UNIVERSITY OF CAMBRIDGE

M. Phil. in COMPUTATIONAL BIOLOGY

Friday, 15 May, 2015 2:00 pm to 4:00 pm

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
Calculator - students are permitted to bring an approved calculator.

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1 Population genetic analyses of genomic data

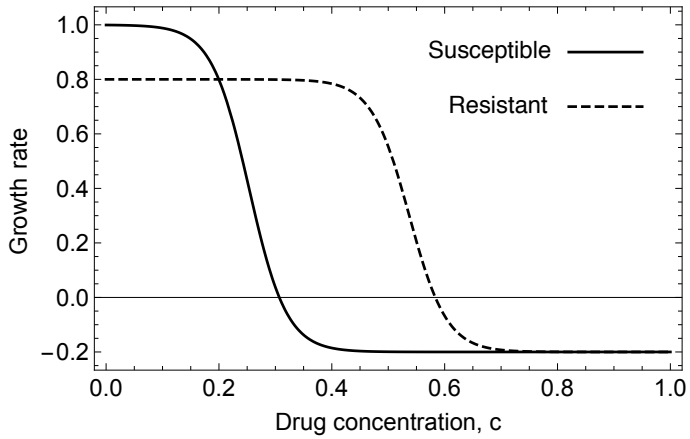
Complete both parts A and B of the question.

A1) Provide definitions for the following terms:

i) Genetic drift
ii) Effective population size
iii) Linkage disequilibrium
iv) Epistasis

A2) Describe two ways via which recombination can increase the fitness of a population over multiple generations.

B1) The figure below shows the relative growth rates of susceptible and resistant strains of *E. coli* when treated with different concentrations of a particular drug.



Plot a graph of the magnitude of selection acting upon the resistant strain (vertical axis) against the drug concentration (horizontal axis).

B2) Consider a case of two antibiotic drugs, denoted A and B. Suppose that *E. coli* has two drug resistance loci in its genome, a and b, with each locus either having the allele 0 or 1, so that the allele 1 at locus a confers resistance to drug A, while the allele 1 at locus b confers resistance to drug B. Considering alleles at the loci a and b, we have four possible genotypes, 00, 01, 10, and 11. We describe the growth rates of these genotypes as:

$$\begin{aligned} f_{00} &= 1 - c_A - 0.5c_B - kc_Ac_B \\ f_{01} &= 0.8 - c_A - 0.05c_B \\ f_{10} &= 0.4 - 0.1c_A - 0.5c_B \\ f_{11} &= 0.1 - 0.1c_A - 0.05c_B \end{aligned} \tag{1}$$

where c_A and c_B are the concentrations of drugs A and B, and k is a real number.

i) In the above system, describe the meaning of the parameter k.

ii) Plot the fitness landscape for the different genotypes for the case in which $c_A = 0$ and $c_B = 0$.

iii) Plot the fitness landscape for the different genotypes for the case in which $c_A = 1$, $c_B = 1$, and $k = 0$.

iv) Relative to the case in which $k = 0$, which values of k will lead to a slower emergence of drug resistance when used to treat the fully susceptible strain? Describe how this occurs.

2 Scientific programming

Question 2 contains three parts; the code within each part is to be studied independently. Hints: `Inf` represents positive infinity, `paste(1:3, collapse=' ')` generates `"1 2 3"` and `floor(x)` rounds `x` down to the nearest whole number.

Part A Networks.

I Each row of the following matrix `g` contains three elements (x, y, w) which defines a directed connection from node x to node y with edge of strength w. Draw the network of nodes and edges (node 1 on the left of the page, node 10 on the right).

```
g <- matrix( c(1, 2, 1,
               1, 3, 3,
               1, 4, 2,
               2, 5, 3,
               2, 6, 1,
               3, 5, 2,
               3, 6, 4,
               3, 7, 3,
               4, 6, 2,
               4, 7, 2,
               5, 8, 1,
               6, 8, 2,
               6, 9, 1,
               7, 9, 1,
               8, 10, 3,
               9, 10, 3), ncol=3, byrow=T)
```

II Execute the following code and show the final value of the matrix `S`:

```
nsteps = 5
nt = length(unique(as.vector(g[,1:2])))
S = matrix(data=Inf, nrow=nsteps, ncol=nt)
prev = c(1);  S[1,prev] = 0
for (step in 2:nsteps) {
  t = vector()
  for (p in prev) {
    d = which(g[,1] == p)
    t = c(t, d)
  }
  for (a in t) {
    x = g[a,1]; y = g[a,2]; w = g[a,3]
    s = S[step-1,x] + w
    if (s < S[step,y]) {  S[step,y] = s  }
  }
  prev = unique(g[t,2])
}
```

III What does the following code print out?

```
prev = which.min(S[nsteps,])
nodes = c(prev)
for (step in (nsteps-1):1) {
  t = which(g[,2]==prev)
  p1 = g[t,1]
  smin = Inf
  for (i in p1) {
    if ( S[step, i] < smin ) {
      prev = i; smin = S[step, i]
    }
  }
  nodes = c(prev, nodes)
  cat(paste('step', step, "p1", paste(p1,collapse=' '),
            'smin', smin, 'prev', prev, '\n'))
}
print(nodes)
```

IV Explain briefly what the code in part II and III does.

Part B Random samples.

I Explain briefly what the following code does; include a graph to help explain what the code is doing.

```
n = 100
x1 = 1; x2 = 2
y1 = 0; y2 = 1
x = runif(n, min=x1, max=x2)
y = runif(n, min=y1, max=y2)
f = sum(y < 1/x)
a = (x2-x1) * (y2-y1) * f / n
```

II How can you use the value of `a` to approximate `ln(8)`?

Part C Recursion.

I What does the function `m(a,b)` return, given two vectors of the same type as input?

```
empty = function(l) { length(l)==0 }

m = function(a, b) {
  res = vector()
  while (!empty(a) & !empty(b)) {
    if (a[1] < b[1] ) {
      res = c(res, a[1])
      a = a[-1]
    } else {
      res = c(res, b[1])
      b = b[-1]
    }
  }
  res = c(res, if(empty(b)) a else b)
  res
}
```

II Execute the following code, and show what is output each time by the `cat` statement. Ensure that you get the order of output correct.

```
ms = function(x) {
  l = length(x)
  if (l < 2) {
    res = x
  } else {
    h = floor(l/2)
    a = x[1:h]
    b = x[-(1:h)]
    a1 = ms(a)
    a2 = ms(b)
    res = m(a1, a2)
  }
  cat(paste("in:", paste(x, collapse=' '),
            "out:", paste(res, collapse=' '), "\n"))
  res
}

ms( c(5, 2, 1, 3, 6, 4))
```

3 Functional Genomics

- Briefly describe the typical contents of the following file-types that are commonly-used in Genomics analyses:

  a) fastq

  b) sam

  c) bam

  d) bed.

- Match the type of DNA involved with the appropriate experiment type

Assay	Type of DNA
1. Whole Genome Sequencing	a) Cross-linked DNA
2. RNA-Seq	b) Purified DNA
3. ChIP-Seq	c) cDNA

- RNA-Seq can be used to measure mRNA transcript abundance. Name three regulatory elements that are involved in determining levels of transcription.

- Your collaborators are interested in the effect of estrogen on the genes in ER+ breast cancer cells over time. After serum starvation of all eight samples, they exposed four samples to estrogen, and then measured mRNA transcript abundance after 10 hours for two samples and 48 hours for the other two. They left the remaining four samples untreated, and measured mRNA transcript abundance at 10 hours for two samples, and 48 hours for the other two. Since there are two factors in this experiment (estrogen and time), each at two levels (present or absent, 10 hours or 48 hours), this experiment is said to have a 2 x 2 factorial design.

  a) Draw a diagram of the experimental design.

  b) Write the design matrix that you would use in limma to fit a linear model with *no interaction* to this experiment.

- Having collected all the samples required for the experiment described above, you collaborator is unsure about whether to perform the study with single-colour microarrays or RNA-Seq. Give three advantages of RNA-Seq over microarrays.

- After the sequencing run has been completed, you are passed a series of single-end fastq files. Describe the steps you would take to produce a list of differentially-expressed genes. Give the names of the tools that you would use and quality control checks that you would perform at each stage.

- What type of experiment can be used to detect where regulatory proteins bind to DNA? Discuss how such an experiment would complement the results of your differential expression analysis.

END OF PAPER