Biol135: Introduction to Bioinformatics programming

Lab 7: Introduction to R programming

For this lab we will begin working with the text based programming language R. For help with basic R syntax and commands I recommend using tutorialspoint. This lab will use the output from our previous scratch lab to quantify the Nucleotide transition frequencies generated by your scratch code. This lab will help us towards building a functioning Hidden Markov model for identifying genomic features.

Part 0: Download Rstudio onto your computer if you have not done so already. Note that rstudio is an interface for R, not the interpreter itself, so it will prompt you to install the interpreter first before installing Rstudio.

Part 1: Using your scratch code, create a list of 100 random DNA characters. Right click on the list that your code created and click export. Name the file "testATCG.txt" If you open it up it should show a series of lines with one nucleotide character in each line.

Part 2: Please create a new R script entitled "[your last name]TransitionMatrix.R" that does the following.

1) Reads a file called "testATCG.txt" from the working directory.
2) Reads the lines one by one and stores them in a vector.
3) Counts the transition frequencies for each possible starting and following character.
4) Stores the transition frequencies as a data frame and prints that data frame to the command line. The data frame should have 16 rows, one for each possible transition. With each row displaying the beginning nucleotide, the following nucleotide, and the absolute count.

Part 3: Just as with our scratch lab, each separate block of code should receive a comment explaining how that code contribute to the overall solution. Include one comment explaining how changing the transition probabilities in your scratch code changed the reported counts in your R output.

Part 4: Save and Upload your script to the MyCourses dropbox. You do not need to include the test file.