**Problem Set 3**

**Problem 1**

(a) See code.

(b) See results files data<N>\_results.txt

**Problem 2**

(a) See evosig\_results.txt

(b) *Are there biases in the sequence properties of either class? If so, where does this bias come from?*

Most frequent k-mers appear to be AT rich. Most conserved k-mers appear to have a more even distribution of all four bases.

*Which of these two lists should we use to direct further our inquiry into yeast transcription factor binding sites?*

Conservation. Even though not all conservation islands will be motifs (many are due to chance or non-motif conservation), we know that individual motif instances are preferentially conserved. You would want to look at conservation in:

1. Intergenic regions
2. Intergenic vs coding regions
3. Upstream vs downstream regions

You would then want to compare conservation to controls.

*From a scan of motifs yeast\_motifs.txt, which known motifs does your scan of 6-mers find?*

The yeast\_motifs.txt file is not well explained. It has letters other than ATCG, so I thought maybe it was encoding amino acids. So I wrote a script to convert the codons into amino acids. However the motifs in yeast are much longer than 2 AA... I’m not sure what the intention was here. The 6-mers are so short that I don’t think finding them in an AA motif is necessarily significant.

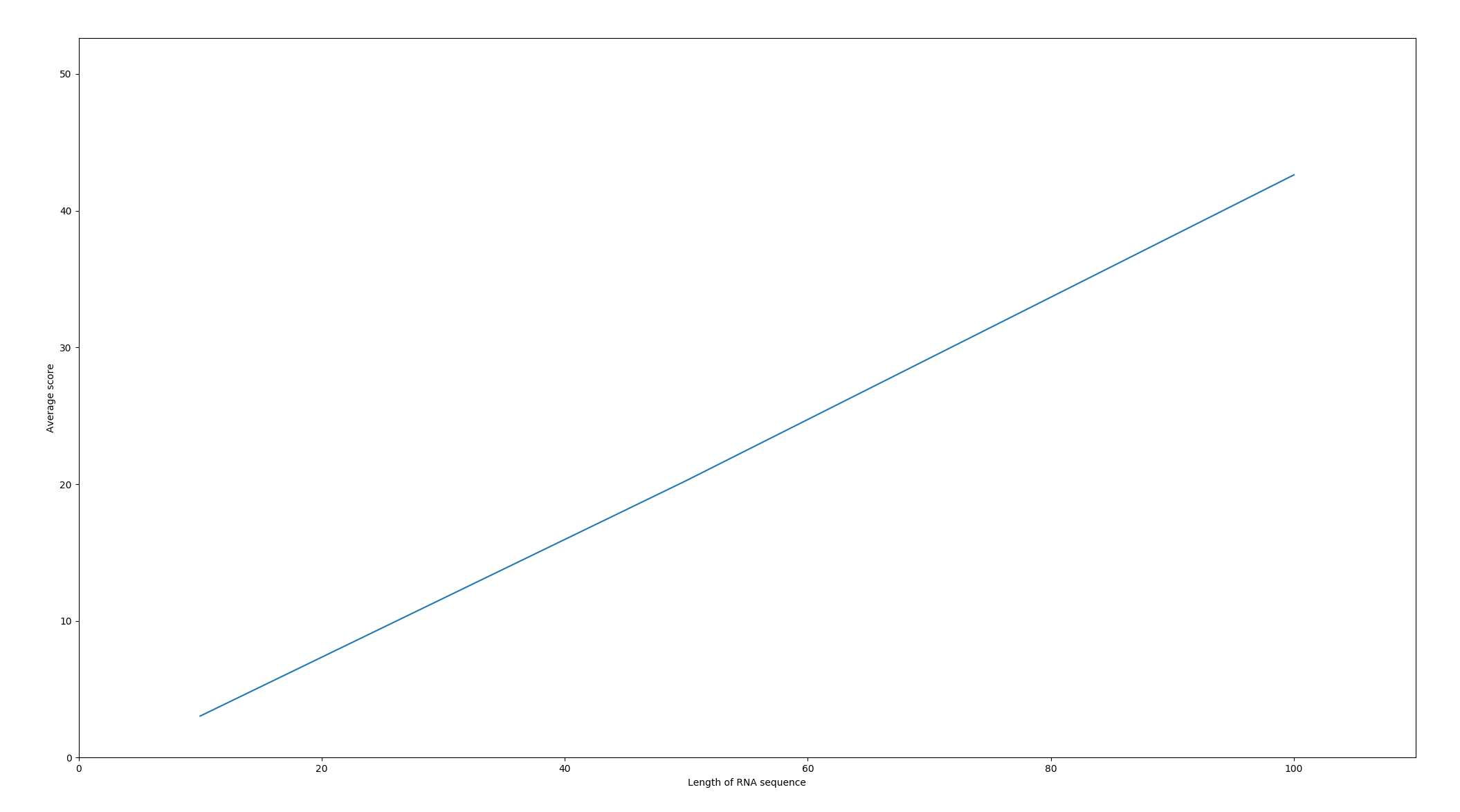
Also, because these motifs are intergenic and are not converted into proteins, it’s odd that they would be expressed as amino acids...

**Problem 3**

(a) See code. Note that I made some changes to the scoring, setting matching base pairs as +1, and all other base pair matches as 0. I also used a maximization function in the scoring rather than a minimization function.

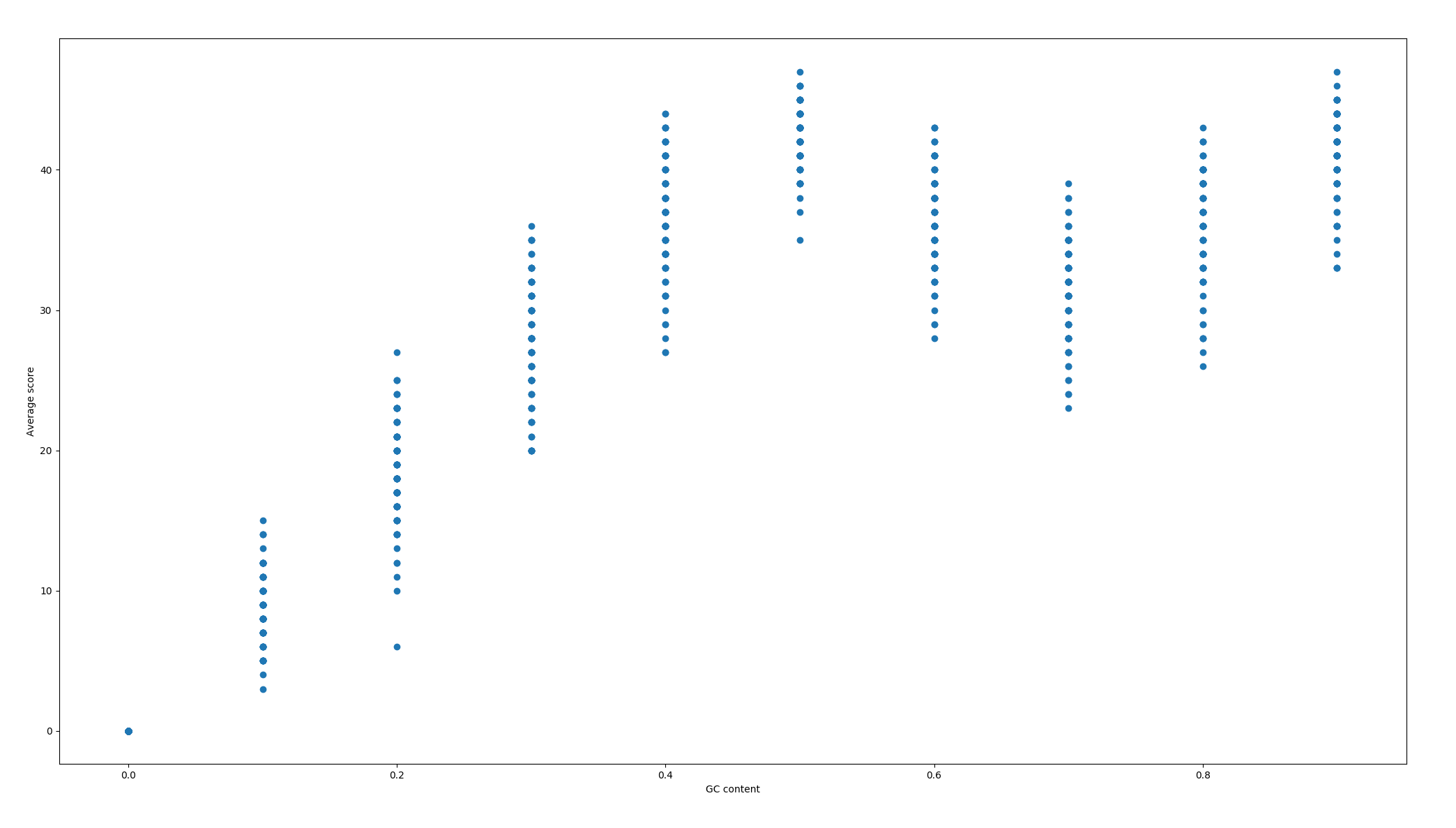
(b) For 100 length randomly generated RNAs, the average score was

(c) Nussinov score vs length



There appears to be a linear relationship between length of sequence and Nussinov score.

(d) Score vs GC content



It appears that when GC < 0.5, the relationship between score and GC is linear. When GC > 0.5, the relationship is either parabolic or flat. This is likely because pairing is maximized when there are equal numbers of ACUG (when CF = 0.5), because A and U cannot bind C and G. However when GC content approaches 100%, the score begins to rise as there are few As and Us.

(e) Going forward, we know that the Nussinov score will increase linearly with sequence length, and will have two maxima at GC content = 0.5 and 1.0.