BIOINFORMATICS & COMPUTATIONAL GENETICS MSc
PROBABILITY AND STATISTICS

S-PLUS EXERCISE SHEET I

There is a data set containing a segment of human chromosome 13 containing the BRCA2 breast cancer gene; it was obtained from the National Center for Biotechnology Information (NCBI) website

`http://www.ncbi.nlm.nih.gov/`

and is held at the Probability and Statistics course website

`http://stats.ma.ic.ac.uk/~das01/BioinformaticsMSc/`

Using a Web browser, go to this link and then click on the link

BRCA2 segment direct from the NCBI site

in the **Data Sets** segment.

Two processed versions of this data set are also available;

- *A processed version with base characters* has just the 127079 base characters
- *Another processed version* contains a numerical representation of the sequence. This is the file to download.

Click on the link, and when the file has fully downloaded, use **File -> Save** to save the file as **z74739.txt** in the **c:\TEMP** directory on the hard drive. Of course, you may ultimately save the data to a floppy disk, or to your home directory. It is possible (and quite straightforward) to read in data from different types of file apart from plain text files, and from files with different formats.

The numerical data in the file **z74739.txt** are coded so that

$$\begin{aligned} 1 &= A \\ 2 &= C \\ 3 &= G \\ 4 &= T \end{aligned}$$

S-PLUS Analysis

Follow the commands to attempt a summary analysis of this sequence:

1. Starting an Splus session

Double click the S-PLUS icon, or start the program from the **Start** menu. When the program has opened up the *Object Browser* window, start a *Commands Window* by using the **Window -> Commands Window** pull down menus.

2. Loading the data into Splus

There are two ways to load the data. First using the **File -> Import Data -> From File** pull down menus, find the **c:\TEMP\z74739.txt** file using the dialog box (remember that the file extension is .txt, so the S-PLUS dialog box may not find the file initially). However, when you find the file and click **Open**, the data will be downloaded into an S-PLUS data frame called **z74739** - if you return to the *Commands Window* and type

```
>length(z74739[,1])
```

(always press **Enter** or **Return** at the end of a line command) you get the response

```
[1] 127079
```

which is the length of the sequence. To read the data into a vector, type the following at the command line:

```
>brca2_z74739[,1]
```

which creates the vector **brca2** containing the data.

The other way to read in the data is to type a command at the command line

```
>brca2_scan(''c:\\TEMP\\z74739.txt'')
```

which creates the same **brca2** vector. To check the steps have worked correctly, type

```
>length(brca2)
```

you should again get the response

```
[1] 127079
```

3. Summary analysis

For a simple summary of the sequence, type

```
>table(brca2)
```

You should get the response

```
     1     2     3     4
 38514 24631 25685 38249
```

which is the breakdown of the sequence by base, that is there are 38514 As, 24631 Cs, 25685 Gs and 38249 Ts. To analyse a sub-sequence, say the first 50000 bases, type

```
>table(brca2[1:50000])
```

that is, look at only the positions 1 to 50000 in the **brca2** vector. You then get the response

```
     1     2    3     4
 14901 10397 9908 14794
```

Therefore, to obtain the sample probabilities, of each base, type

```
>brca2.p <- table(brca2)/length(brca2)
>brca2.p
```

which calculates the probability vector **brca2.p** vector by dividing the values obtained by the **table** command by the total length of the sequence. You should get the response

```
         1         2         3        4
 0.3030713 0.1938243 0.2021184 0.300986
```

Note that the probabilities are not (even approximately) equal. Note also that by typing

```
>brca2sub.p <- table(brca2[1:50000])/length(brca2[1:50000])
>brca2sub.p
```

which calculates the probability vector **brca2sub.p** vector by repeating the calculation for the first 50000 bases, you get the response

```
      1       2       3       4
 0.29802 0.20794 0.19816 0.29588
```

indicating a similar collection of probabilities.

4. Adjacent base pairs analysis

Suppose now that an analysis of adjacent pairs in the sequence is required Type the following sequence of commands (take care not to miss any brackets):

```
>brca2.mat <- matrix(0,4,4)
>brca2.pmat <- matrix(0,4,4)
>brca2.totals <- table(brca2)
>for(i in 2:length(brca2))
+ {brca2.mat[brca2[i-1],brca2[i]] <- brca2.mat[brca2[i-1],brca2[i]]+1}
>for(i in 1:4) {brca2.pmat[i,] <- brca2.mat[i,]/brca2.totals[i]}
```

(the + sign in line 4 appears automatically when you hit return from line 3). These commands construct a matrix counting the number of adjacent pairs of each type are present in the sequence, that is a matrix

		Next base				
		A	C	G	T	
	A	n_{AA}	n_{AC}	n_{AG}	n_{AT}	n_A
Base	C	n_{CA}	n_{CC}	n_{CG}	n_{CT}	n_C
	G	n_{GA}	n_{GC}	n_{GG}	n_{GT}	n_G
	T	n_{TA}	n_{TC}	n_{TG}	n_{TT}	n_T

where the row totals are just the total numbers of bases of that type (as calculated in step 2 to produce vector **brca2.totals**. Don't worry too much about the exact meaning of each command at this stage.

The third step (which essentially counts the number of adjacent AA, AC, AG etc pairs) may take a while. The fourth step divides each row in the matrix of counts by the row totals, to produce the final matrix of probabilities:

```
> brca2.pmat
          [,1]      [,2]       [,3]      [,4]
[1,] 0.3451732 0.1628239 0.23978294 0.2522200
[2,] 0.3525638 0.2431895 0.05440299 0.3498031
[3,] 0.3011485 0.2029200 0.23262605 0.2633054
[4,] 0.2300975 0.1871421 0.23880363 0.3439567
```

Note that the row sums in this matrix are 1 (by construction). From this analysis we see that "transitions" from C to G in the sequence are relatively rare.

Two things to consider:

(a) are the transition probabilities approximately equal in different segments of the entire sequence ?

(b) are the coding regions/exons/introns (that can be identified from the file from the Bioinformatics MSc.page (or the NCBI site) fundamentally different in terms of their composition by base ?.

5.. Finishing the session.

When you want to close the S-Plus session, use the **File -> Exit** pull down menus; you are then prompted by a **Save Database Changes** dialog box, on which you should click **Select All** and then **OK.** It is possible (but a little tricky) to keep a record of the commands issued and the results obtained during the session; an easy way to do this is to click, drag and (hence) highlight the work you want to save, and the copy and paste the work to a text file that you can save on your hard disk