**Statistical and Numerical Methods for Bioinformatics: Exercises 1**

**E1** Cars pass a certain tree on a quiet country road at a rate of one every two minutes.
(i) What is the probability that exactly five cars pass the tree in ten minutes?
(ii) What is the probability that exactly five cars pass the tree in ten minutes given that in the first five of these minutes one car passes the tree?
(iii) What is the probability that all of the periods 10.00-10.02, 10.02-10.04, 10.04-10.06, 10.06-10.08, 10.08-10.10 contain at least one event of a car passing the tree?
(iv) Would a Poisson process be a good model for a busy road?

**E2** Ten unemployed former students join a job club. As soon as one gets a job, they leave. Suppose that each is equally likely to get a job and receives offers (which are always accepted) at rate 3 per year.
a) What is the probability that all the students have got a job after 2 years ?
b) What is the probability that the last student to get a job gets it in the third year ?
c) How long must we wait before the probability that all the students have accepted a job is greater than 0.5 ?

**E3** Consider a simple birth and death process with birth rate 2 and death rate 1, starting with two individuals. What is the probability that;
(i) the population becomes extinct ?
(ii) the population reaches 5 before it becomes extinct ?
(iii) the population becomes extinct, given that it reaches 5 ?
(iv) the population reaches 5, given that it becomes extinct ?
(v) the population is extinct at time 2 ?
(vi) What is the expected number of offspring for a given individual ?

**R1** If the number of messages arriving at a telephone exchange occurs according to a Poisson process of rate 3 per minute, find the probability that
(i) no messages arrive in a minute
(ii) greater than two messages arrive in a minute
(iii) the 3rd message arrives in the second minute.

**R2** A population starting with 16 individuals following a death process of rate 3 per year has been running for six months. Find the probability that:
(i) the population size is 3.
(ii) the population size is no more than 3.

**R3** A population following a birth process of rate 0.5 per day starts with 3 individuals. Find the probability that
(i) there are still 3 individuals after one day (i.e. there have been no births)
(ii) there are more than 5 individuals after one day
(iii) there are between 100 and 200 individuals after ten days.

**Statistical and Numerical Methods for Bioinformatics: Exercises 2**

**E1** Find the stationary distribution of the Markov chain with transition matrix

$$\begin{vmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{vmatrix} \tag{1}$$

**E2** In the Markov chain defined in E1, find the probability that the occupied state at time 3 is $E_1$, given that the initial distribution (at time 0) is (0.8, 0.2).

**E3** Consider the random walk described by the following matrix (similar to that in the notes, except that when either state 0 or $n$ is reached, the next step is forced to be 1 from 0, or $n-1$ from $n$.

| | 0 | 1 | 2 | 3 | ... | $n-3$ | $n-2$ | $n-1$ | $n$ |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 0 | ... | 0 | 0 | 0 | 0 |
| 1 | $q$ | 0 | $p$ | 0 | ... | 0 | 0 | 0 | 0 |
| 2 | 0 | $q$ | 0 | $p$ | ... | 0 | 0 | 0 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| $n-2$ | 0 | 0 | 0 | 0 | ... | $q$ | 0 | $p$ | 0 |
| $n-1$ | 0 | 0 | 0 | 0 | ... | 0 | $q$ | 0 | $p$ |
| $n$ | 0 | 0 | 0 | 0 | ... | 0 | 0 | 1 | 0 |

(2)

a) Show that this Markov chain is periodic (you only need to show that at least one state can only be reached at given regularly spaced times).
b) Construct equations for the stationary distribution $\phi$ and solve them for the case where $p = 0.5$

**R1** Simulate a continuous time Markov chain with the following transition ($\mathbf{Q}$) matrix, up to time 30, obtaining a series of states and the times of each transition. Repeat the process twice and compare the results.

$$\begin{vmatrix} -0.2 & 0.1 & 0.1 \\ 0.5 & -0.6 & 0.1 \\ 0.1 & 0.2 & -0.3 \end{vmatrix} \tag{3}$$

**R2** A continuous time Markov chain has all possible transitions between the states $i, j = 1, 2, 3$ $i \neq j$. Observations were made at the following times

Times - 0, 0.1, 0.3, 0.7, 0.8, 0.9, 1.2, 1.4, 1.6, 1.7, 2.1, 2.2

The following sequence gives the observed states

States - 1, 1, 1, 2, 2, 1, 1, 1, 3, 3, 3, 1

Obtain estimates of the transition rates between states and the mean time spent in any given state in one visit.

**R3** Consider the aneurism data in the data file *aneur*. 838 male patients over 65 were followed, with transitions occurring in sequence between grades of aortic aneurisms Aneurism free, Mild aneurism, Medium aneurism, Severe aneurism.

Obtain estimates of the transition rates between the states.

**General** Repeat the simulation from R1. Using the data that you obtain estimate the transition PROBABILITIES between the states. By comparing the real transition rates to the formulae in the notes, find the true transition probabilities, and comment.

**Statistical and Numerical Methods for Bioinformatics: Exercises 3**

**E1** Suppose that we wish to cover a sequence of length 10000 bases with N pieces each of length 500 bases, randomly chosen.
(i) What is the expected proportion of coverage for the sequence if N=50?
(ii) What value would $N$ have to be to give 99% coverage?

**E2** A sequence contains 100000 bases, and the frequency of $a$ in the sequence is 0.3. What is the probability that the maximum number of $a$'s in a row in the whole sequence is at least 15?

**E3** Using the simple scoring system
*number of matches-number of mismatches-number of gaps*
find the best alignment of $x = cttgac$ in $y = cagtatcgtac$
(i) where gaps are not allowed (simply state the highest score obtained),
(ii) where gaps are allowed (it will take a long time to do this exhaustively; try to find good alignments by eye).

**R1** Investigate the data *woodmouse* which is a set of 15 sequences of woodmouse DNA data, using the R help system. Describe the data in more detail.

**R2** Find the base frequencies for each of $a, g, c$ and $t$ for the woodmouse data.
Use this to test the hypothesis of equal frequencies within the data (Hint: if you can find the number of observations from the data file you can use the frequencies given to calculate the total number of each base observed).
What conclusions do you reach?
Are there any problems with this testing idea? If so, how would you modify your test?

**R3** Generate the PAM1 matrix in R, and hence find the PAM6 matrix using matrix multiplication.

**General** What is the probability that a given amino acid A does not occur at a given location at each of 6 successive PAM time units if all transitions between amino acids are equally likely? How would you work this out if they were not?

**Statistical and Numerical Methods for Bioinformatics: Exercises 4**

**E1** Define a Hidden Markov model $\Lambda$ with the following parameters:
Three states $S_1, S_2, S_3$, alphabet $A = \{, 1, 2, 3\}$, transition matrix $P$ given by

$$\begin{vmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{vmatrix} \tag{4}$$

$\pi = (1, 0, 0)$
$b_1(1) = 0.5, b_1(2) = 0.5, b_1(3) = 0$
$b_2(1) = 0.5, b_2(2) = 0, b_2(3) = 0.5$
$b_3(1) = 0, b_3(2) = 0.5, b_3(3) = 0.5$

What are all possible state sequences for the observed sequences $O$, and what is $P(O|\Lambda)$, for the following sequences?
(i) $O = 1, 2, 3$
(ii) $O = 1, 3, 1$

**E2** A Hidden Markov model is defined as follows:
There are two states $S_1, S_2$, alphabet $A = \{1, 2\}$, transition matrix $P$ given by

$$\begin{vmatrix} 0 & 1 \\ 0.5 & 0.5 \end{vmatrix} \tag{5}$$

$\pi = (0.6, 0.4)$
$b_1(1) = 0.5, b_1(2) = 0.5$
$b_2(1) = 1, b_2(2) = 0$

If the observations $1, 1$ occur, what is the most plausible underlying sequence of states?

**E3** Consider the five amino acid sequences $WRCCTGC$, $WCCGGCC$, $WCGCC$, $WCCCGCC$, $WCCGC$. Suppose that their respective paths through a protein model HMM of length 8 are

$$\begin{array}{cccccccccc} m_0 & m_1 & i_1 & m_2 & m_3 & m_4 & m_5 & d_6 & m_7 & m_8 \\ m_0 & m_1 & m_2 & m_3 & m_4 & m_5 & m_6 & m_7 & m_8 & \\ m_0 & m_1 & m_2 & d_3 & d_4 & m_5 & m_6 & m_7 & m_8 & \\ m_0 & m_1 & m_2 & m_3 & m_4 & m_5 & m_6 & m_7 & m_8 & \\ m_0 & m_1 & m_2 & m_3 & d_4 & m_5 & d_6 & m_7 & m_8 & \end{array} \tag{6}$$

Give the alignment of the sequences that these paths determine.

**R1** Generate 100 random observations from a gamma distribution with parameters 2 and 1, and then perform a one sample bootstrap on the data set, taking 1000 replicates. Find a 95% confidence interval for the true mean and provide a histogram of the sample means.

**R2** Repeat R1 using the exponential distribution with parameter 100.

**R3** Generate two data samples using random numbers with Normal distributions with
a) mean 1, variance 1
b) mean 0, variance 1
Carry out a 2-sample bootstrap for the difference of the means of the distributions, giving a 95% confidence interval for this difference and plot a histogram of the sample mean differences.

**General** Generate a new random sample as in R3 a). Plot the empirical distribution function for this sample and the true distribution function for the underlying distribution. Repeat this for the distribution in R2. Comment.

**Statistical and Numerical Methods for Bioinformatics: Exercises 5**

**E1** Using the Metropolis Hastings algorithm, find a 3-state Markov chain with a stationary distribution (0.2,0.5,0.3).

**E2** Suppose that we have a two dimensional vector $\mathbf{Y}$, where each element can take values 1 or 0 only, so there are four possible vectors only
(0,0), (0,1), (1,0) and (1,1).
Suppose that these vectors occur in relative proportions (0.2,0.4,0.45,0.05).

Using the Gibbs sampling method, construct the appropriate Markov chain with this stationary distribution.

**E3** For the multiple sequence alignment problem from the Gibbs sampling example, show that if we can consider $q_{ij}(s)$ and $q^*_{ij}(s)$ as approximately equal, the relative entropy between $q^*_{ij}(s)$ and $p_j$

$$\sum_{i=1}^{W}\sum_{j=1}^{20} q^*_{ij}(s) log\left(\frac{q^*_{ij}(s)}{p_j}\right)$$

has an approximate linear relationship with the logarithm of the probability of being in state $s$

$$C\prod_{i=1}^{W}\prod_{j=1}^{20}\left(\frac{q^*_{ij}(s)}{p_j}\right)^{c_{ij}(s)}$$

[Note that $X$ and $Y$ have a linear relationship if for some constants $a$ and $b$, $Y = aX + b$].

The three R-based questions all concern the routine *MCMCmetrop1R* in the MCMCpack library. Visit the help system and read about this routine. In particular the first example is about logistic regression which serves all three questions. This example is divided into three main components; the function to optimise, generation of the data (these are simulated) and the application of the routine with some related commands.

**R1** Copy the central part of the example (from the line beginning x1 to that beginning yvector) and run this in R. Obtain plots of the data to get a feel for how $y$ depends upon the x-values (noting that $y$ are Bernoulli variables, taking values only 0 or 1) .

**R2** Now copy first the example function (logitfun to }) and then run it, with related commands, (using post.samp to summary(post.samp)). Give estimates of the three variables, and briefly explain the other output.

Note that the function being optimised is

$$\sum (ylog(p) + (1-y)log(1-p))$$

where

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}$$

**R3** Repeat R2, but this time choose a different example function. You can do this by entering each line in turn and making a single change to the form of '$p$' (but make sure that this always lies between 0 and 1, as $p$ is a probability).

**General** Construct a continuous-time Markov chain whose transition RATES correspond to the transition probabilities that you found in E1. Hence simulate a sequence of transitions of the process from E1 How else could you perform such a simulation without working out these rates?.

**Statistical and Numerical Methods for Bioinformatics: Exercises 6**

**E1** Find the proportion of each zygote $A_iA_j$ under Hardy-Weinberg equilibrium for the following sets of alleles and frequencies;
(i) $A_1$ occurs in frequency 0.7 and $A_2$ in frequency 0.3.
(ii) $A_1$ occurs in frequency 0.45, $A_2$ in frequency 0.33 and $A_3$ in frequency 0.22.

**E2** The zygotes $A_1A_1$, $A_1A_2$ and $A_2A_2$ occur in a sample with the following frequencies:
Case (i) 37, 126, 132
Case(ii) 72, 75, 68
For each of the above, perform a chisquare test for Hardy-Weinberg equilibrium.

**E3** Repeat question E2, but this time perform Fisher's test in each case.
Compare your results and comment.

**R1** Suppose that a population has two alleles $A$ and $B$ at a given locus, and that the numbers of each genotype are as follows
AA 23, AB 55, BB 58
Test for Hardy-Weinberg equilibrium using
(i) a $\chi^2$ test,
(ii) Fisher's exact test.

**R2** Suppose that a population has three alleles $A, B$ and $C$ at a given locus, and that the numbers of each genotype are as follows
AA 17, AB 32, BB 71, AC 44, BC 63, CC 49
Test for Hardy-Weinberg equilibrium using a $\chi^2$ test.

**R3** Estimate the level of disequilibrium, and find a 95% confidence interval for this level, for
(i) the data from question R1
(ii) the same data, but with the number of AAs being 53 instead of 23
(iii) the same data, but with the number of AAs now 8 and the number of BBs 38 (instead of 58).

**General** Simulate a new generation of individuals descended from those in R1, assuming that the population follows Hardy-Weinberg.

**Statistical and Numerical Methods for Bioinformatics: Exercises 7**

**E1** Obtain the distances between each of the nodes in the following tree and verify that these distances satisfy the distance properties (i)-(iii) from the notes.

**E2** For the five species $a, b, c, d$ and $e$ with distances given in the following table, reconstruct the tree using the algorithm in 17.3, starting with two species and adding species sequentially.

| | a | b | c | d | e |
|---|---|---|---|---|---|
| a | 0 | 2 | 8 | 8 | 8 |
| b | | 0 | 8 | 8 | 8 |
| c | | | 0 | 4 | 4 |
| d | | | | 0 | 2 |
| e | | | | | 0 |

**E3** Use the UPGMA algorithm to construct the tree between the five species with distances given in the following table.

| | a | b | c | d | e |
|---|---|---|---|---|---|
| a | 0 | 9 | 8 | 7 | 8 |
| b | | 0 | 3 | 6 | 7 |
| c | | | 0 | 5 | 6 |
| d | | | | 0 | 3 |
| e | | | | | 0 |

**R1** Consider the phylogeny data set *bird.orders*. Use *R* to show that the tree is ultrametric and binary and obtain a plot of the tree.

**R2** Generate a random tree with
a) 7 species
b) 25 species.
Plot the tree for each case.
In each case convert the tree to an ultrametric one, and obtain plots of the new trees.

**R3** The data file *bird.orders* and the random trees above are data sets of the class *phylo*. By looking at the plots of the trees from questions R1 and R2, and the information given when you type the name of the data file, explain the information in a 'phylo' data set.

**General** Repeat R2 with a tree with 5 species, find the distance matrix between the species, and then use one of the methods from the notes to reconstruct the tree from the distance matrix.

**Statistical and Numerical Methods for Bioinformatics: Exercises 8**

**E1** Verify that the stationary distribution for the Kimura model is (0.25, 0.25, 0.25, 0.25)

**E2** The Jukes-Cantor model is a special case of the Kimura model (so that if you choose $\alpha = \beta$ in the Kimura model you get the Jukes-Cantor) and also of the Felsenstein model. Is either the Kimura model or the Felsenstein model a special case of the other?

**E3** Discuss the example in section 18.4 of the notes. Some groupings were consistent across all methods, others were not. By examining the data, explain why particular groupings would be consistent and others not. in particular why is the location of 'human' so different?

**R1** For each of the ultrametric random trees and the *bird.order* data from the previous question sheet, obtain the matrix of distances associated with the tree.

**R2** Use a Kimura evolutionary model to obtain a distance matrix for the data *woodmouse*. Repeat this with a Jukes-Cantor model.

**R3** Use a Kimura evolutionary model to obtain the distance matrix for the following three DNA strings, which are subsequences of the first three sequences of the woodmouse data from the previous question.
ATCAGTCACT
ATCAACCACT
ATCAATCACT

**General** Draw the tree relating to the sequences from R3 (by hand). Comment on any problems that you have, and why they might be occurring. Repeat R3 with different sequences, and compare your results.

**Statistical and Numerical Methods for Bioinformatics: Exercises 9**

**E1** A genetic algorithm has six individuals in each generation. The fitness of the current six individuals are 0.7,1.2,1.3,1.5,1.6 and 1.7 respectively.
(i) What is the probability that the first individual in the mating population is a copy of the fittest individual?
(ii) What is the probability that exactly two out of the six individuals in the mating population are copies of the fittest individual?

**E2** Two individuals from the mating population have genomes (1,1,2,1,1) and (1,2,1,1,2). List
(i) all of the possible genomes under the random allocation model,
(ii) all of the possible pairs of genomes under the single crossover model.
(iii) For each of parts (i) and (ii), what is the probability that a genome will contain a sequence of (at least) four 1s?

**E3** The values of the first allele of two mating individuals, from a continuous set, are 1.342 and 1.765 respectively. Assuming that random allocation occurs in recombination, and that mutations occur with probability $r = 0.15$ according to a normal distribution with a mean of the current allele and variance 1:
(i) What is the probability that, if the chosen value from recombination is 1.342, the new allele will have value less than 1?
(ii) What is the probability that the new allele has value less than 1, if the chosen value of the allele after recombination is unknown?