Beliveau Lab Summer Project: Exploratory Analysis on NUPACK "prob" function and KAS-seq

Jumana Fathima

Summer 2020

Introduction

As published by Wu et al.[1], Kethoxal assisted single-stranded DNA sequencing (KAS-seq) has recently been used in understanding transcription dynamics in situ. Specifically, this method showed that enhancers that were single-stranded could be characterized by KAS-seq, and showed different behavior than other enhancers. KAS-seq acts by labelling ssDNA with N3-kethoxal, and the labelling was specific to ssDNA due to Watson-Crick labelling in double stranded DNA. Enhancer regions that were labelled as single-stranded using KAS-seq methods showed higher enhancer activity when filtering for distal epigenetic marks such as H3K27ac, H3K4me1, and H3K4me3. KAS-Seq peaks also correlate with ChIP-seq and ATAC-seq peaks in enhancer regions and at transcription start site.

In this research project, I aimed to use an in-silico method of finding single stranded DNA, and to match up results with those seen from KAS-seq profiling. The "prob" function of NUPACK, which is a popular bioinformatics tool, calculates the probability of a given secondary structure for a sequence of nucleotides. This probability is determined by free energy calculations done on the sequence. The purpose of analyzing the "prob" values, and looking for correlations with KAS-seq peaks, is exploratory—if there is a correlation between the in silico and in situ methods of detecting single stranded enhancers, it would be interesting to explore and analyze in other experimental methods, and it would be interesting to see the correlation between in silico methods of enhancer activity correlating with other epigenetic transcription markers.

Goals

The goals of this project are three:

1. I aim to build a Python-based pipeline to analyze outputs of "prob" function on NUPACK across various enhancers
2. Then, I aim to match these regions with KAS-seq outputs from the Nature paper mentioned above, to see if there is a correlation between high single stranded probability and KAS-seq reads
3. Finally, I aim to use other data, such as ChIP-seq peaks, to identify if there are interesting biological factors tied to KAS-seq reads

Methods

All parts of the pipeline can be found at the Beliveau Lab Github repository, linked here.

KAS-seq data was made available through the supplemental data provided by the paper mentioned above. This paper also linked publicly available ChIP- and ATAC- seq data that they used in their experimental methods, which I also used in my analyses.

In this initial analysis, mm10 was used consistently throughout all data to ensure that coordinates match up well and results are accurate. The results that will be displayed below were done only on chromosome 1.

The first part of the pipeline involved extracting enhancer sequences for one chromosome, given a FASTA file of the entire chromosome, as well as a list of chromosome start and stop locations for various enhancers[2]. The bed file that has the start and stop locations is in the Github repository, and the chromosome FASTA files were downloaded from UCSC (https://hgdownload.soe.ucsc.edu/goldenPath/mm10/chromosomes/). The start and stop locations for each enhancer were mapped to specific locations in the FASTA file, and the sequence of the enhancer was then extracted and output into a file. The Python script associated with this step is labelled Bed_to_FASTA.py.

This output file, with the nucleotide sequences of each enhancer, was then fed into a script that calculated the probability that the enhancer sequence could be found in any secondary structure, such as a random coil, using the NUPACK "prob" function. These "prob" values were calculated for windows of 17 nucleotides each per enhancer. These "prob" values were then output into a file. The Python script associated with this step is labelled nupack_prob_test.py, and the output file is labelled chr1_enhancer_prob.csv. Other windows were also tested, and the output files are located in the Github, labelled chr1_enhancer_prob_WINDOW.csv.

Finally, these "prob" values were analyzed in a wide variety of ways, along with KAS-seq, ChIP-seq, and other information about the enhancer region and its activity. This was all done in a Jupyter Notebook, labelled nupack_analysis_jupyter.ipnyb. Note that part of this analysis includes running the output file from the previous step into a csv to bed converter, labelled csv_to_bed.py. The converted file from the csv to bed conversion is labelled ____. Some files were used from the Wu et al. paper in the Jupyter notebook—specifically, files classifying the start and stop locations of classes 1-4 genes were used, and these files are labelled class_k_genes.csv, where k ranges from 1 to 4.

Statistical analysis was conducted by generating 100 random files of the same size as the prob file that was initially produced and analyzed. These files were then analyzed in another jupyter notebook with similar analysis as the original nupack_analysis_jupyter notebook, and the distributions of the randomly generated sequences were compared to the enhancer sequences to see if the effects seen were statistically significant. The script used to generate the 100 random prob files is labelled control_prob.py, and the analysis can be found in a jupyter notebook called prob_statistical_analysis.

Note: this analysis could not be fully conducted due to some last minute issues, but can be easily generated and run by others.

Results

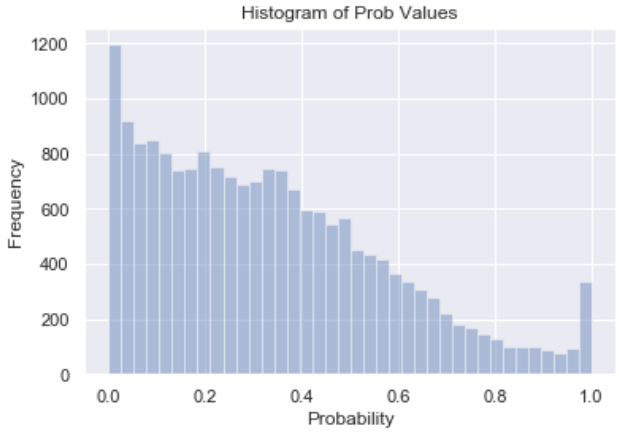
In the results shown below, the prob value of an enhancer refers to the probability that the structure lacks any secondary structure (signified in NUPACK using a string of dots).

Selecting 17 nucleotide analysis window

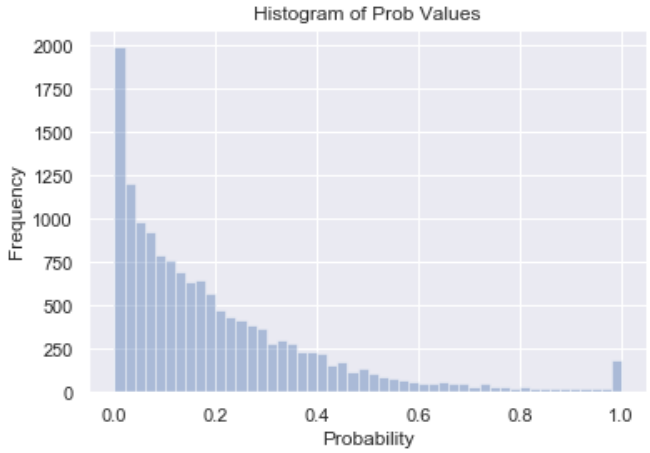
As mentioned earlier, a 17 nucleotide window was selected for the probability analysis conducted with NUPACK's "prob" function. This window was selected because it produced the most distributed probability values. With windows that were larger, the distribution was more left skewed, meaning there was more enhancer slices that had low "prob" values. On the other

hand, with windows that were smaller, the distribution was right skewed. 17 nucleotides didn't produce a perfect distribution, but it was more evenly spread through all "prob" values ranging from 0 to 1.

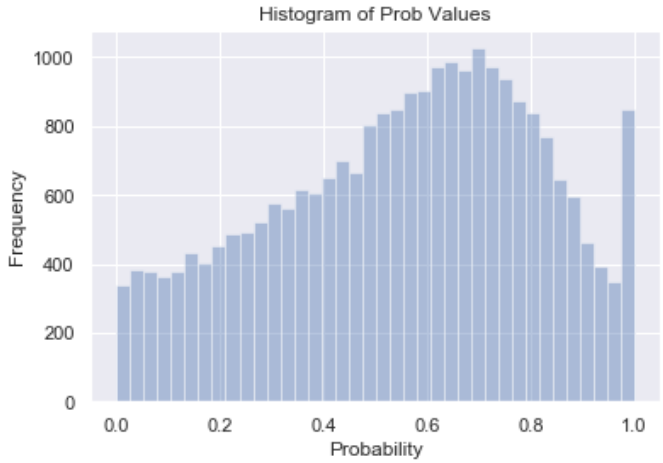
Below is the histogram of "prob" values for a window 20 nucleotides wide:



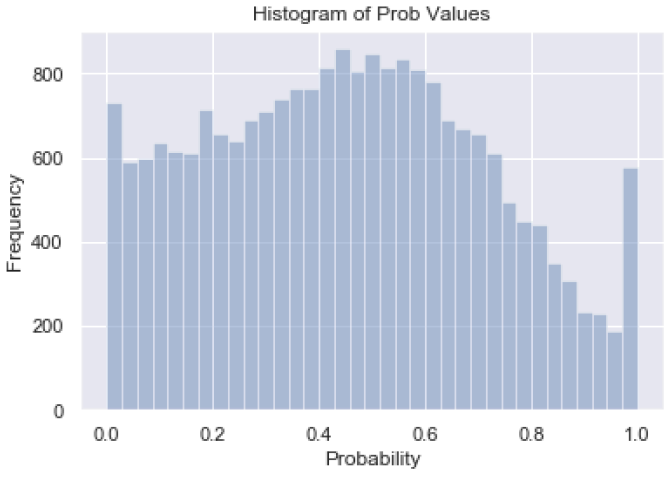
With a 25 nucleotide wide window:



Below is the histogram of "prob" values for a window 15 nucleotides wide:

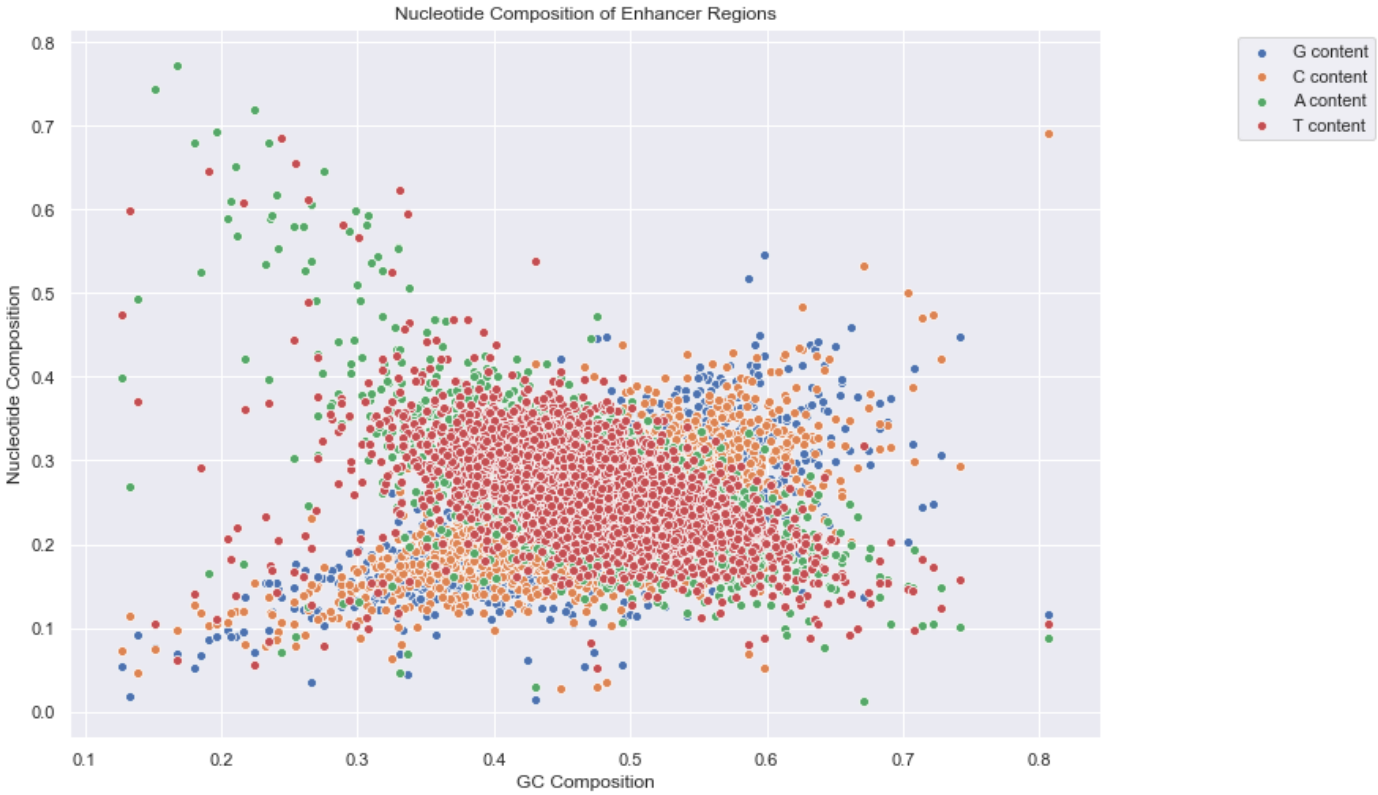


Finally, this is the histogram of "prob" values for a window 17 nucleotides wide:

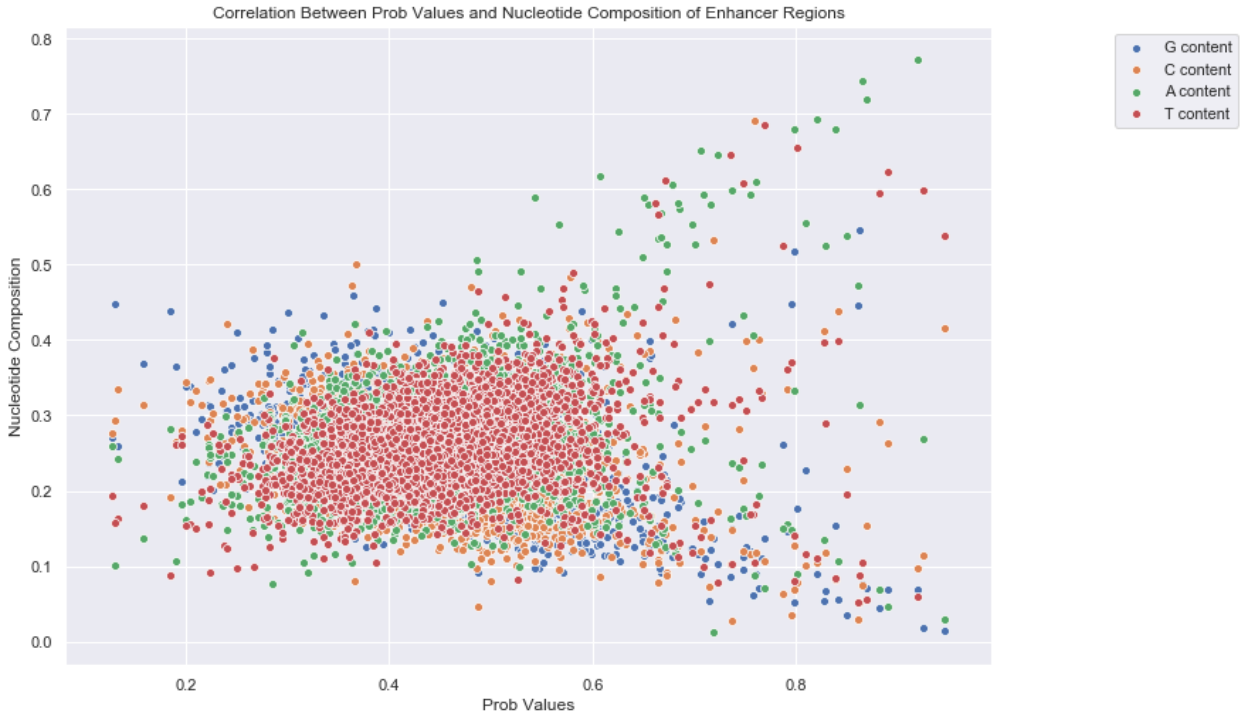


GC Content of Enhancers and Correlation with Probability of Secondary Structure

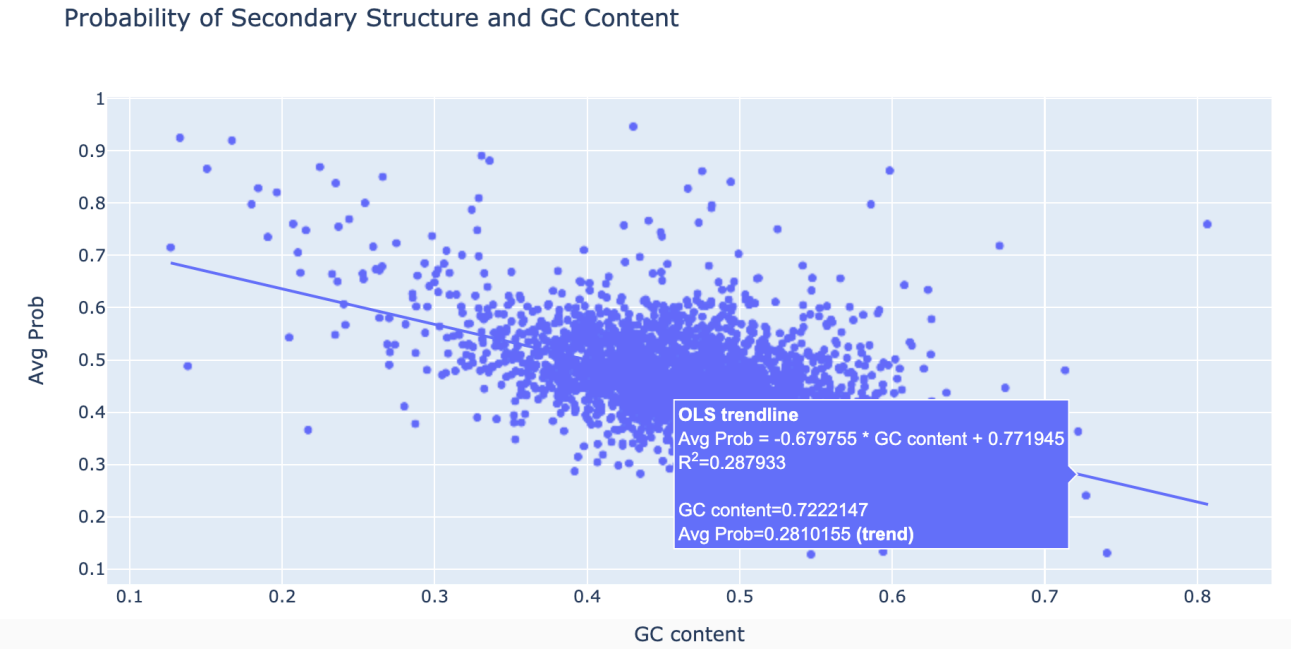
First, I explored nucleotide composition and GC content to see that results were as expected for the enhancers that I explored in this experiment. Here, I plotted the G, C, A, and T content for enhancers and their correlation with the GC content of the enhancer. As expected, G and F content looked to trend positively with GC content, whereas A and T content looked to trend with a negative slope with GC content.



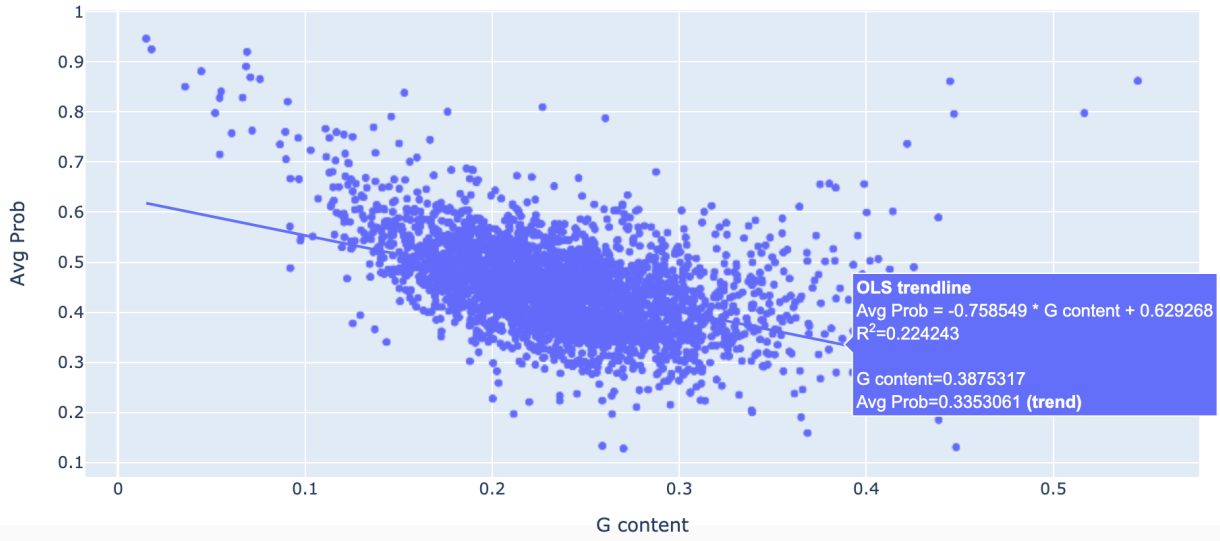
Next, I looked at the correlation between prob values and the overall nucleotide composition of each enhancer. This plot shows all the aspects of nucleotide composition (A, C, G, T content) and the prob values for each enhancer. By examining this plot by eye, there seems to be no strong correlation between the nucleotide composition and the prob values for each enhancer. However, the prob values and seem to be contentrated around 0.2-0.6, and the nucleotide composition (scaled to range from 0 to 1) was concentrated between 0.2-0.4, as shown below.



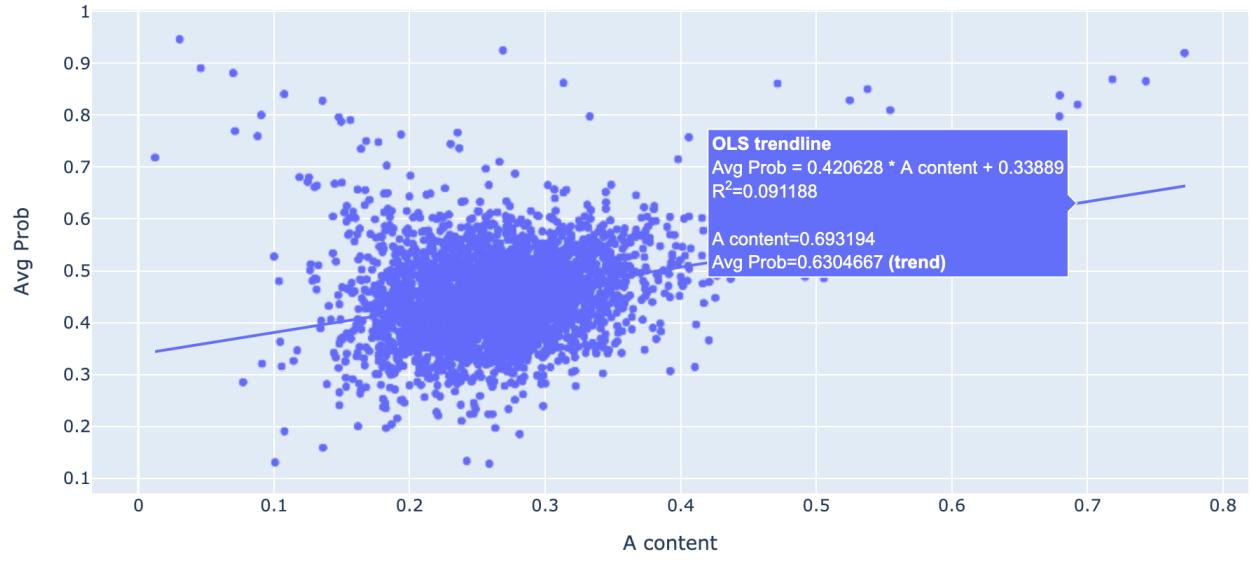
Each of the nucleotides were then explored separately in relation to the average prob value for each enhancer. These plots are shown below, along with a trendline that shows the correlation between the prob values and the nucleotide composition.

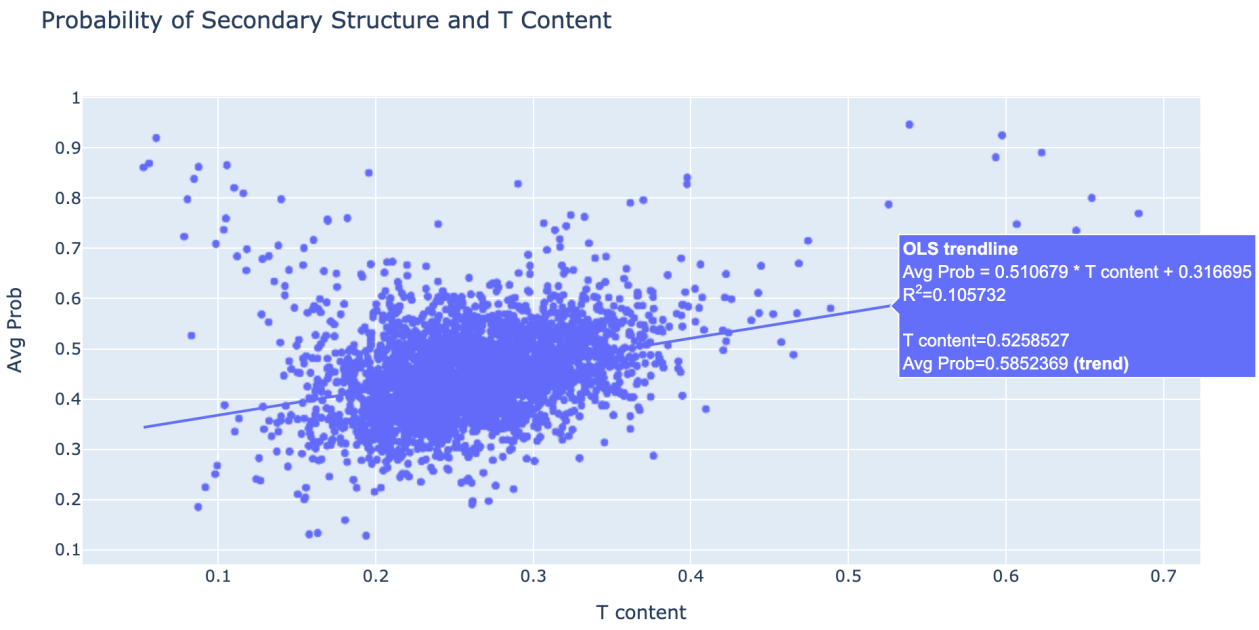


Probability of Secondary Structure and G Content



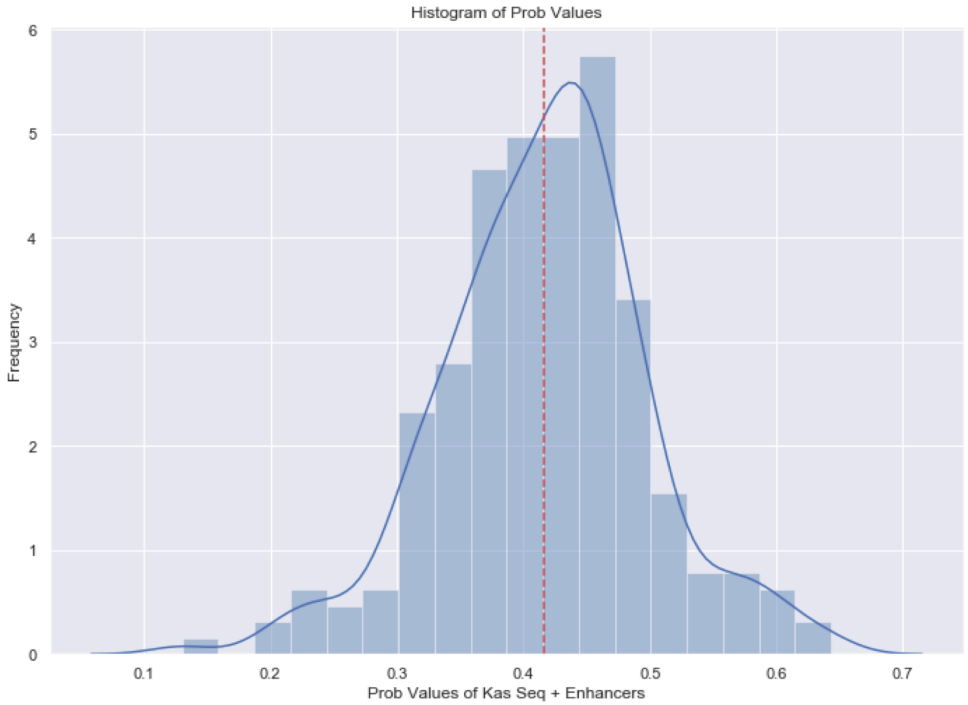
Probability of Secondary Structure and A Content



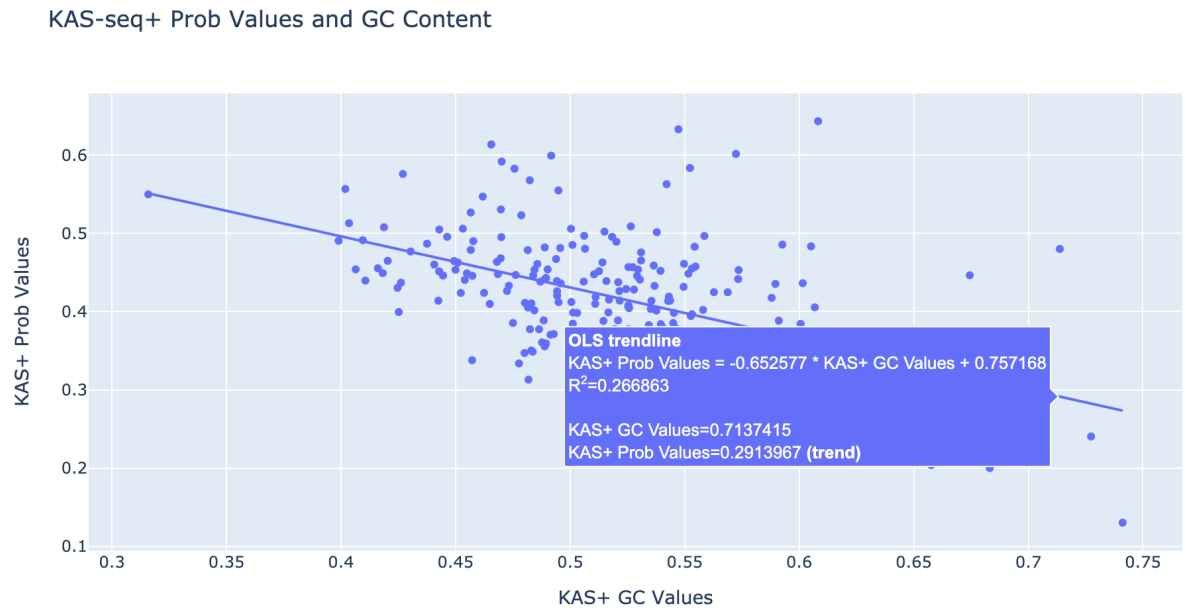


Correlation Between KAS-seq + Enhancers and Probability of Secondary Structure

The probability of enhancers that were KAS-seq+ were calculated and plotted on a histogram, shown below. The prob values for these enhancers were centered around 0.41, and generally these enhancers didn't have prob values that were greater than 0.7 and less than 0.2.

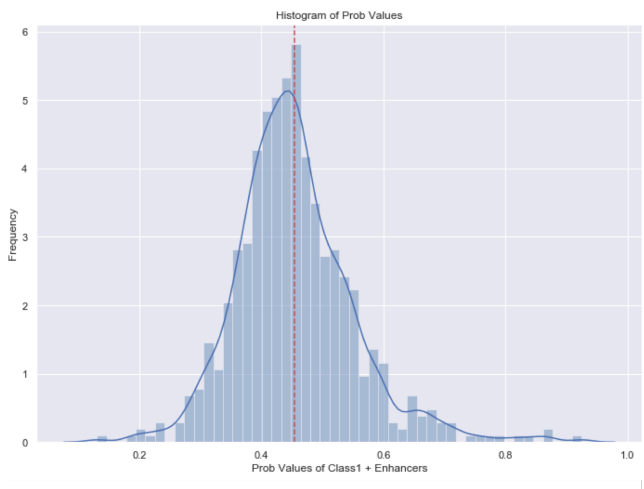


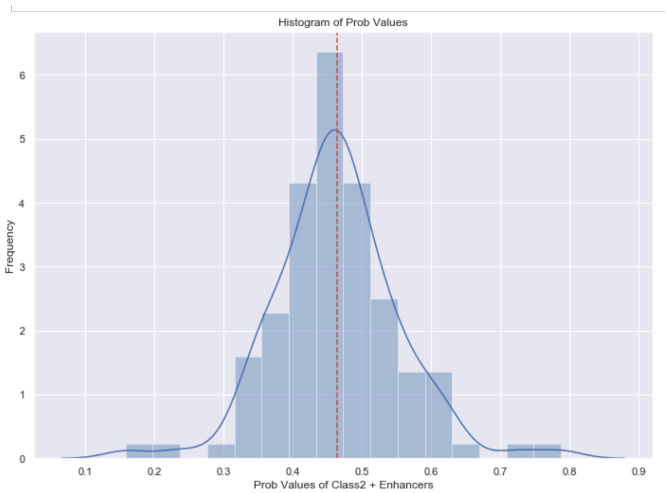
The prob values of KAS-seq+ enhancers correlated with their nucleotide composition, specifically with GC content, as shown below.

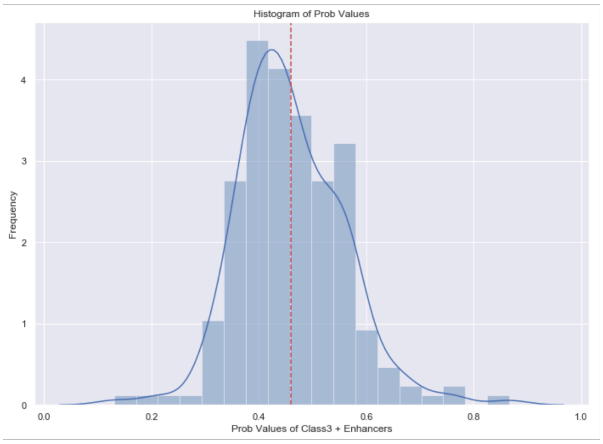


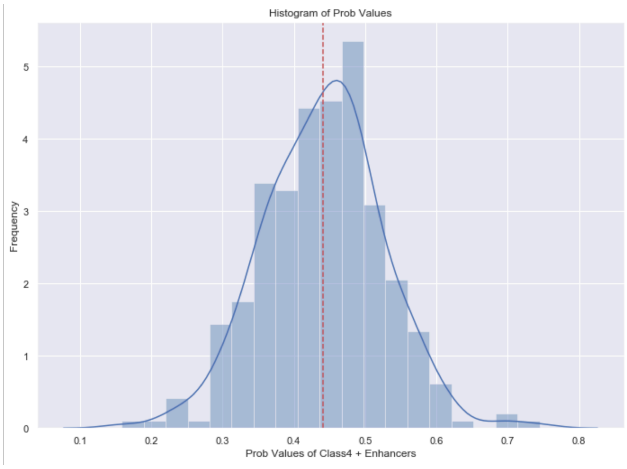
Correlation Between Active, Paused, Inactive, and Not Paused Enhancers and Probability of Secondary Structure

The paper from Wu et al. classified four classes of genes based on their transcription states. Class 1 refers to genes that are paused and active, class 2 refers to paused and inactive genes, class 3 refers to not paused and active genes, and class 4 refers to not paused and inactive genes. Prob values of enhancers from all four classes were plotted on a histogram, and the results are shown below. As seen with KAS-seq + enhancers, the prob values of the enhancers from all four classes show prob values that are centered between 0.4 and 0.5.









Statistical Analysis

The Mann-Whitney U test was performed on the distributions of the prob values seen above. The null hypothesis held for this test was that the distributions are equal, and since the p values fell below alpha = 0.05 for all the distributions, his null hypothesis could be rejected for all the different distributions.

More statistical analysis can be conducted with the provided Python script and Jupyter Notebook.

Discussion

The results seen regarding the window of enhancer length that produced the most equally distributed prob values showed that at higher or lower enhancer lengths, the results skewed dramatically. This shows that small changes in the length changes the probability of secondary structure, or the lack of secondary structure, which can be useful in future analyses using the prob function.

The correlation between prob values and nucleotide composition, as seen above, is as expected. As GC content increased, the prob values of the enhancers generally decreased, because higher GC content generally has been shown to be correlated with increased secondary structure formation. Since the prob value represents lack of secondary structure, low prob value represents a high probability of secondary structure formation. On the other hand, A and T content increasing showed higher prob values, signifying that high AT content in an enhancer results in lower probability of secondary structure formation.

The prob function is does not fully encapsulate biological conditions, as it only requires specification of temperature, salt concentration, and magnesium concentration. These were set to 37 degrees Celsius, 0.157 M, and 0 M, respectively. There could be many more aspects of biological conditions that can influence DNA folding and formation of secondary structure, which could result in differences in correlation seen with the in silico and in situ results using KAS-seq and prob. Adding more qualifiers into the calculation of the probability of secondary structure formation, could influence results and clearer correlation to in situ methods.

This analysis can be extended to cover the rest of the chromosomes in mm10, as well as other organisms to explore similar correlations and to run more statistical analyses.

References

1. Wu, T., Lyu, R., You, Q., & He, C. (2020). Author Correction: Kethoxal-assisted single-stranded DNA sequencing captures global transcription dynamics and enhancer activity in situ. *Nature Methods, 17*(7), 749-749. doi:10.1038/s41592-020-0881-1

2. Dalby, M., Rennie, S., & Andersson, R. (2018, September 07). FANTOM5 transcribed enhancers in mm10. Retrieved August 27, 2020, from https://zenodo.org/record/1411211

Appendix

A list of all the analyses conducted in the Jupyter Notebook can be found below:

1. Histogram of prob values
2. Correlation between length and prob
3. Correlation between prob and G content
4. Correlation between prob and C content
5. Correlation between prob and GC content
6. Correlation between prob and A content
7. Correlation between prob and T content
8. Correlation between kasPeaks and prob values
9. Correlation between class 1 genes (See Wu, et al) and prob values
10. Correlation between class 2 genes and prob values
11. Correlation between class 3 genes and prob values
12. Correlation between class 4 genes and prob values
13. Heatmap of different analyses
14. Correlation between kasPeaks, length, and prob value
15. Correlation between kasPeaks, G content, and prob value
16. Correlation between GC content and Nucleotide Composition (G, C, A, and T content)
17. Correlation between class 1 genes, GC content, and prob values
18. Correlation between class 2 genes, GC content, and prob values
19. Correlation between class 3 genes, GC content, and prob values
20. Correlation between class 4 genes, GC content, and prob values