BIOLOGY 664: Integrated Bioinformatics Using R for Both Wet and Dry Scientists

Problem Set 9: Analyzing Sequences

Due: At the beginning of class Thursday, April 24th.

Chapter 9 Exercises

Problem 1 – Use the JASPAR2014 and TFBSTools packages to download the PWM for the tumor suppressor protein BRCA1 from the JASPAR website. Mutations in BRCA1 (and its paralog BRCA2) are highly associated with early-onset breast cancer. Download the new Krijnen.chapter9.r script for an example of how to use these tools. Generate the sequence logo representing the sequence specificity of the BRCA1 transcription factor. Use the BRCA1 PWM to find putative binding sites on chromosome 4 of the Drosophila genome. (Hint: You can confirm your sequence logo by querying for BRCA1 at the JASPAR website http://jaspar.genereg.net/)

Problem 2 – Write a dynamic programming function in R to perform pairwise overlap alignment as discussed in class. Remember that the overlap alignment is a modification of the pairwise GLOBAL alignment algorithm - using elements from the pairwise LOCAL alignment method. Use your new R function to overlap align the amino acid sequences HEAGAWGHEE and PAWHEAE. For the alignment, you should use the BLOSUM50 substitution matrix with gap opening cost 0 and gap extension cost -8 (just like the Needleman-Wunsch and Smith-Waterman examples in the Krijnen.chapter9.r script). Remember that you can check your solution using the pairwiseAlignment() function. (Hint: look at the Needleman-Wunsch and Smith-Waterman code in the new Krijnen.chapter9.r script. Also, the correct optimal overlap alignment is presented in the Chapter 9 lecture notes.)

Problem 3 – Answer exercise 7 in the book - but instead of performing many pairwise alignments, perform one multiple alignment of the 8 sequences using the MUSCLE R package. To validate your alignments you can upload or paste your sequences into the MUSCLE web service at https://www.ebi.ac.uk/Tools/msa/muscle/.

At the very least, you need to hand in an R Markdown File (.Rmd) that logically delineates and presents you R code as well as the R output and plots generated by your code for each exercise. For more information on how to create an Rmd file look here:

http://www.rstudio.com/ide/docs/authoring/using_markdown

Hardcopy

Also, in addition to submitting a softcopy through blackboard I'd also like to get a printed hardcopy of your Problem Set. Please print double-sided and use Arial font size 7 or 8 in order to save trees.

Extra Credit

I'll be giving extra credit for those Problem Sets that use Knitr and LaTeX to create printed PDF reports that include your R code as well as the R output and plots generated by your code. I'm also going to give extra credit for plots that have been "fancied up" to produce more publication-like figures. These plotting enhancements include colors, titles, legends, axis labels, p-values, and R-squared's. Being able to create publication-quality figures in R is an important skill, and this is an opportunity to earn extra points while learning to do so. Look at the "Quick-R" and "Producing Simple Graphs with R" links on the online syllabus for good online references for how to fancy-up R plots.