In this project I created a program to find under- and over-represented DNA palindromes (or k-mers). DNA palindromes are k-mers that are the same as their reverse compliment (the middle letter will not match when k is odd, but we still count this as a DNA palindrome if all other letters match). We used the null model that gave an expected count, $E(C(W))$, for the palindrome or k-mer $W = W_1W_2 \dots W_n$, where

$$E(C(W)) = \frac{\text{counts}[W_1W_2 \dots W_{n-1}] \cdot \text{counts}[W_2 \dots W_{n-1}W_n]}{\text{counts}[W_2 \dots W_{n-1}]}$$

and counts is the observed counts for the different subwords.

The program worked by first generating a list of DNA palindromes (or just k-mers, if –all was specified in the command line) where $\min_k \leqslant k \leqslant \max_k$ ($\min_k$ and $\max_k$ were specified by the user). After this a dictionary containing k-mer keys and counts as values was created from one or more fasta files (any of which could have been gzipped), where $\min_k - 2 \leqslant k \leqslant \max_k$ (this is all that is needed for the null model). Once a dictionary of counts was produced, statistics for each of the palindromes (or k-mers) produced in the first part were tested for under- or over-representation. This was done using the observed and expected counts to compute a Z-score. Using the erfc function, this was then converted to a p-value, which was then converted to an e-value. If this e-value was below a certain threshold (specified by the user) it was outputted.

There were a couple of interesting phenomena that I observed running this program. First I ran the program on the file "H.influenzae.fa.gz" ($\max_e$ was set at 0.01 for all tests). At first I only looked at palindromes of length 6. When I looked at palindromes of length 5 I noticed that there was an interesting symmetry. There were many 5-mres that had the same observed and expected value. For example, both CACTG and CAGTC occurred 2128 times and according to the null model we expected to see this palindromes 2660.08 times. This appear to happen for all odd k-mers, where the k-mers with conjugate middle bases had the same expected and observed counts. This also occurred when looking at the files "P.abyssi.fa.gz," "P.furiosus.fa.gz," and "P.horikoshii.fa.ga." This would motivate future updates to the program to include an option to merge counts for words with either a wild card base for all four nucleic bases, or perhaps to merge A and T bases into an S and C and G bases into a W, as this was observed when the middle bases were conjugates.

Another interesting observation is that in both the "P.abyssi.fa.gz," "P.furiosus.fa.gz," "P.horikoshii.fa.ga" data sets and in the "H.influenzae.fa.gz" data sets that the palindrome "TATA" was under-represented. This is surprising as this palindrome is a very import one used in transcription. However in "H.influenzae.fa.gz" it was only observed 12264 times, while we expected to see it 17624.75 times. In "P.abyssi.fa.gz," "P.furiosus.fa.gz," and "P.horikoshii.fa.ga" we only saw it 20674 times, but expected to see it 22628.61 times. Furthermore, when testing for palindromes of lengths 3 and 4 in "H.influenzae.fa.gz", we find that 27 out of 32 are under-represented. When testing for palindromes of lengths 3 and 4 in "P.abyssi.fa.gz," "P.furiosus.fa.gz," and "P.horikoshii.fa.ga," we find that 28 out of 32 are under-represented. This would suggest that this null model may need to be revamped, as the majority of palindromes of length 3 and 4 and under- or over-represented (including import ones like TATA).