

THE UNIVERSITY OF AUCKLAND

FIRST SEMESTER, 2015

Campus: City

COMPUTER SCIENCE

Computational Science

(Time allowed: THREE hours)

NOTE: Attempt *all* questions

Use of calculators is NOT permitted.

Put your answers in the answer boxes provided below each question.

You may use the blank pages at the end of the exam script for scratch work, which will not be marked.

<i>Section:</i>	A	B	C	Total
<i>Possible marks:</i>	15	20	25	60
<i>Awarded marks:</i>				

SURNAME:

FIRSTNAME:

ID:

CONTINUED

ID: _____

**Section A: Linear Systems and Root Finding
(15 marks)**

1. (a) Define what it means for x to be a root of the function $f(x)$. Describe the bisection method for finding the roots of f . [2 marks]

- (b) Give a reason to prefer Newton's method for finding the roots of f over the bisection method and a reason to prefer the bisection method over Newton's method. [1 mark]

2. (a) Suppose that $\mathbf{A}\mathbf{u} = \mathbf{b}$ is an over-determined linear system. What does over-determined mean and describe the solution we find by solving the associated normal equation, $\mathbf{A}^T\mathbf{A}\mathbf{u} = \mathbf{A}^T\mathbf{b}$? [1 mark]

CONTINUED

ID: _____

- (b) Describe the size and properties of the matrices \mathbf{Q} and \mathbf{R} that are the result of the orthogonalisation of \mathbf{A} . [1 mark]

- (c) Let \mathbf{q}_i be the i th column of \mathbf{Q} and \mathbf{a}_i be the i th column of \mathbf{A} . Show how \mathbf{q}_1 and \mathbf{q}_2 are derived from \mathbf{A} using the Gram-Schmidt process. [2 marks]

3. Suppose $\mathbf{A} = \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 0 & 1 \end{bmatrix}$.

- (a) The eigenvectors of $\mathbf{A}^T \mathbf{A}$ are

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

corresponding to eigenvalues 10 and 11 respectively.

What does it mean for $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ to be an eigenvector of $\mathbf{A}^T \mathbf{A}$ with eigenvalue 11? [1 mark]

CONTINUED

ID: _____

- (b) The first two unnormalised eigenvectors of $\mathbf{A}\mathbf{A}^\top$, given in order of decreasing eigenvalues, are

$$\begin{bmatrix} -1 \\ 3 \\ 1 \end{bmatrix} \text{ and } \begin{bmatrix} 3 \\ 1 \\ 0 \end{bmatrix}.$$

Normalise these eigenvectors.

[1 mark]

- (c) What are the eigenvalues corresponding to the normalised eigenvectors in the Question 3b? [1 mark]

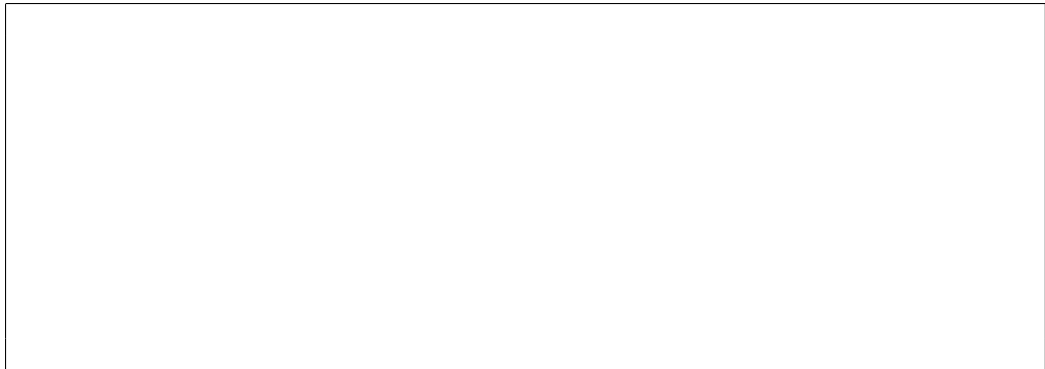
- (d) Write down the singular value decomposition for \mathbf{A} . [2 marks]

- (e) Write down the pseudo-inverse of \mathbf{A} . [2 marks]

CONTINUED

ID: _____

4. Let \mathbf{A} be a $m \times n$ measurement matrix where each column represents the result of an experiment where m measurements are taken, so each row represents a particular measurement taken across n experiments. The principal components of \mathbf{A} can be found via the singular value decomposition of matrix \mathbf{B} which is a centred version of \mathbf{A} . How is \mathbf{B} derived from \mathbf{A} ? [1 mark]



ID: _____

Section B: Sampling and Markov models (20 marks)

5. (a) Given a method `randunif()` that returns a uniform random variate between 0 and 1, write a pseudo-code method, `sample(x, prob)`, which takes the arrays `x` and `prob` of equal length such that the sum of the elements of `prob` is 1 and returns `x[i]` with probability `prob[i]`. [2 marks]

- (b) Given a method `randexp(λ)` that produces exponentially distributed samples for parameter λ , write a pseudo-code method `randpoiss(λ)` for sampling Poisson random variables with parameter λ . [2 marks]

- (c) Given methods `randunif`, `sample`, `randexp` and `randpoiss`, sketch a pseudocode method `mutate(A, L, mu, t)` that takes a sequence `A` of length `L` and mutates it according to the Jukes-Cantor model over a period of time `t` with mutation parameter μ . Allow mutations from a base to itself. [3 marks]

CONTINUED

ID: _____

6. What does it mean for a sequence of random variables X_0, X_1, X_2, \dots to have the Markov property? Express your answer in plain English and in mathematical notation. [1 mark]

7. Suppose we are modelling regions of DNA sequences as being either high or low in CG content. We use an HMM with two underlying states, H for high CG content regions and L low. An H state is followed by an L state with probability 0.05 and an L state is followed by an H state with probability 0.10. There is an 80% chance that the sequence starts in the L state.

Each state can emit any of the 4 bases A, C, G or T . Emission probabilities are given by the following table.

	A	C	G	T
H	0.1	0.4	0.4	0.1
L	0.3	0.2	0.2	0.3

- (a) Sketch a diagram of the HMM, showing all states, possible transitions and transition probabilities. Include the begin state. [2 marks]

- (b) Given a state sequence π , explain why the length of a run of L s in π is geometrically distributed with parameter $(1 - a_{LL})$. [2 marks]

CONTINUED

ID: _____

- (c) What is the joint probability $P(x, \pi)$ of the state sequence $\pi = LH$ and the symbol sequence $x = AC$? (You can leave your answer as a product or sum of numbers). [1 mark]

- (d) The forward algorithm calculates the quantity $P(x)$ by iteratively calculating the quantities $f_H(i)$ and $f_L(i)$. For the symbol sequence $x = ACCG$, what probability does $f_H(4)$ represent? [1 mark]

- (e) Given the forward matrix, F , below for the sequence $x = ACCG$, what is $P(x)$? (You can leave your answer as a product or sum of numbers).

	A	C	C	G
$F : H$	0.01	0.0146	0.007496	0.00320496
L	0.27	0.0487	0.008912	0.00167912

[1 mark]

- (f) For the HMM discussed here and a sequence x of length L , what is the algorithmic complexity of the forward algorithm? What is the algorithmic complexity of directly calculating $P(x) = \sum_{\pi} P(x, \pi)$ using brute force methods? [1 mark]

CONTINUED

ID: _____

- (g) Suppose we did not know the emission or transition probabilities of the HMM but had observed a state sequence, π , and a symbol sequence, x . Illustrate how to estimate the emission and transition probabilities and HMM with the same structure as the one discussed here given $\pi = HHLLL$ and $x = ATGGT$. [2 marks]

- (h) The Baum-Welch algorithm is used to estimate the parameters of an HMM. Give a basic pseudo-code description of the version of the Baum-Welch algorithm that uses the Viterbi path as the imputed state sequence at each iteration. Include the initialisation procedure, how each iteration works and the stopping condition. [2 marks]

CONTINUED

ID: _____

Section C: Alignment and phylogenetics (25 marks)

8. Let the symmetric matrix

$$D = \begin{pmatrix} 0 & & & \\ 8 & 0 & & \\ 8 & 2 & 0 & \\ 4 & 8 & 8 & 0 \end{pmatrix}$$

specify the pairwise distances, D_{ij} , between the four sequences x_1, \dots, x_4 .

(a) Construct a UPGMA tree from this distance matrix, showing your working. [3 marks]

(b) What biological assumptions are we making when we assume that UPGMA will construct the correct tree from given biological sequences? [1 mark]

CONTINUED

ID: _____

- (c) Under what circumstances will neighbour joining find the correct tree where UPGMA fails and why do we sometimes prefer to use UPGMA anyway? [1 mark]

9. (a) Why do we use heuristic techniques to solve the multiple alignment problem rather than extend the dynamic programming approach we apply to pairwise alignment? [1 mark]

- (b) What role does the neutral character X play in the Feng-Doolittle method for progressive alignment? [1 mark]

10. Consider the four aligned sequences, W, X, Y, and Z:

W: GGTC
X: ATCC
Y: AACC
Z: GTTC

- (a) When is a site un-informative to a parsimony analysis? Specify which sites in this alignment are parsimony uninformative. [2 marks]

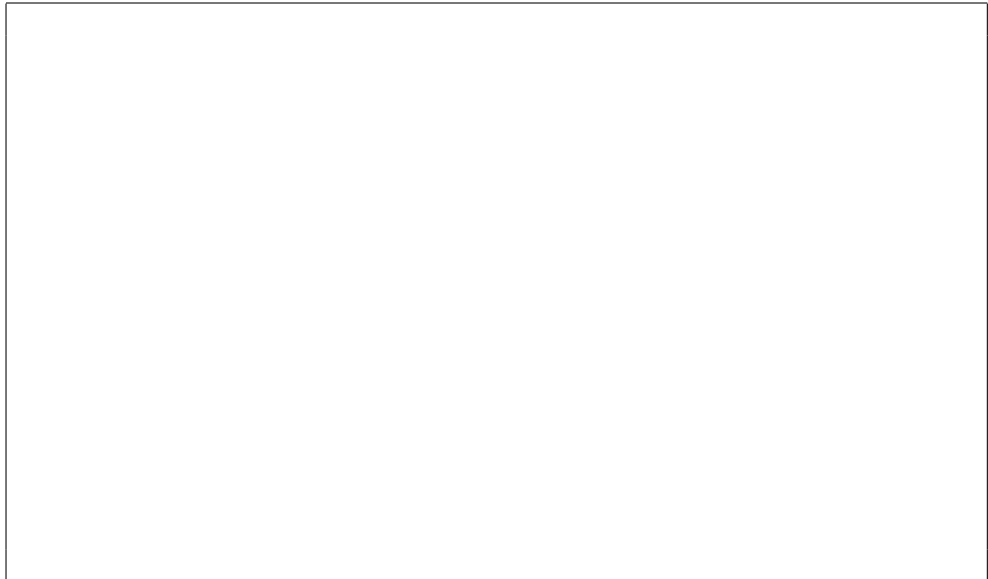
CONTINUED

ID: _____

- (b) By enumerating all possible trees on 4 taxa and finding the parsimony score for each tree, identify the maximum parsimony tree for this alignment. [3 marks]



11. (a) Sketch the steps of a heuristic algorithm that can be used to find a tree with a low parsimony score. [2 marks]



- (b) Why is this method not guaranteed to find the maximum parsimony tree? [1 mark]



CONTINUED

ID: _____

- (c) Give two reasons to prefer likelihood based tree reconstruction methods over parsimony methods and one reason to prefer parsimony methods. [2 marks]

12. The partially completed F matrix for calculating the local alignment of the sequences GAC and TAACT is given below. The score matrix is given by $s(a, b) = -1$ when $a \neq b$ and $s(a, a) = 6$ while the gap penalty is $d = -3$.

		T	A	A	C	T
	0	0	0	0	0	0
G	0	0	0	0	0	0
A	0	0	6	6	3	0
C	0	0	u	v	x	y

- (a) Complete the matrix F by finding values for u, v, x and y . [2 marks]

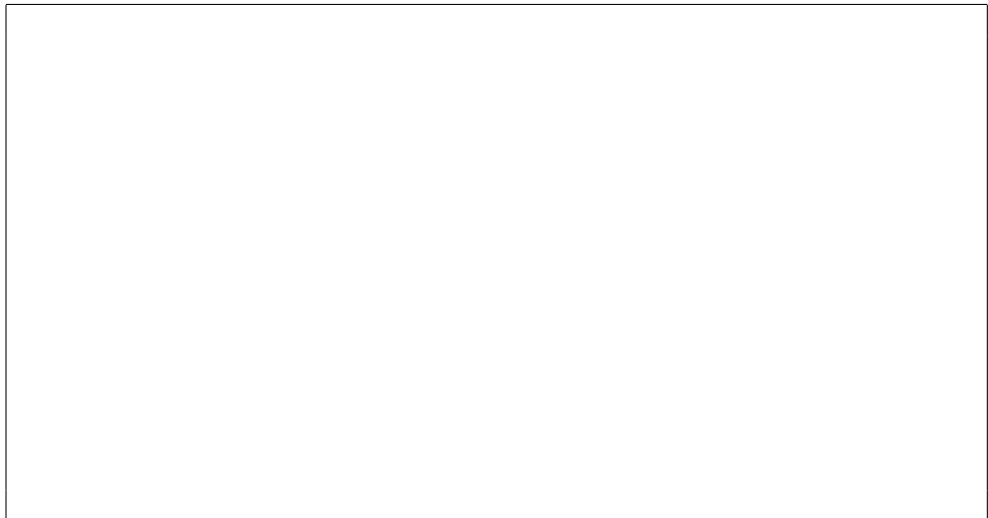
- (b) Give the score for the best local alignment of these two sequences and provide an alignment that has this score. [2 marks]

ID: _____

- (c) Describe the changes required to transform the local alignment algorithm into the global alignment algorithm. [2 marks]



- (d) The gap penalty used here is linear. Define an affine gap penalty and state why it is preferable from a biological point of view. [2 marks]



CONTINUED

ID: _____

Blank page 1 — will not be marked

CONTINUED

ID: _____

Blank page 2 — will not be marked

CONTINUED

ID: _____

Blank page 3 — will not be marked
