Lab #10: LAST LAB! Next week will be study guide for the final.

This one will test your programming skills in R (or the language of your choice), so may be more challenging than previous labs. Last labs at the end the semester are meant to stretch your skills. Do your best. (Or you might find it easy!)

You can do this project entirely in R, or you can use the language of your choice (Python, Java, Perl) if you prefer. You’ll probably want to use R for the graph at the end, but you can as an alternative use the graphing language of your choice.

Please e-mail code, graphs and answers to questions to [afodor@uncc.edu](mailto:afodor@uncc.edu)

Please have lab submitted (whatever you have) before lab on Wed., April 27.

In this lab, we will use sliding windows to identify GC islands in a region of Chromosome #1 of the Human Genome.

The sequence to work on (part of Chromosome 1 of the Human genome) is here:

<http://afodor.github.io/classes/stats2016/chr1_substring.txt>

Download it into a file. You can read it into R with something like:

fileName <- 'chr1\_substring.txt'

myChar <- readChar(fileName, file.info(fileName)$size)

Our goal is to scan this sequence for GC islands in windows off 200 base pairs starting at position 1 and then every 50 nucleotides. So the first window will be 1-200, the second window will be 51-250 and so forth.

We will use the GC island di-mer “substitution matrix” defined in the Durbin book ( see Lecture #23 ; slide 10).

We can represent this in R as:

vals<- t( matrix(c( -0.740, .419, 0.580,-0.803, -0.913, 0.302, 1.812, -0.685, -0.624, 0.461, 0.331, -0.730, -1.169, 0.573, 0.393, -0.679) , nrow=4))

or in Java as:

**private** **static** **double**[][] *VALS* = { {-0.740, .419, 0.580,-0.803}, { -0.913, 0.302, 1.812, -0.685 } , { -0.624, 0.461, 0.331, -0.730}, {-1.169, 0.573, 0.393, -0.679 }

The assignment:

For every 200 basepair window in the sequence:

Sum up the score for all the dinucleotides.

So if a sequence was AACG the score would be:

( -0.74 + 0.419 + 1.812) which is the score for AA + AC + CG

Graph the score (on the y-axis) vs every starting position for every 200 basepair window along the sequence. Show this graph in your answer.

Is there evidence for any GC islands in this region of the human genome?