Joshua Rhoades
“77-jrhoades”
MAST697
Python Programming
Final Project
13DEC16

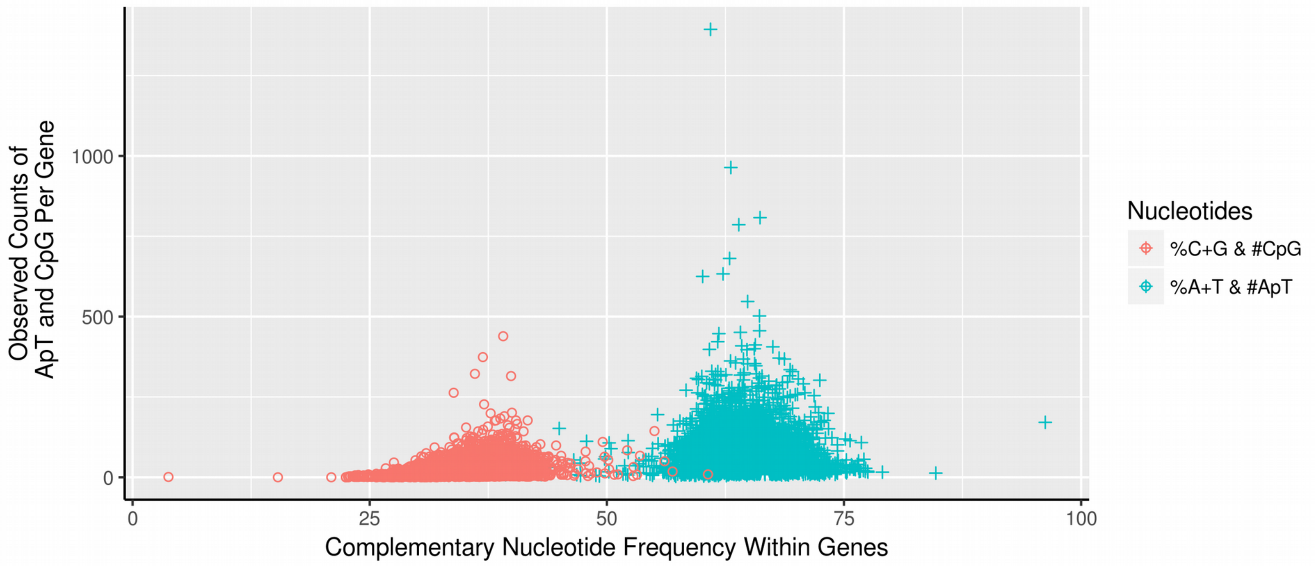
I chose one of the *Bacillus thuringiensis* genomes, as I have a background in Entomology and Agriculture research and have regulatory bioinformatic experience with various Cry proteins and their role in transgenic crops. The first part of the assignment calls for the creation of four figures, which I was able to combine into two.

My first figure displays the observed count of ApT and CpG motifs on the Y-axis and the sum frequency of complementary nucleotides within genes on the X-axis. A normal distribution of both A+T and C+G nucleotide frequencies around their means is displayed. The observed counts of CpG are generally lower than ApT, possibly due to the higher frequency of A+T in the genome. There are several outliers, having either high ApT counts, or an extreme nucleotide frequency. These outlying genes may be interesting areas of investigation.

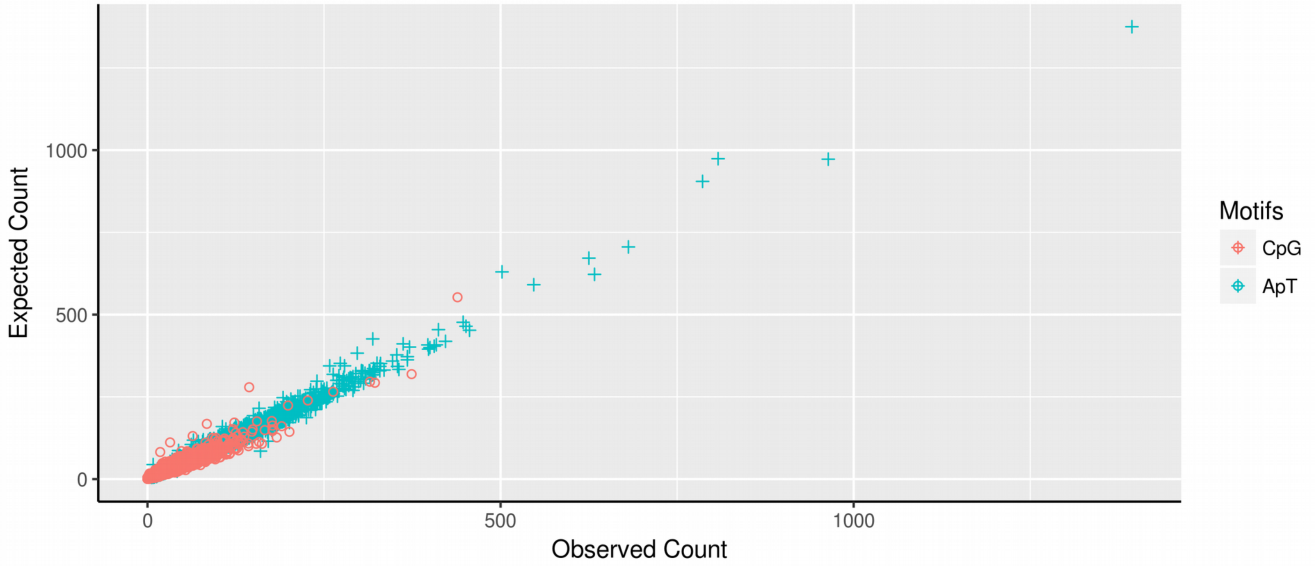
The second figure displays expected vs observed counts of ApT and CpG. The data displays a linear 1:1 relationship, suggesting that these two variables are in agreement and expected counts may be an accurate predictor of observed counts. However there are two interesting trends in this figure. First, the two motifs cluster at slightly different places, demonstrating the higher frequency of ApT motifs in the genome, again possibly due to the higher A+T nucleotide percentage. The second trend is that both CpG and ApT have outliers above their clusters, but still on a near 1:1 linear relationship.

The third figure compares the observed count of 2mer amino acids within CxxC motifs to the expected frequency of all 2mer amino acid combinations across the *Bacillus thuringiensis* proteome. The more frequent a 2mer, the more often it should theoretically appear within a CxxC motif, conversely the less frequent 2mer’s should be found in CxxC motifs less often. When we look at the third figure, that principle hold true. The linear model returns a narrow confidence band, suggesting that it is a good fit for the data. The positive slope of the line agrees with the hypothesis that as frequency of 2mers increases, so does the frequency of it being found within CxxC motifs. Lastly, we note some very high observed counts of 2mers with low to moderate expected frequencies. These common 2mers within CxxC motifs may have biologic relevance and interesting targets for future investigation.

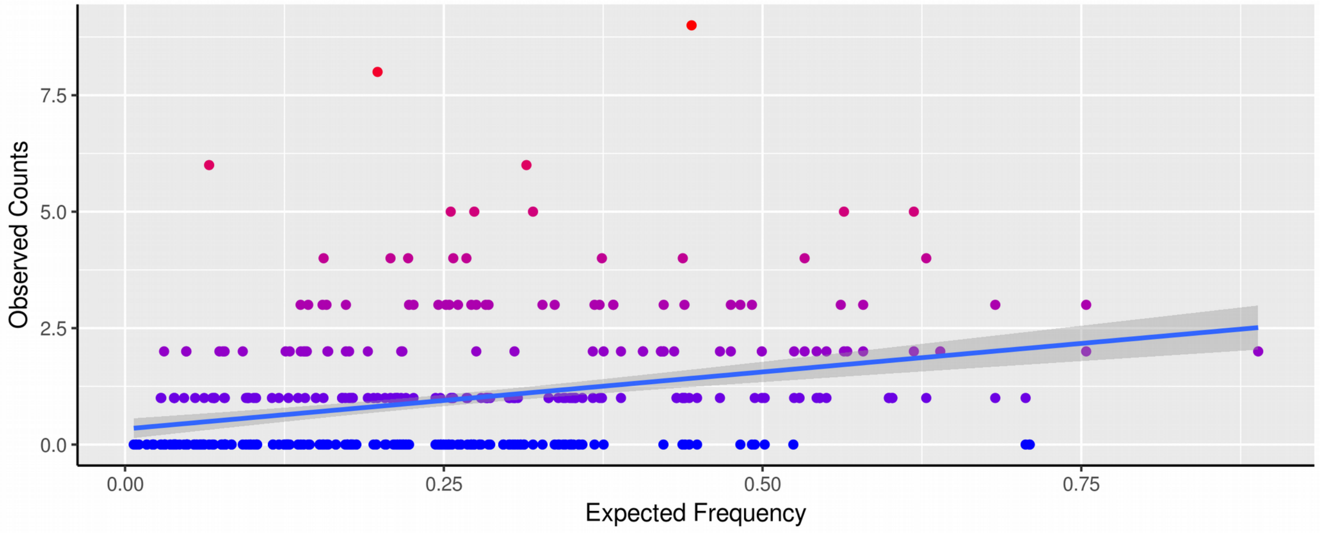
Bacillus_thuringiensis_YBT_1518_uid229419 Gene Compostion



Bacillus_thuringiensis_YBT_1518_uid229419 Genome Observed vs Expected Motif Counts



Expected Frequency vs Observed Counts of all possible 2mer Amino Acid Combinations
 within CxxC Motifs in the Bacillus_thuringiensis_YBT_1518_uid229419 Genome


Observed Counts
Expected Frequency