by BLAST for the following high scoring triple? What is the BLOSUM score of the HSP?

```
Q Q M I L V E Q N N
D E L I L V Q M M L
```

[Use the partial BLOSUM-62 matrix given above]

b Suppose you have the following Multiple Sequence Alignment that you want to construct a Hidden Markov Model for:

```
E-N-DE-
Q-NDNQE
Q-EDNN-
QQEE-N-
```

i) Including the start and end states, how many columns would your HMM have?

ii) Using the data from the MSA above, what would the emission probability distribution be for match node 1?

iii) Again using the data from the MSA above, draw a HMM for the above MSA. Put in any edges with a non-zero probability, and any nodes with a non-zero probability on the edge going into them. Omit edges or nodes with zero probability.

iv) Draw a path with a non-zero probability of occurring through the HMM for the sequence below. Calculate the probability of the sequence being generated via that path.

```
EQNDEE
```

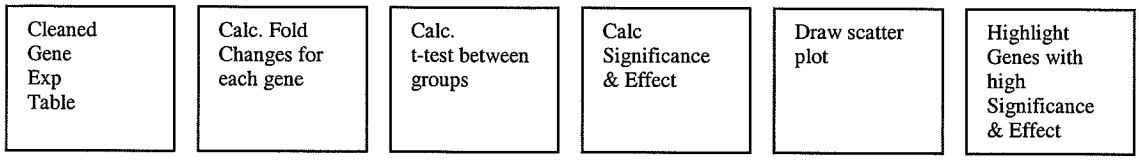
c i) Give an overview of the Baum-Welch algorithm. Include a description of how the algorithm handles the problems of local minima and overfitting.

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

3a Give a brief description of gene-expression chip technology and mention two commonly used analysis methods for gene expression data.

b Describe the application of statistical test methods in gene-expression analysis.

c The following workflow defines a gene-expression analysis process for a gene-expression profile of two populations (gene-expressions in untreated tissue samples vs. gene-expressions in its corresponding treated tissue samples),



i) Discuss the concept of fold change and give the method to compute such fold changes.

ii) Discuss the concept of significance and effectiveness of the change and describe the formula for computing the significance.

iii) Discuss possible ways to perform further analysis of the gene highlighted as significant and effective.

The three parts carry, respectively, 20%, 20% and 60% of the marks.

4a Describe the weights for perceptrons which define the following functions:

i) Not a AND b.

ii) a OR NOT b.

b 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.

c You are given 10 positive examples and 10 negative examples of 4-helical bundle proteins. How do you:

i) Machine learn rules for recognising 4-helical bundle proteins?

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

iii) Estimate the error in predictive accuracy?

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