Bios 555: High-throughput data analysis using R and Bioconductor

Homework 3

The written assignment is due on **September 29, Wednesday before 11:59pm.**

I. Short answer questions, 10 points each. Try your best to be specific and provide concrete answers in a technical way. For example, instead of saying “use function foo”, say “foo(x=a, y=b, something=TRUE)”. Partial credit will be given.

1. Given a DNA sequence, how to compute the AT content (the percentage of bases of being either A or T)?
2. How to find the occurrence of a specific sequence pattern “ACCGTT” from a long DNA sequence?
3. What is run length encoding method? Why is it useful?
4. What’s the run length encoding results for the following vector c(1,1,1,2,3,3,3,3)? How to achieve it in R?
5. If you are provided two lists of genomic intervals in text files, each file has three columns for chromosome, start and end positions. How to compute the percentage of intervals in the first list overlapping intervals in the second list.

II. Based on the results obtained from the lab, write a short report to present the exploratory analysis results of human sequence compositions. Imagine you are writing a scientific paper to be submitted to *Science*. You should provide descriptions, figures and statistical analysis results to address the findings that the human genome is depleted in C, G and CG, but the depletion is less severe at gene promoter regions. Introduce the definition of CpG Island, and discuss their relationship with transcriptional starting sites (TSS). (50 points).