UNIVERSITY OF CAMBRIDGE

M. PHIL. IN COMPUTATIONAL BIOLOGY

Friday, 10 May, 2019 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 Functional Genomics

(a) Explain why the following sentences are true or false:

- *We want to estimate the effect of a treatment of the expression of a given gene. If we were to conduct an experiment and obtain a 95% confidence interval for the fold change under the treatment, there is a 95% probability that the true value of the fold change is included within the bounds of the confidence interval.*
- *We want to estimate the effect of a treatment on the expression of a given gene. We have conducted an experiment and obtained a 95% confidence interval for the fold change under the treatment. There is a 95% probability that the true value of the fold change is included within the bounds of the confidence interval.*
- *We have performed 1000 hypothesis tests and we have obtained 90 p-values smaller than 0.05. The expected proportion of false positives amongst them would be 90% approximately.*

(b) In the context of pathway analysis:

- Explain what are the goals and the steps needed to run a gene ontology analysis and a gene set enrichment analysis.
- When we try to quantify the involvement of a pathway in our data set, what is the difference in using just the list of genes, or split by overexpression and underexpression?

(c) In the context of somatic copy number analysis in tumour samples,

- What is the relationship between normal cell contamination and normalised log ratios to detect copy number gains and losses?
- What is the relationship between normal cell contamination and allelic frequency to detect one copy losses?

(d) In the context of cluster analysis:

- Explain methods that help you deciding the number of clusters present in your data and validating a cluster analysis.

(e) In the context of the following experiment, where we want to estimate the expression of a given gene under the values of ER (level of oestrogen receptor with two possible values, positive + and negative -), and the dose (different doses of radiation):

Sample	**ER**	**Dose**
Sample 1	+	37
Sample 2	-	52
Sample 3	+	65
Sample 4	-	89
Sample 5	+	24
Sample 6	-	19
Sample 7	+	54
Sample 8	-	67

- Write down the design matrix for a model with no interaction between ER and Dose.
- Write down the design matrix for a model with interaction between ER and Dose.

2 Genome sequence analysis

(a) A pair of dice are thrown. Let X be a random variable representing the higher of the two dice, and let Y be a random variable representing their sum.

(i) Find the expected value of the higher of the two dice.

(ii) Find the expected value of the higher of the two dice given that their sum is 9 or more.

(b) Consider a HMM with hidden variables X_0^N, emission variables Y_0^N, state space $\{s_1, \ldots, s_J\}$, transition matrix A and emission distributions $b_i(v)$. Define $\beta_n(i) = \mathrm{P}(Y_{n+1}^N \mid X_n = s_i)$.

(i) Show that the following recursion relation holds:

$$\beta_n(i) = \sum_j \beta_{n+1}(j) b_j(v_{n+1}) A_{ij}$$

(ii) It can be shown that

$$\mathrm{P}(X_n = s_i \mid Y_0^N) = \frac{\alpha_n(i)\beta_n(i)}{\sum_j \alpha_n(j)\beta_n(j)}$$

How is $\alpha_n(i)$ defined here, and what is the significance of this result for inferring the hidden states of a HMM given some data?

3 Scientific Programming

1. Study the following R code:

```
s = function(n) {
  a = b = 0
  while( n > 0 ) {
    n = n - 1
    if ( a == 1 ) {
      a = 0; b = b + 1
    } else {
      a = 1
    }
  }
  c(b, a)
}

g = function(a, b) {
  if (b==1) {
    a
  } else {
    r = s(b)
    ifelse(r[2], a, 0) + g(a+a, r[1])
  }
}

## Four cases to study
g(9,8) # 1
g(8,9) # 2
g(6,10) # 3
g(7,9) # 4
```

(a) What do `s(6)` and `s(7)` return? What is the purpose of the function `s`? [15%]

(b) What does the function `g` generate for the four cases at the bottom of the code? Show your working. [20%]

(c) What is the purpose of the function `g`? How does it work? [15%]

2. Study the following R code:

```
f = function(n, y) {
  stopifnot(all(diff(y)>0))
  a = rep(0, n)
  for (c in y) {
    for (j in 1:n) {
      if (c <= j) {
        l = j - c
        w = ifelse(l==0, 1, a[l])
        a[j] = a[j] + w
      }
    }
    print(a)
  }
  a[n]
}

### Three cases
f(7, c(1,2,4))    ## 1
f(12, c(1,2,4))   ## 2
f(12, c(1,5,10))  ## 3
```

(a) What are n and y assumed to be in the function f? [10%]

(b) What is printed and what is the output from the three cases at the bottom of the code? [30%]

(c) What is the function f doing? [10%]

END OF PAPER