UNIVERSITY OF CAMBRIDGE

M. Phil. in COMPUTATIONAL BIOLOGY

Friday, 13 May, 2016 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 Scientific Programming in R

(a) Define each of the following terms in the context of the R programming language: [25%]

1. data frames
2. vectorization
3. the apply family of functions
4. vector recycling
5. formal arguments and local arguments

(b) The following function finds prime numbers. When `e(14)` is evaluated, what is printed, and what is returned? [25%]

```
e = function (n) {
  s = 2:n
  p = c()
  for (i in 2:n) {
    print(s)
    if (any(s==i)) {
      p = c(i, p)
      s = s[(s%%i) != 0]
    }
  }
  return(p)
}
```

Hints: `a %% b` returns the remainder when a is divided by b; `a %/% b` returns a/b ignoring any remainder. Hence `7 %% 2 = 1, 7 %/% 2 = 3`.

(c) What does the following function return when called with `b(148)` and `b(1240)`? What does the function do in general for non-negative integer values of n? How would you change the function to improve the output for `b(0)`? [25%]

```
b = function(n) {
  ## n is a non-negative integer of length 1.
  r = c()
  ok = n>0
  while (ok) {
    l = n %% 10
    r = c(l, r)
    n = n %/% 10
    ok = n>0
  }
  r
}
```

(d) Study the following function (`b` is defined in the previous part). Show what the function returns in the following cases:

`d(1210), d(1211), d(21200), d(21201), d(3211000)`

Can you state what the function is doing in general terms? (Hint: the looping variable `i` deliberately indexes from zero.)

```
d = function(n) {
  v = b(n)
  l = length(v)
  ok = TRUE
  for (i in 0:(l-1)) {
    c = v[i+1]
    s = sum(v==i)
    ok = ok && (s==c)
  }
  ok
}
```

[25%]

2 Genome Informatics

(1) (a) Perform a global alignment between the sequences 'RACE' and 'RANCE' using dynamic programming using the following scoring parameters:
Match +1, Mismatch -1, Gap -2
Show your working (i.e. the matrix you use to calculate the alignment and the path through it). [28%]

(b) The above alignment uses a simplified scoring method. How can the scoring method be improved for better alignments? [10%]

(2) List the steps needed to go from a set of raw reads to an annotated genome. Summarise how each step may be carried out (some steps can use multiple different methods). [34%]

(3) (a) What are SNPs and how frequently do they occur in the human genome? [7%]

(b) What are the different types of SNP and what effects can they have? [21%]

3 Population Genetics

(a) In a laboratory-grown population of organisms, successive genome sequencing reveals a change over time in the frequency of an allele at a certain locus. Assuming that the observation is accurate, describe three potential causes for such a change. What factors would you take into account in order to determine what was the most likely cause of the change? [30%]

(b) Suppose that you intend to conduct a study to identify the time to the most recent common ancestor from which any pair of students in the 2015-16 cohort are descended. What data would you plan to collect? What theory would you use to estimate this value? What factors could you take into account in setting up an evolutionary model? [70%]

END OF PAPER