582606 Introduction to bioinformatics, Autumn 2007

Exercise session 1
12.9.2007

In the CS department Linux environment, you are able to access R with the command

```
$ /opt/R/bin/R
```

You can find more information on the R project's webpage at `http://www.r-project.org/` or from numerous tutorials on the web, such as `http://www.cyclismo.org/tutorial/R`. The course book appendix A "A Brief Introduction to R" might also be useful.

In assignments 1-3, you need to describe the steps you took in detail. Be prepared to present the input, the steps (R commands) you took and the output of R commands in the exercise session.

1. *(Basics of R language: data input)*

   Create a vector of character strings in R from the text file `http://www.cs.helsinki.fi/mbi/courses/07-08/itb/exercises/ex1names.txt`.

   Then, create a matrix of integers from the text file `http://www.cs.helsinki.fi/mbi/courses/07-08/itb/exercises/ex1data.txt`.

   Print out your vector and matrix. Rows of the matrix correspond to students' grades from five assignments, while the vector contains student names.

   In R you can name the rows and columns of your matrices. Assign names for the rows of your matrix from the vector of character strings so that the i*th* string in the vector will be the name of the i*th* row of the matrix.

   Save data in your matrix to an external file called `matrix.txt`. What is the format of the saved file?

   Compute the means of the scores of the first four assignments (columns) for every student.

2. *(Basics of R language: sampling)*

   Make a vector variable Δ of the characters $\{A, C, G, T\}$ by assigning the values to a new variable.

   Create a random string of characters of length 100 by randomly sampling the vector Δ, with the probability for each character being the same. How can you count the occurences of each character in the string?

   Repeat the previous step 1000 times. Calculate the mean proportions of character occurences in the strings. *Hint: try the* `replicate` *command.*

3. *(Basics of R language: plotting)*

   Plot a histogram of the character occurence distribution you obtained earlier.

   How can you set the title and axis labels for your plot? How can you create a PostScript file of your plot?

4. Read the article The Diploid Genome Sequence of an Individual Human by Samuel Levy *et al.*, *PLoS Biology* Vol. 5, No. 10, 2007. The journal is freely accessible from the computers at the University of Helsinki at `http://biology.plosjournals.org`.

   Answer the following questions.

   - What was the main result presented in the article?
   - How was the result was obtained (in general terms)?
   - What is the significance of the result?

   To present your answers, be prepared to give a short (about 5-10 minutes) presentation. You do not have to understand the whole article to complete the assignment. Instead, try to extract the main ideas and explain them in your own words.

5. (Chapter 2, Exercise 1) The base composition of a certain microbial genome is $p_G = p_C = 0.3$ and $p_A = p_T = 0.2$. We are interested in 2-words where the letters are assumed to be independent. There are $4 \times 4 = 16$ 2-words.

   (a) Present these 16 probabilities in a table. Do your 16 numbers sum to 1.0?

   (b) Purine bases are defined by $R = \{A, G\}$ and pyrimidine bases by $Y = \{C, T\}$. Let E be the event that the first letter is a pyrimidine, and F the event that the second letter is A or C or T. Find $P(E), P(F), P(E \cup F), P(E \cap F)$ and $P(F^c)$.

   (c) Set $G = \{CA, CC\}$. Calculate $P(G|E), P(F|G \cup E), P(F \cup G|E)$.

Please report any inaccuracies and errors in the exercises to Esa Pitkänen!