Assignments to Lecture 3 – Reading data into R

Danny Arends

Reading data

0a) Download the data file: "Assignment 3 - Data.zip"
0b) **Unzip the data** to a folder on your harddrive
1) Read in the different text datasets (txt, fasta, vcf), using the read.table or read.csv function:
- Make sure that the separator is set correctly
- Make sure numbers are read in as numbers
- Make sure strings are not factors, when we want strings
- Make sure strings are factors when they need be
- Make sure row and column names are correct
- Make sure that number of rows loaded is correct (compare with the file)

2a) Read in the text file 'Lorem Ipsum.txt' line by line, using a **file connection** and a **while loop**
2b) **Copy the code** from assignment 2a. Adjust it so that each line in the file, the number of words on the line us counted using the ***strsplit*** function. Use the ***cat*** function to print out to screen: 'line x, contains y words' where x and y are the line number and the number of words on that line, respectively.

biomaRt (internet required, if you don't have internet you can skip ahead)

3a) install the package biomaRt from the bioconductor repository
3b) Load the package, and list all the marts available
3c) Connect to the SNP database for mouse
3d) Load in the SNPids.txt file
3e) Query biomaRt for the SNPids, and retrieve: "refsnp_id", "allele", "chr_name", "chrom_start"

Continue a long running analysis

4a) Set your random seed to a fixed value (so that we generate the same matrix with random numbers each time), and create a matrix holding random values to calculate correlation on (size ~ 10000 x 1000).
4b) Decide where to store the results on disk (location and filename)
4c) Check if this file is empty (e.g. using an ***if*** statement, or the ***tryCatch*** function). If the file is not empty we need to load in the data from the previous computation. If the file is empty we need to create an empty matrix (in the ***else*** branch or in the error clause of the ***tryCatch*** function). Load in the results using the read.table or read.csv function.
4d) Use a **for** loop to go through the columns of the matrix, calculate the correlation (using the ***cor*** function) of the current column against all the other columns (including itself). NOTE: If the file from 4c was found on disk continue from the line numbering of the file, otherwise start at 1
HINT: The correlation function can do 1 vs many in one call of the function: ***cor(matrix[,x], matrix)***
4d) Inside the loop store the result from correlation and write the new results to the file (using the ***cat*** function) and store it to disk

Reading binary data

5a) Load the provided bmp file in R use *what='raw'* (check the slides for the example)
5b) Do not forget to throw away the first 54 bytes (the header of the image)
5b) Extract the red/green and blue channel data from the BMP image
5c) Create images of the different color channels using the ***image*** function

Advanced (since we didn't talk about plots yet)

We are going to create a pointillism painting in R using the 3 color channels of the image.

6a) First we need to setup a plot window (200 x 200) (no type, no axis, no plotting), we can do this by using the ***plot*** function and setting the paramters 't', 'xaxt', 'yaxt', 'xlab', and 'ylab'
6b) Setup the locations [1,1] [1,2] [1,3] … [1,200] [2,1] [2,2] … [200,200], and plot them to the empty plot window using the ***points*** function.
6c) Now add the colors to the plot using the col parameter, and the **rgb** function