**Case Study 3: Search for Unusual Cluster in the Palindromes**

*Math 189 Homework 3*

Aditya Anandkumar

Andrea Cabezas

Bethan Lilly Wynne-Cattanach

Carlo Mazzafaro

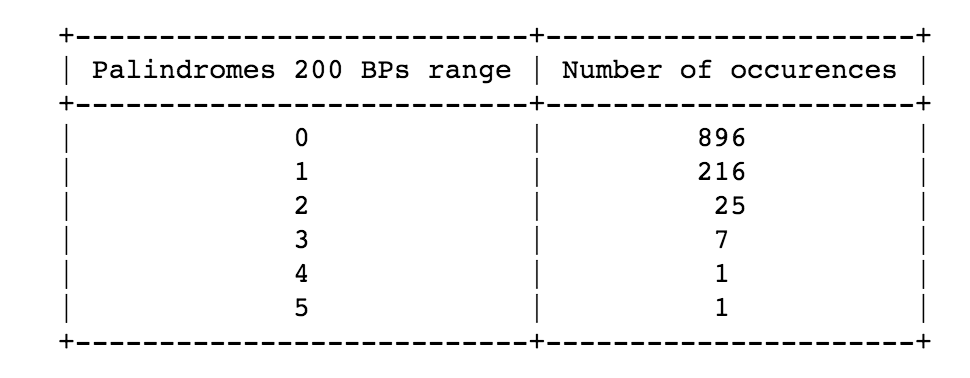
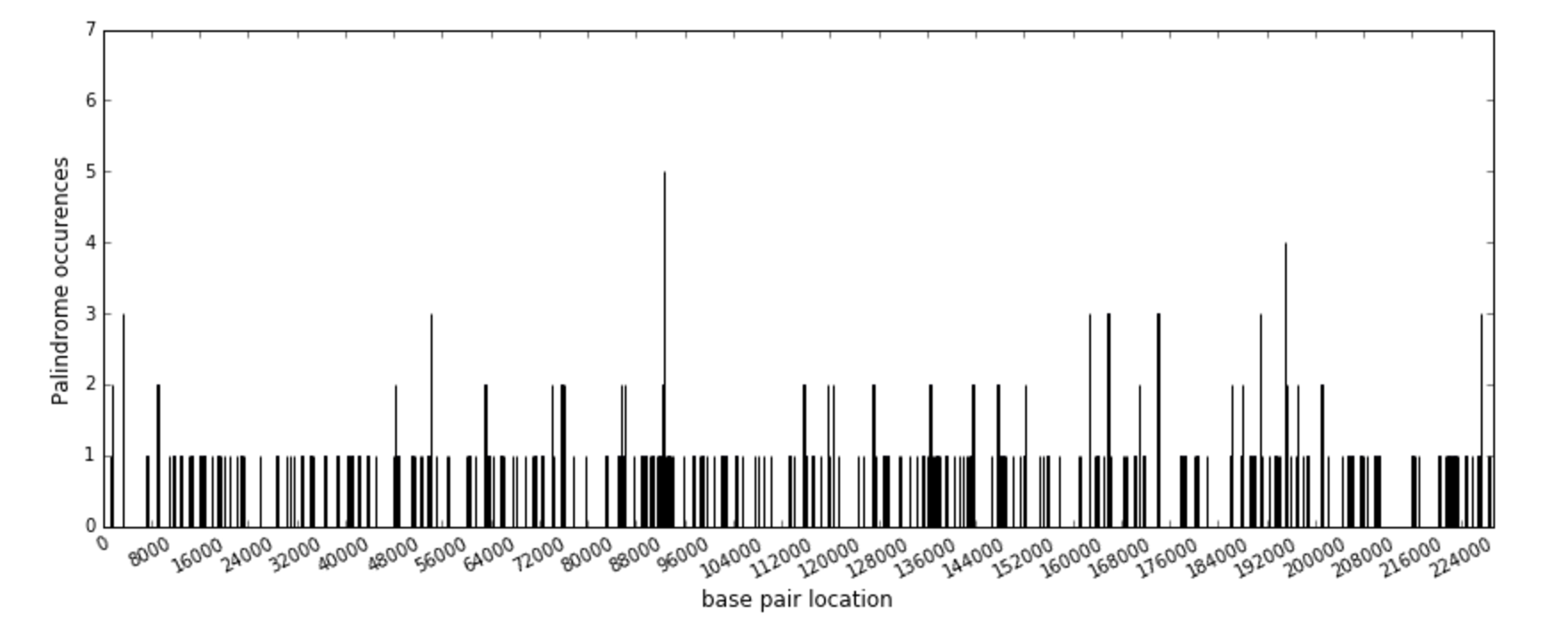
Jake Ehlers

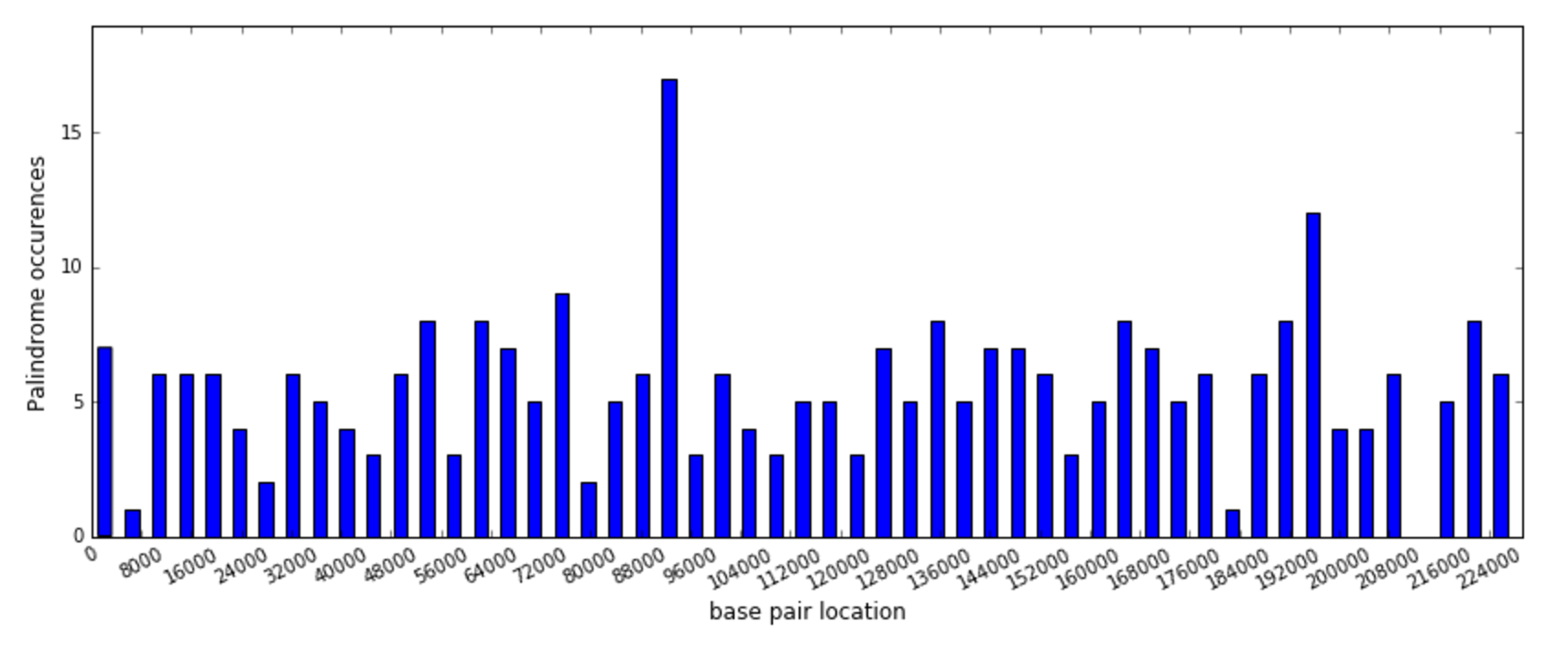
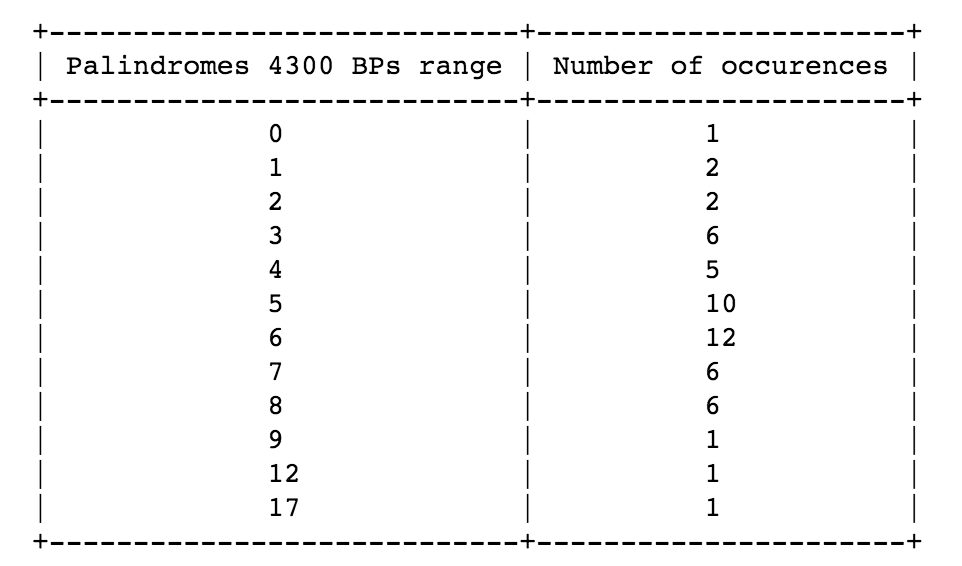
Yi Ma

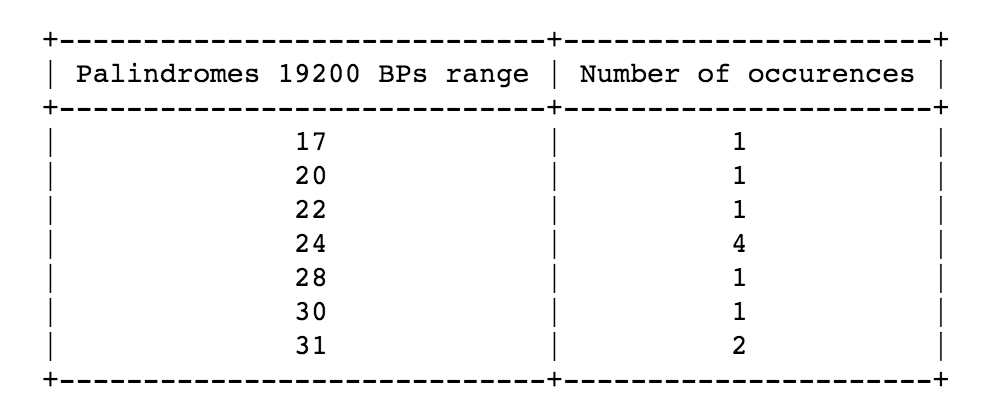
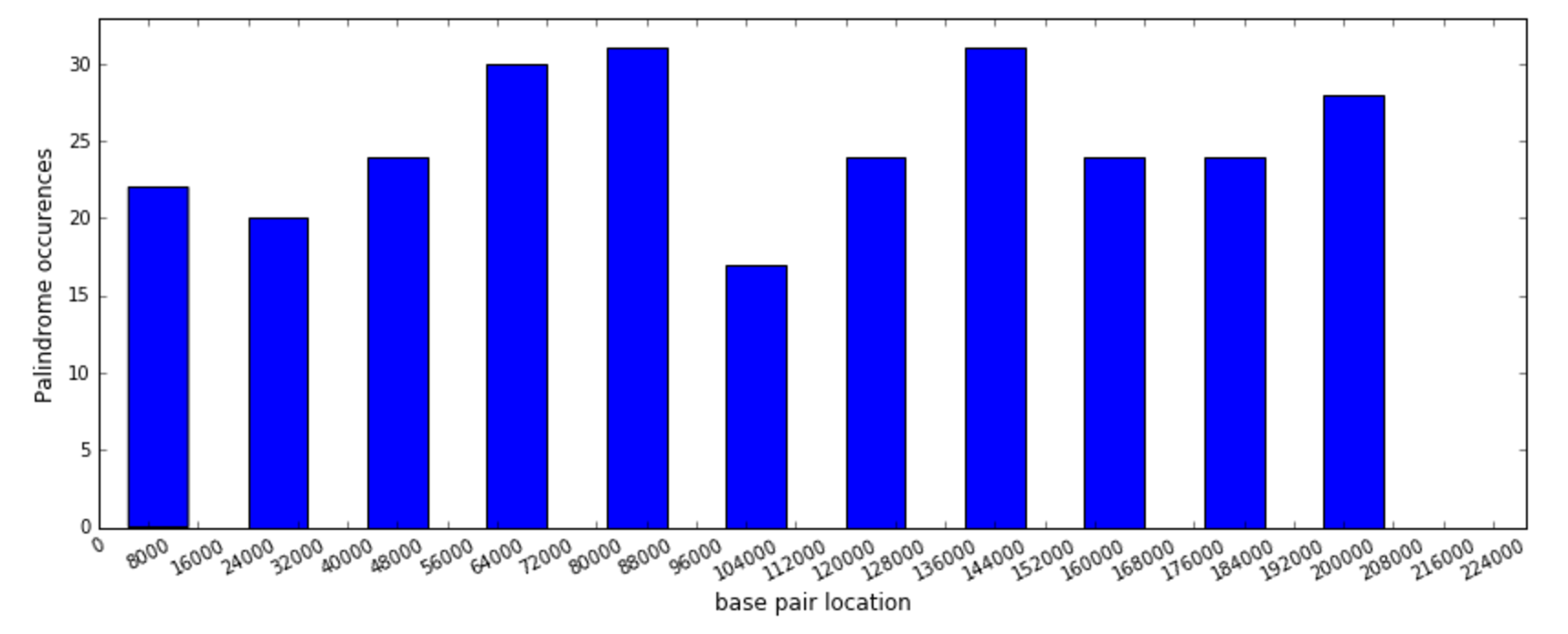
**Introduction**

Our study determines the DNA sequence intervals where biologists are most likely to find the replicative origins of the CMV virus. Following the hypothesis that the replicative origins of viruses in the congruent family are signaled by large palindrome clusters, we inspected the largest of such concentrations within the virus’ DNA sequence. We also simulated random scatters of palindromes, and confirmed that the probability of our highlighted clusters were low under those distributions to insure their significance. To determine an appropriate distribution for our data, we conducted tests to gauge its fit with the Poisson, Exponential, and Gamma distributions. After confirming the non-randomness of our large clusters, we performed additional analysis on the distances and counts of palindromes to reinforce our results. We close our case study with a brief discussion on the coherence of our findings with greater academia and our recommendations for further research.

**Data and Method**

Our data consists of the CMV Virus’ DNA sequence, which is composed of 229,354 base pairs. 296 palindromes were identified that were at least 10 letters long – the longest ones were 18 letters long and found in moments 14719, 75812, 90763, and 173893 of the sequence. Palindromes shorter than 10 letters were ignored. We begin our study by perusing palindrome frequencies across the sequence (‘bins’), which can be viewed below. In order to determine the best sequence size in which to group the DNA base pairs, we drew histograms containing the frequency of palindrome occurrence per interval using intervals of varying size ranging from 200 base pairs to 38400 base pairs. This will give us some preliminary idea of which bin size will work best and also increase our odds of finding a cluster. Represented here are the histograms for bin sizes of 200, 4300 and 19400 base pairs.



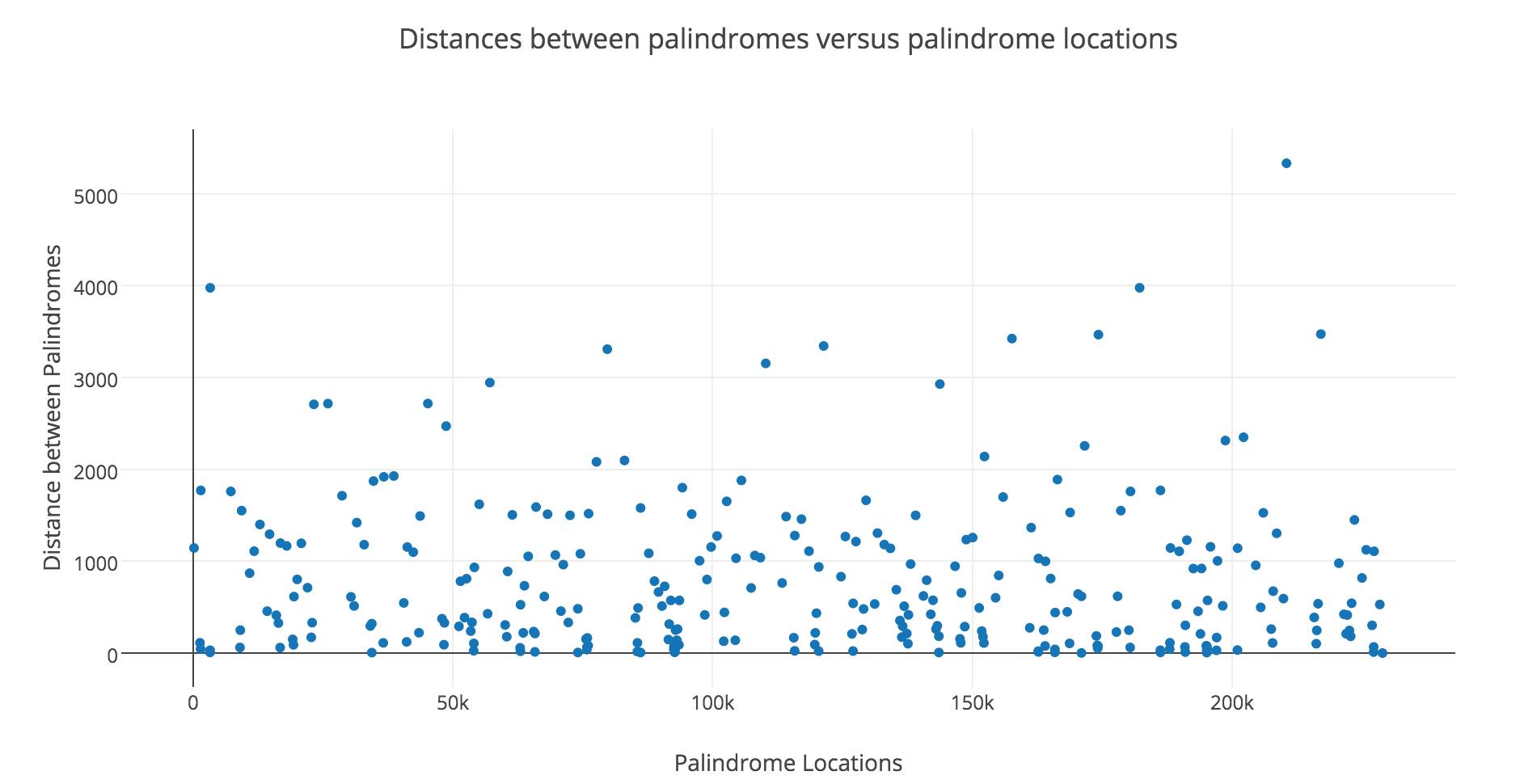


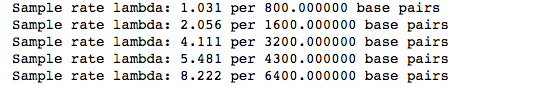
It is clear that the first and last plots do not provide any insightful information – the number of palindromes per bins is either too small or too large. However, by observing the plots for other bin sizes (omitted), it was possible to clearly distinguish an unusually high number of palindromes grouped around base pair 88000 regardless of the selected bin size. To verify the non-randomness of that congregation, we determine the distribution of our data and assess the probability that such a cluster would occur in that congregation. In our analysis we will focus on using bin sizes such as the one in the middle (4300 base pairs per bin).

**Random Scatter**To begin with we want to see how our data compares to the random uniform scatter, to see whether or not our initial guess of the palindrome cluster at 88000 is unusual.   
First, we used the data for the locations of the palindromes. We simulated 296 random locations, using a random number generator between 1 and 229354. To compare the data we plotted our random simulations against the actual data, and tested for linearity.

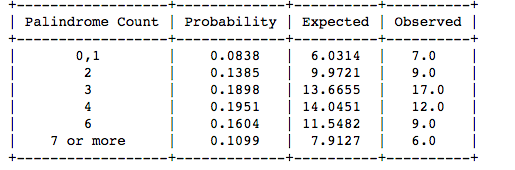
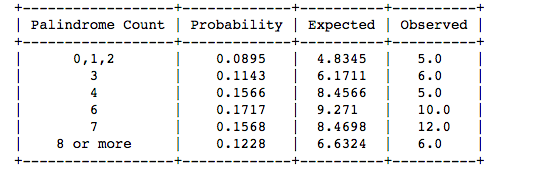
We found the correlation coefficient for the relationship between our data and the random data to be 0.9974, strongly suggesting our data is related to a uniform random scatter and that it is almost random.

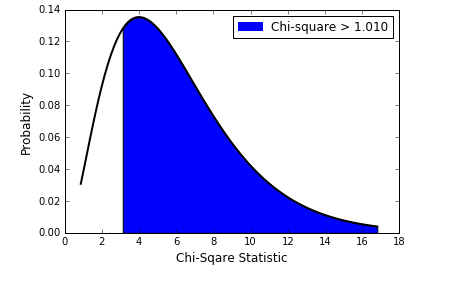
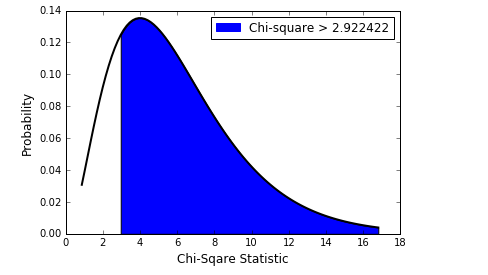
**Locations and Spacings**Here we have plotted a graph of the spacing between consecutive palindromes according to their locations.

  
This graph does indicate that around the 88000 mark again, we see smaller distances between the palindromes, suggesting again that this could be an unusual cluster.   
Although the methods used so have do suggest that we do have a cluster at approximately the 88000th base pair, the results are not very precise. So we examined different distributions and performed more formal statistical tests to compare to our data.   
**Randomized Models of Distribution**

Palindrome occurrence shares a number of similarities with Poisson processes, the most notable being that palindromes are scattered randomly across DNA, their frequencies in different regions are independent of each other, and their probabilities of materializing are uniform across sequences. The first potential distribution we examine is the Poisson distribution. We believe this distribution models the locations of the palindromes. The Method of Moments (MM), as well as the Maximum Likelihood Estimation (MLE) (appendix, 7) both yield the same result when the estimation of the rate parameter lambda is attempted. The lambdas for the different interval sizes are listed below.

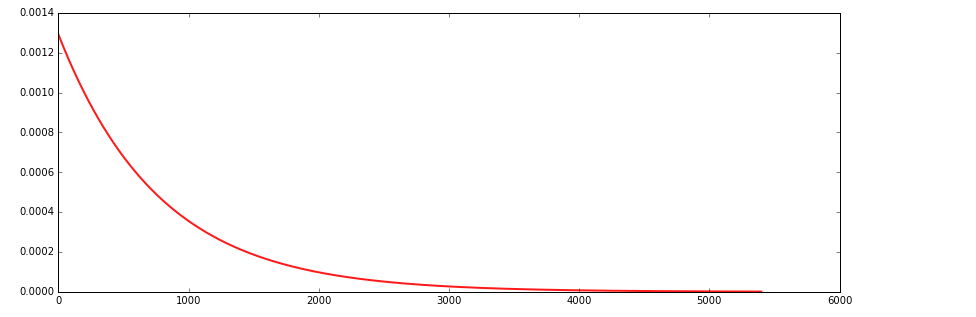
Using the above rate parameters and the Poisson probability mass function, we derived the expected palindrome frequencies for the 3,200 and 4,300 bin sizes from a truly random Poisson process having the same parameters as the ones we found, and produced the comparable tables below.

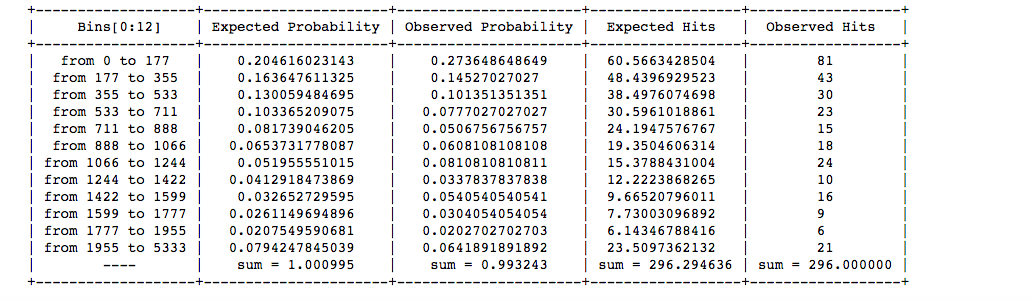


****The expected values we derived also allow us to perform a Chi-square test, which we conducted with both bin sizes.

The test with bin size 3200 yielded a chi-square statistic of 2.922 and p-value of .819, whereas its counterpart with bin size 4200 yielded a chi-square statistic of 3.146 and a p-value of .79. Unsurprisingly, both results suggest our data may be driven by a Poisson process, albeit not definitively. This demonstrates that our data may possess some non-randomness.

**Exponential Distribution and the Distance between Palindromes**

To further assess our data, we compared whether the distances between palindromes in our sequence fit the spread of an exponential distribution. We began this process by first deriving a rate parameter by dividing the number of palindromes with the total number of base pairs, and arrived at a value of .00129. That parameter was then modeled into an exponential distribution, and we discovered that the expected distance between palindromes for an exponentially distributed data set with our rate parameter is 774.84459. This result fits well with the mean distance observed in our data, which was 772.89189, and so we decided to perform a chi-square test. To derive the expected values necessary for the chi-square test, we plotted the distances of an exponentially distributed random variable with our rate parameter of .00129.

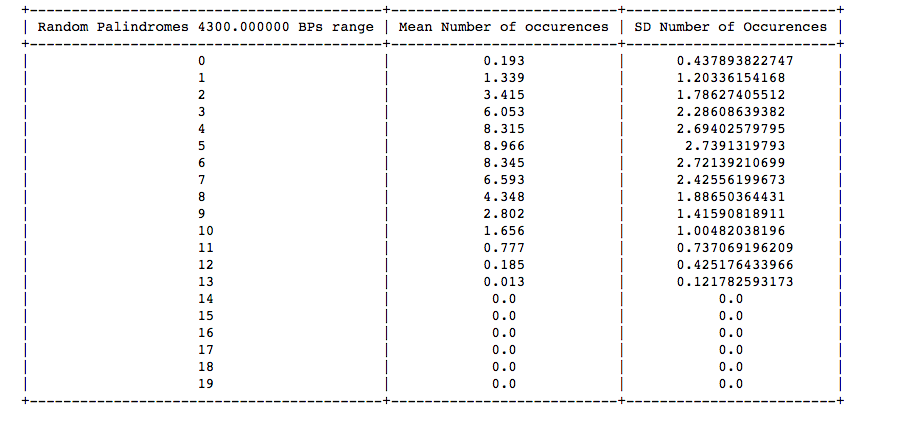
Because the occurrences at the tail of the distribution are extremely low, we grouped the results into 12 bins. We then assessed the coherence between our observed data and the expected values of the exponential distribution above.

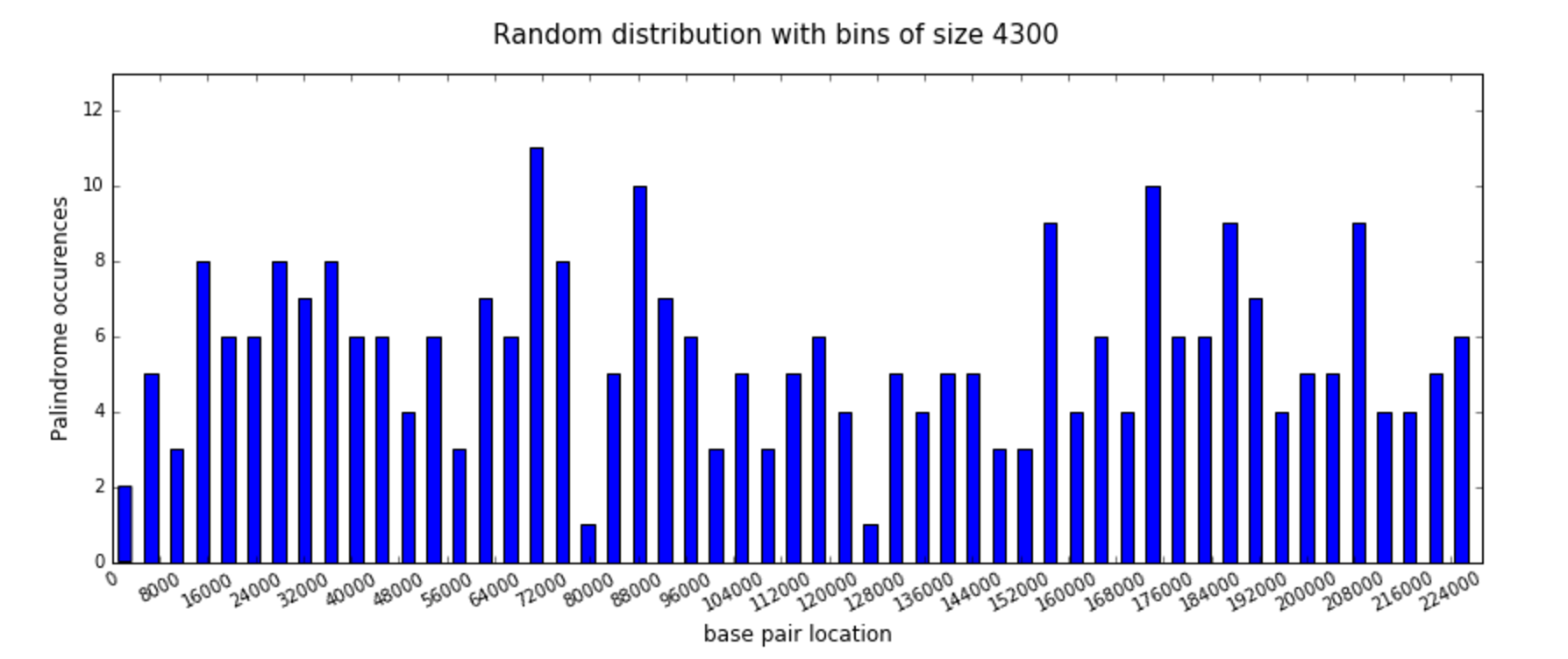
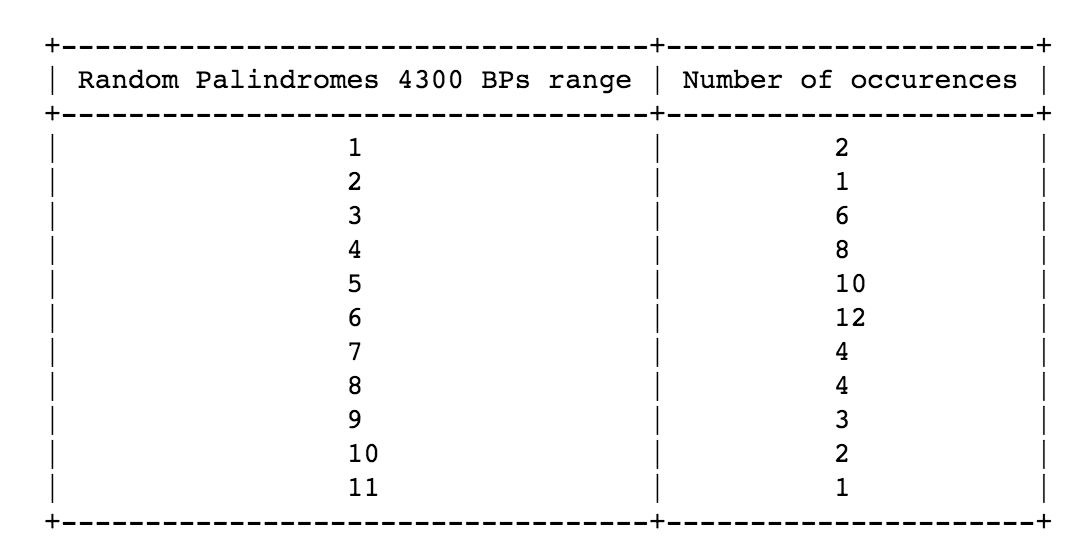
Having derived the necessary expected values, we performed a chi-square test to determine if the distances between palindromes in our sequence were exponential distributed. Surprisingly, the test yielded the contrary result. Our chi-square statistic was 24.723683, and our p-value was .010004. This rejects our null hypothesis of the distances being exponentially distributed. However, it turns out the biggest contributor to the chi-square statistic was the large number of hits observed for small distances, particularly for the range from 0 to 177 (81 observed versus 60.6 expected). This is a strong indicator that clusters are present. Having obtained evidence for clustering, we decided not to proceed with a gauge of our data’s fit with the gamma distribution.

**Construction of the Random Poisson Distribution**

We’ve now rejected the hypothesis that the palindromes are exponentially distributed, but the assumption that the palindromes arise from a Poisson distribution still holds (chi-square statistic from our test was not significant). Hence, we now construct the random Poisson distribution that resembles the specifics provided in our data. To begin, we randomly selected 262 numbers within the range of 90 and 229,354 in order to represent a random distribution of palindromes. The distribution for the 4,300 base pair bin size can be viewed below.

To get an idea of how many palindromes to expect in the 3200 and 4300 intervals, we used a bootstrap procedure with 1000 iterations. To derive the expected values necessary for the chi-square test, we plotted the distances for an exponentially distributed random variable with our rate parameter. Below is a table displaying the expected number of palindromes for the 4,300 base pair intervals.





From this data, we can see that the probability of having 13 or more palindromes in a single bin of size 4300 is extremely small (0.019). In our data, there’s one specific outlier of 17 palindromes in one single bin, which in a Poisson random process repeated 1000 times happened with probability less than .02. From this test, and from the fact that the data does not obey an exponential distribution, we can infer with a high degree of certainty (p<0.02) that the data contains a cluster around the base pair 88000 range.

**Analysis**

Our results indicate an unusually large cluster of palindromes around the 89th thousand and 96th thousand base pair. The non-randomness of that cluster is affirmed by our resampling bootstrap technique after determining their best approximated theoretical distribution (Poisson in this case). We affirmed that our data resembles this distribution through chi square tests. The results stated that the probability for a random Poisson distribution to display a cluster as large as ours is less than 0.02%. We further established the presence of a cluster by observing an overwhelming number of palindromes being closely spaced in a fashion not predictable by the exponential distribution, the model used to predict the spacing between random Poisson event.

**Relation to Greater Research**

We were pleased to discover that a cohort of researchers from Stanford University led by professor Marie Masse in 1992 arrived at results that mirrored our own. While the group used a much larger array of scientific and statistical tools, they conducted their analyses on the same DNA sequence and also filtered for palindromes of 10 letters or more. The group concluded that the unusually large palindrome cluster occurring between base pair 92,000 and 94,000 was responsible for the virus’ replicative mechanisms (Masse, 48).

**Conclusion**

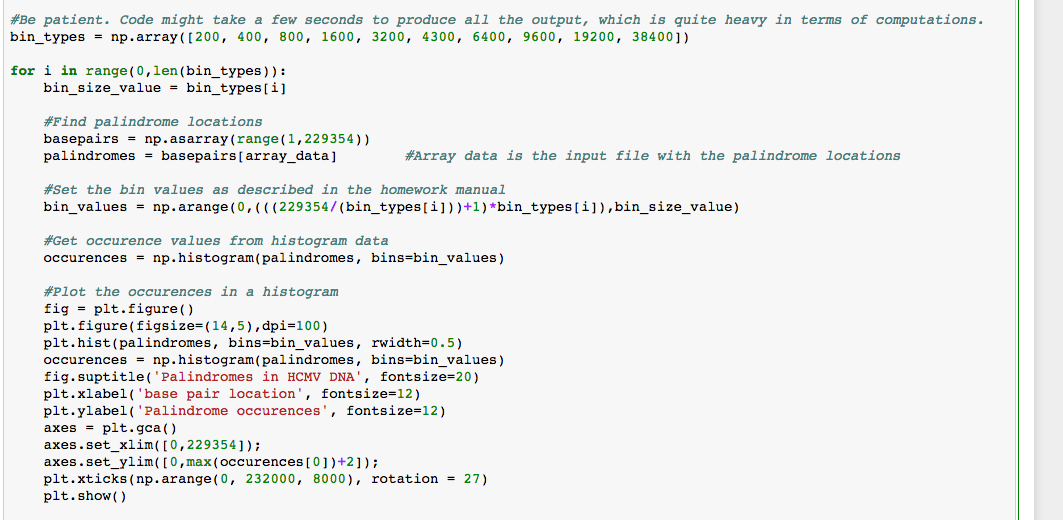
While we have confirmed the presence of a large palindrome cluster in the 89,000 and 96,000 base pair interval, the exact role that the cluster structure plays in the replication mechanism is still unclear – we recommend using our results as a base for this next step in research. Afterall, even though searching within our recommended interval may save biologists ample resources, their efforts may still be for nought if the CMV virus has replicative mechanisms that differ from those of its family viruses. Therefore, we believe the most crucial step moving forward in battling the virus is the illumination of the relationship between large palindrome clusters and the replicative origins of the Herpes simplex.

**Appendix**

1) The size of the DNA was 229,354 base pairs with the last palindrome occurring at 228,954. We did not want to exclude this from our analysis, and hence we increased the bin sizes to slightly more than 4000.

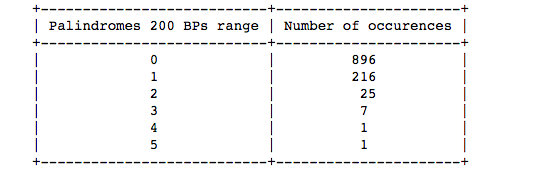
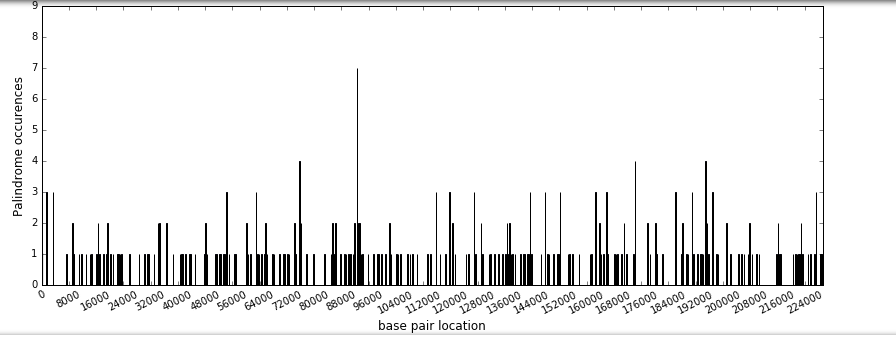
200, 400, 800, 1600, 3200, 4300, 6400, 9600, 19200, 38400 were the sizes of our bins used, and this gave us a preliminary idea of which bin size would work best

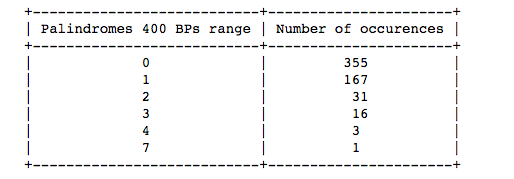
We ran a code to create a table of the number of occurrences of palindromes, and to plot the occurrences in a histogram, for every bin size used.

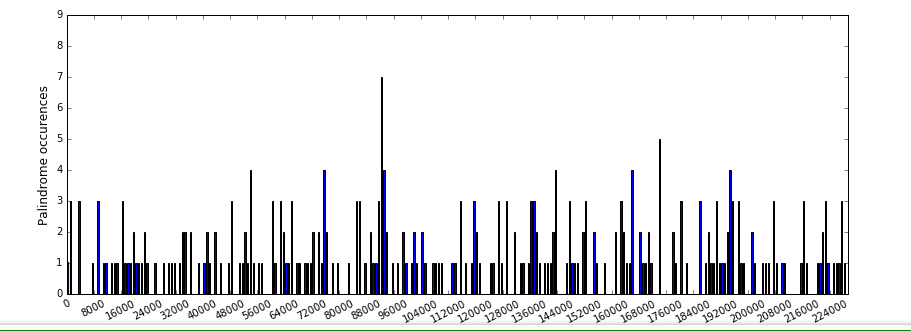
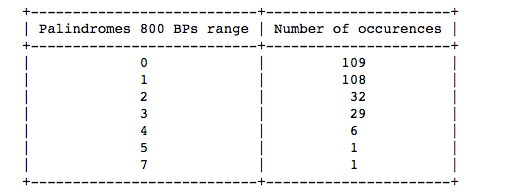
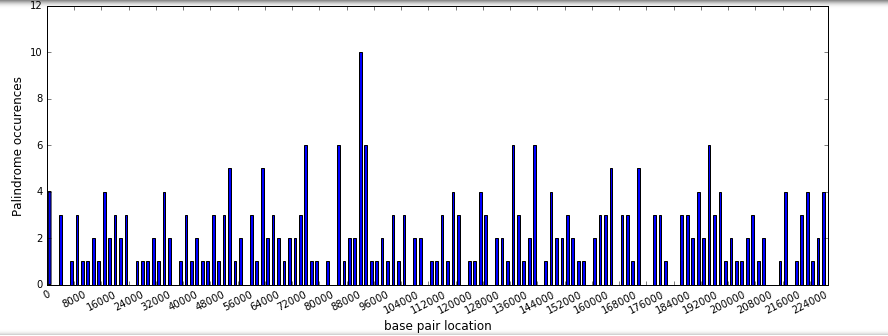
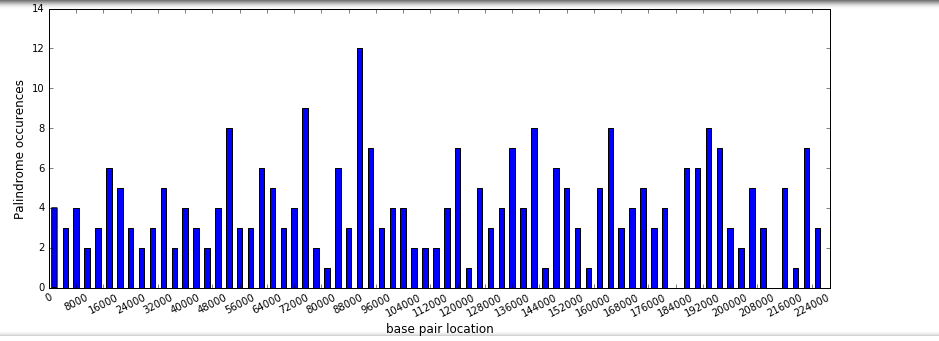
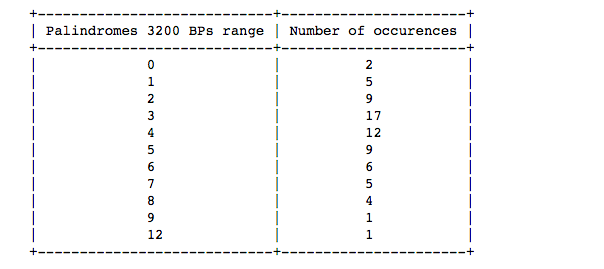
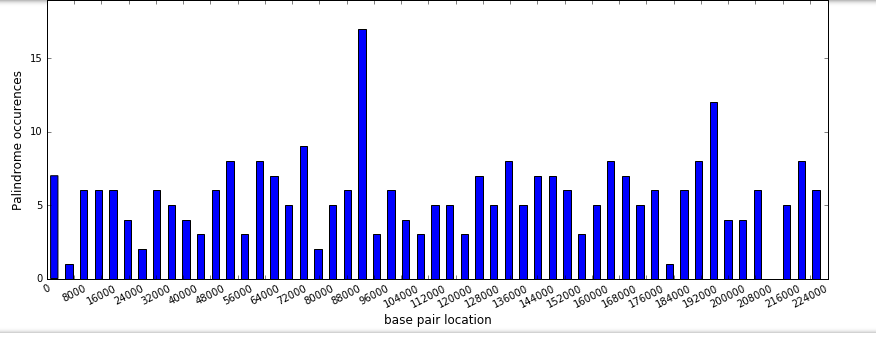
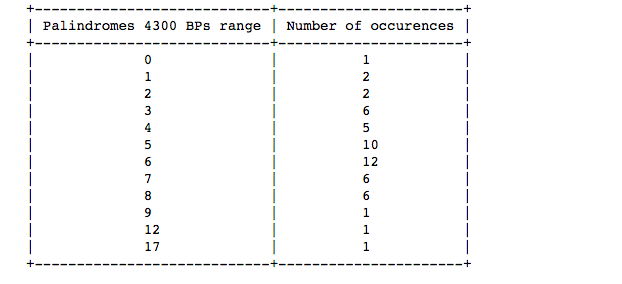
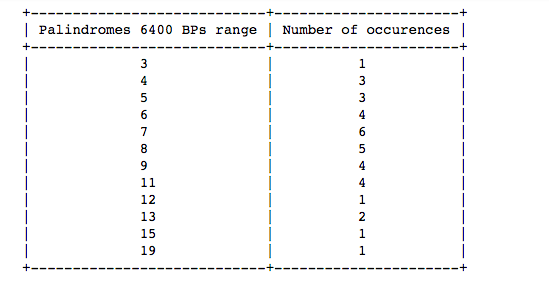
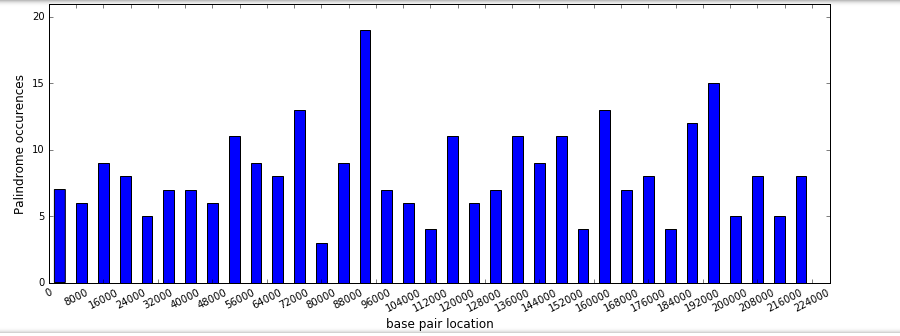
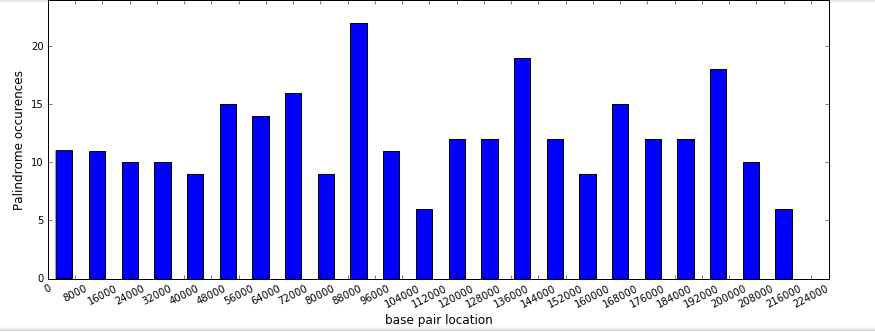


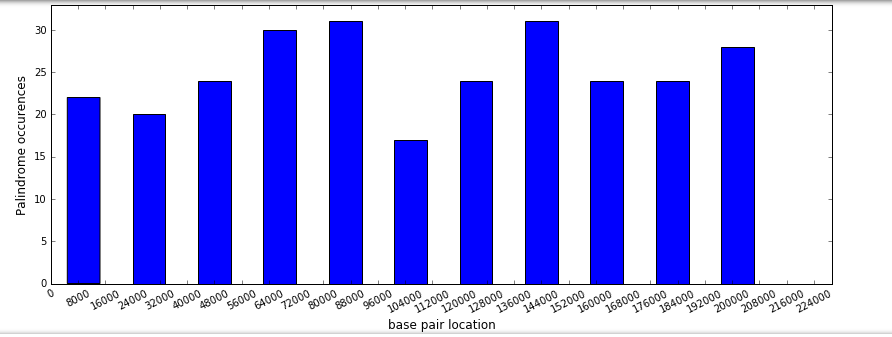
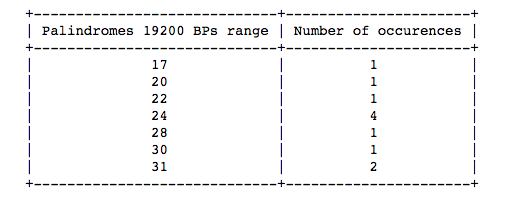
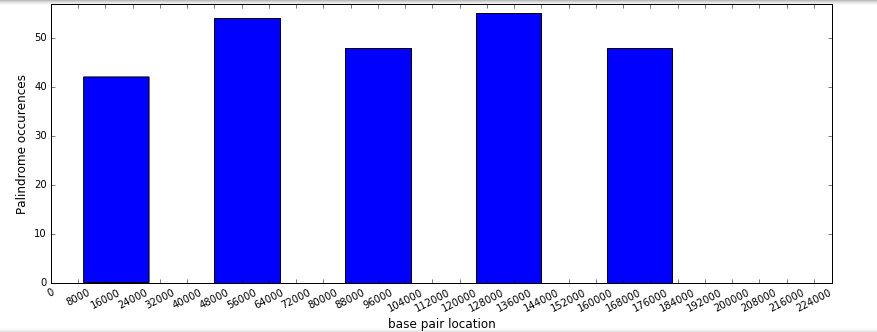
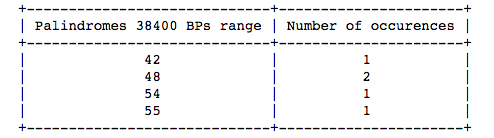












# 2) Random model for the distribution of palindromes. Since the palindrome occurrences share a variety of similarities with the poisson random process, we modeled it using a poission process. The characteristic features of the Poisson Process are as follows:

a) The underlying rate at which points, called hits, occur and is such that it doesn’t change with location (homogeneity).

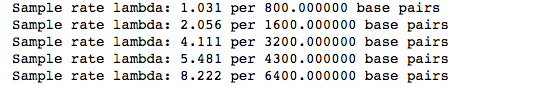
b) The number of points falling in separate regions are independent.

c) It isn’t possible for any 2 points to land in the place

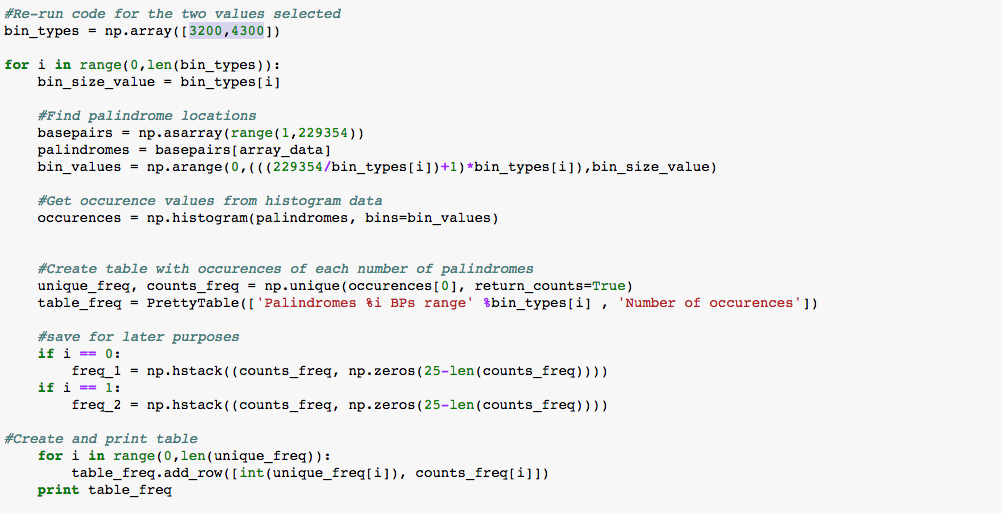
Since the strand of the DNA can be thought of as a line, and the location of a palindrome can be thought of as a point on the line, palindromes are scattered randomly and uniformly across the DNA, the number of palindromes in any small piece of DNA is independent of the number of palindromes in another non overlapping piece, and the chance that one tiny piece of DNA palindrome in it, is the same for all tiny pieces of the DNA.

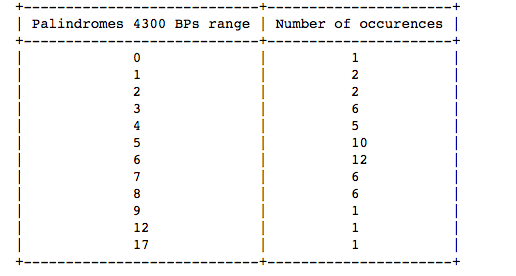
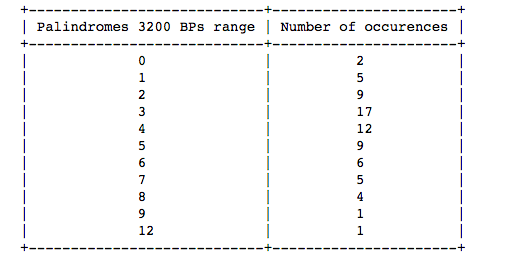
3) Statistics of the Poisson Process. The Method of Moments (MM), as well as the Maximum Likelihood Estimation (MLE) both gave the same result when we attempted to estimate the rate of the parameter, lambda. Its value was estimated in this case by dividing the number of palindrome occurrences by the number of bins.

# 

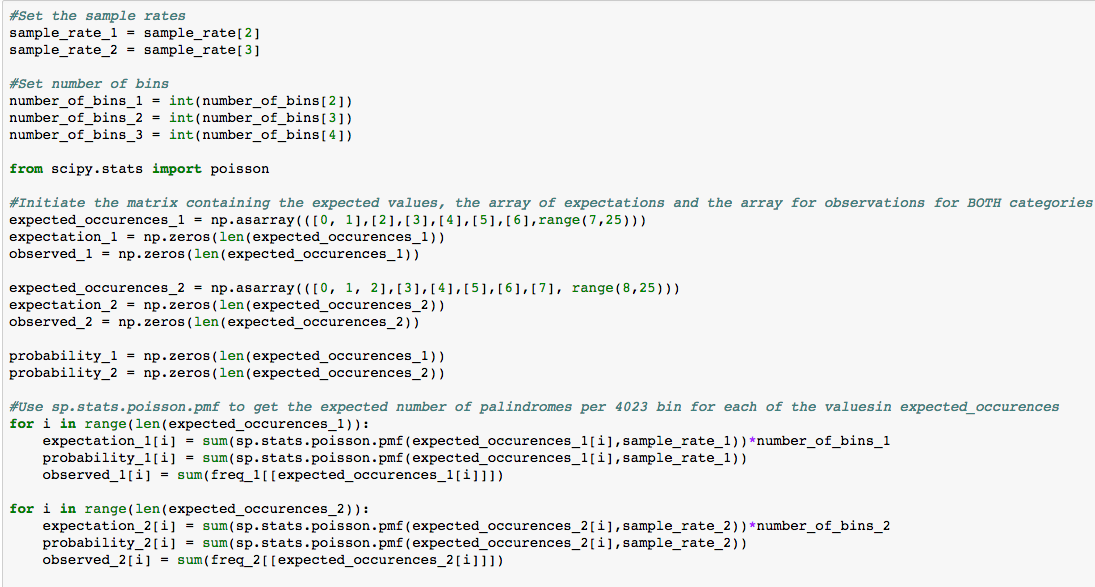


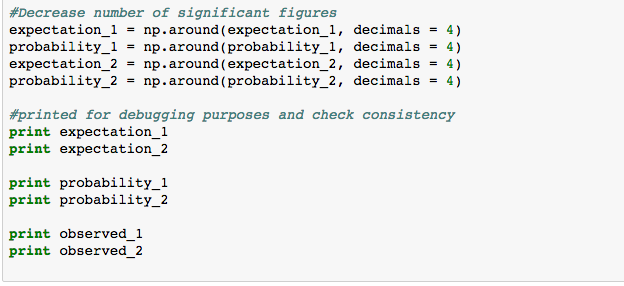
We selected bin sizes 3200 and 4300, and re-ran the code for both of them to generate tables and histograms like we did in appendix 1.

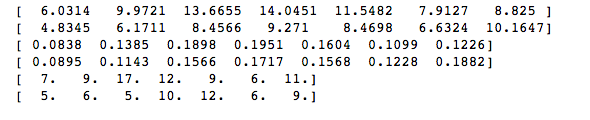




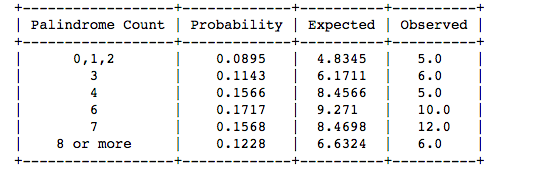
Then, we created a vector containing the expected values, given sample rate and number of observations using the poisson pmf function.

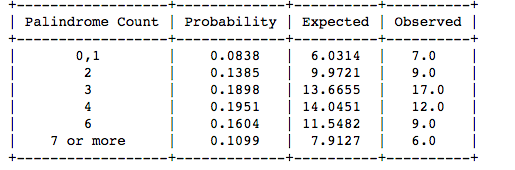






Furthermore, we created a table containing expected and observed number of palindromes for each number of observations.





# 4) The Chi-Square Test. First, we plotted the chi-square distribution for bin size 3200 and calculated the chi-square values.

# 

The test revealed a chi-squared statistic of 2.922, and a p-value of 0.819. Clearly, the

values we found could be related to a poisson random process, but not in a definitive

way. This showed that our data may possess some non-randomness.

# Next, we plotted the chi-square distribution for bin size 4300 and calculated the chi squared values.

# 

The test revealed a a chi-squared statistic of 3.146, and a p-value of 0.790 Again, the

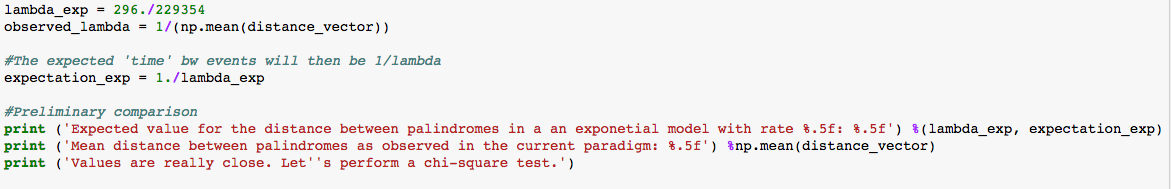
values we found could be related to a poisson random process, but not in a definitive

way. This showed that our data may possess some non-randomness In

fact, a p-value of such size indicates a similarity to the poisson, but not a definitive one.

5) Modelling the exponential distribution. We modeled and compared exponential

distributions. Since we had 296 palindromes scattered in 229354 bps, the rate parameter in this case was estimated by its event rate, or number of palindromes/bp of dna : 296/229354.



Expected value for the distance between palindromes in a an exponetial model with rate0.00129 was **774.84459**

The mean distance between palindromes as observed in the current paradigm was

**772.89189**.

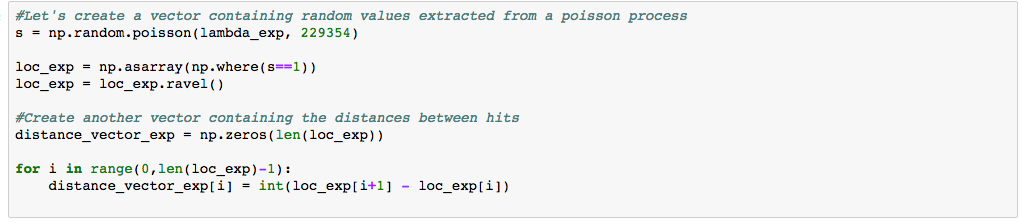
Since the values are really close, we performed a chi-square test.

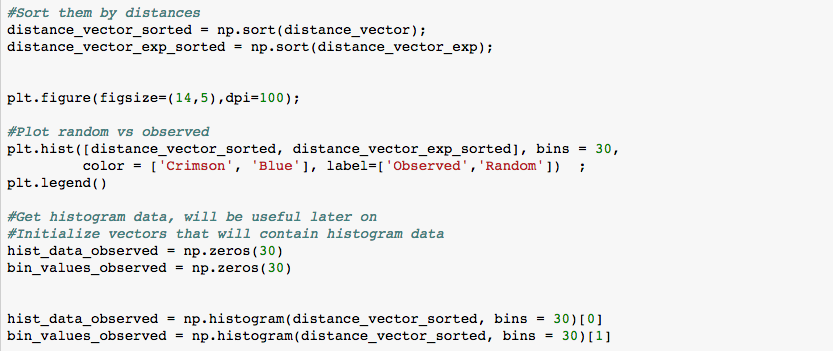
Hypothesis Testing**:** Under the null hypothesis (Ho), we have that the distance between hits is exponentially distributed. We checked data consistency using graphical methods and plotted distances of a random poisson process against our observations.

# Plot of cdfs of functions with different lambdas (observed vs random):

# 

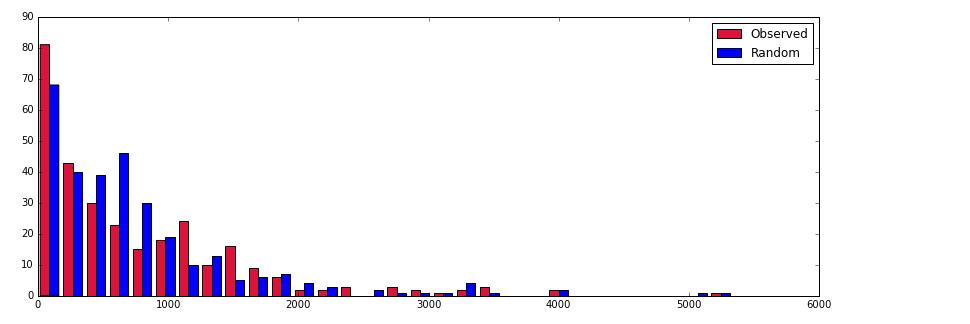
Once again, this was done to check if our data was in accordance with a poisson random process and distance between hits exponentially distributed.



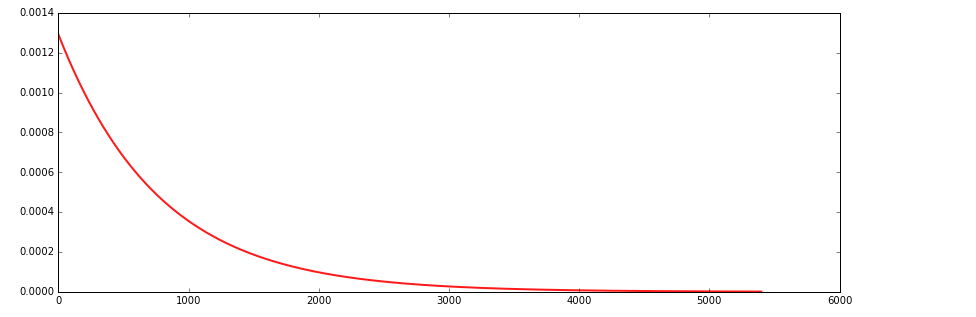
****

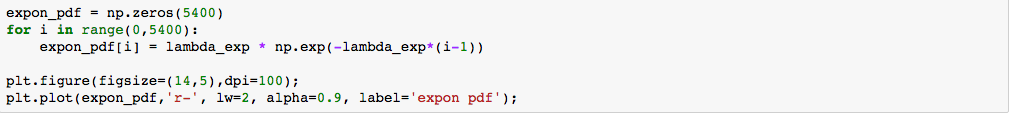
As a preliminary analysis, we can see that the number of palindromes occurring close toeach another (1st bin) is higher than on our observation set than on a random scatter.

Hence we investigated further.

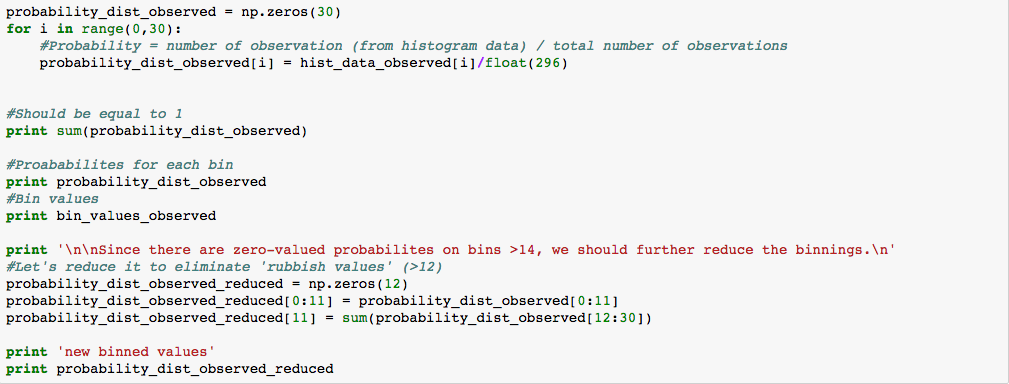


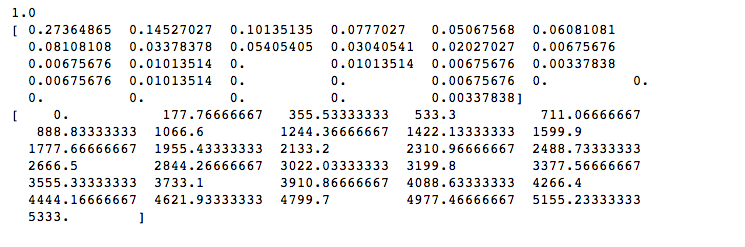
We ploted for an exponential random variable having our designated lambda. We then put it in a bin and compared it to the observed distribution. (Putting it in a bin was required as at the tail of the distributions of the occurrences were extremely low as we can see on the previous histogram).





Furthermore, we created a vector containing the probabilities for the number of observation falling into 30 (an arbitrary number) different bins. We compared this using a chi-squared test to the same probabilities of an exponential function having the expected lambda.





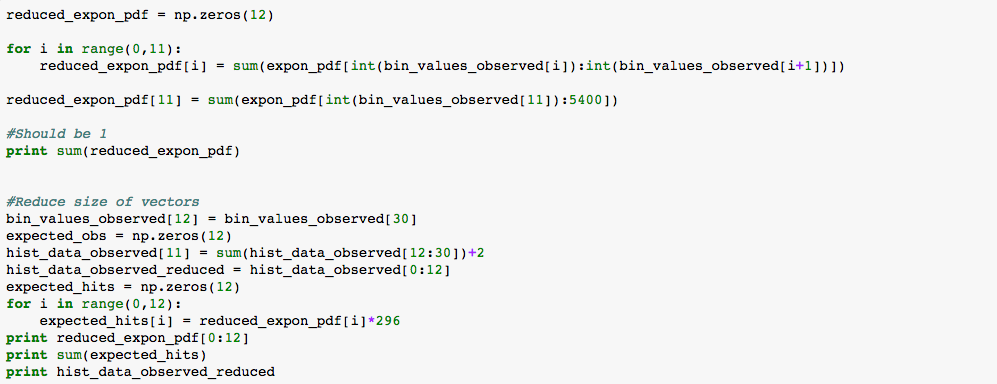
Since there are zero-valued probabilities on bins >14, we should further reduce the

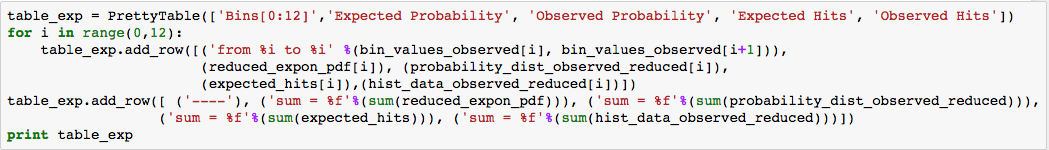
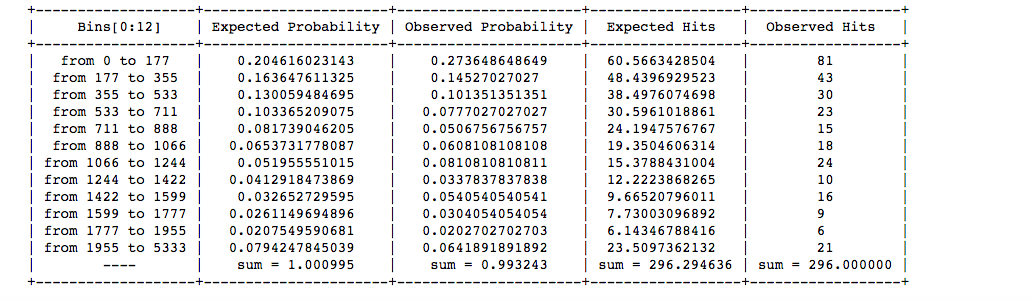
Bin sizes.



We then reduce the theoretical exponential distibution by values according to the bins determined previously (30 bins, variable stored as bin\_values\_observed)

The number of bins was standardized to 12, to make it look more consistent and reasonable.





6) The Actual Chi Squared Test. The chi-square test for the number of hits of an

exponential distribution revealed that the data wasn't actually exponentially distributed. Therefore, we rejected the null hypothesis.

Values: chi-square test statistic = 24.723683, p-value: 0.010004

The data is not exponentially distributed. We felt that was enough to justify the use of

other methods for analysis to confirm that cluster indeed exists.

What was of particular interest was the fact that the biggest contributor to the test

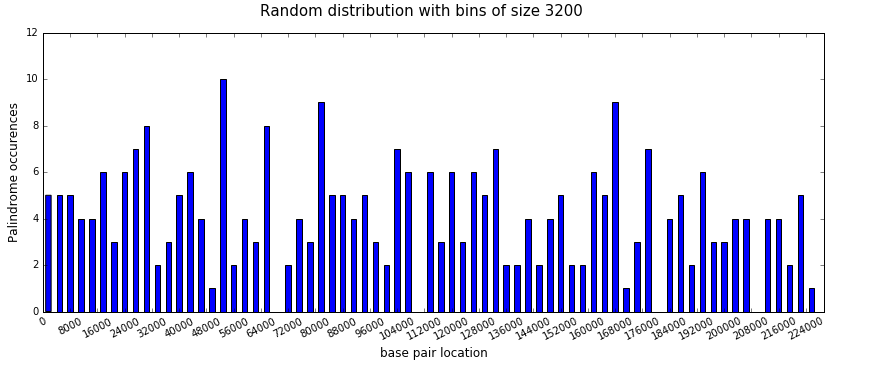
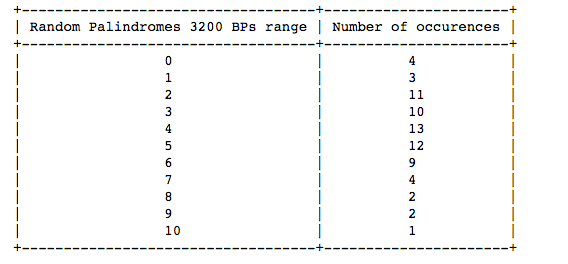
statistic was the difference in number of hits observed for small distances

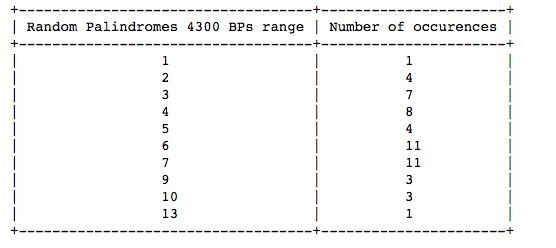
(from 0 to 177). This was an obvious indicator that clusters are present.

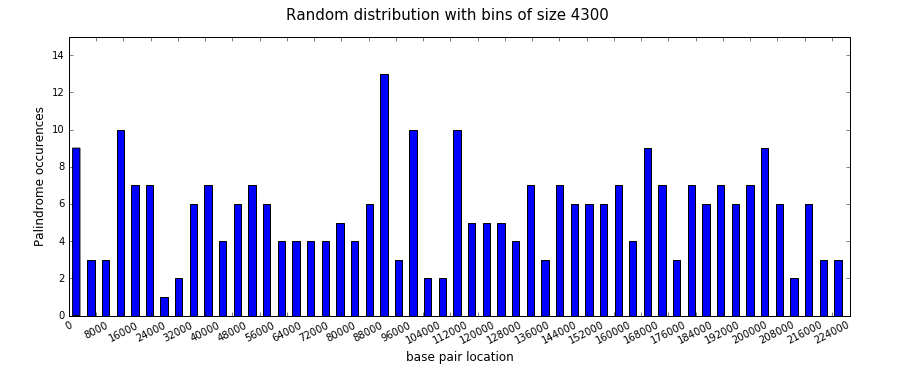
We created a random distribution of palindromes that follow the Poisson distribution and proceeded by creating a distribution that resembles the one given by the data and assessed how closely they could be related. i.e, we created two distributions and categorized them using bin sizes of 3200 and 4300.

First we will randomly selected 262 random numbers in the range 90, 229 354 that represented a random distribution of palindromes.



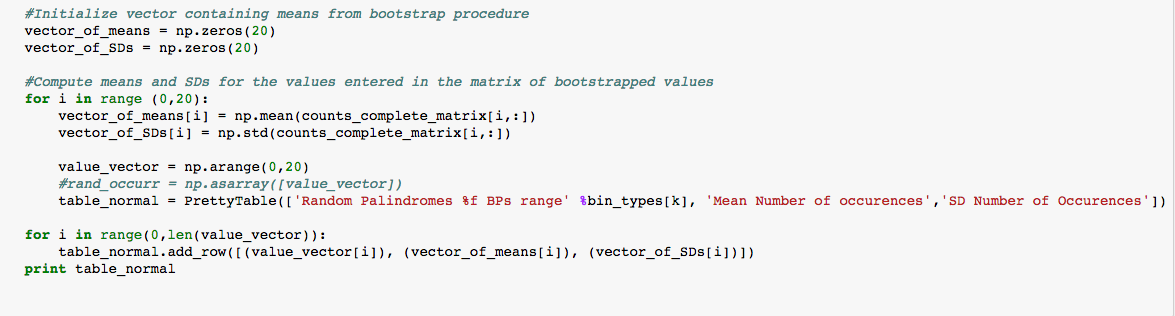


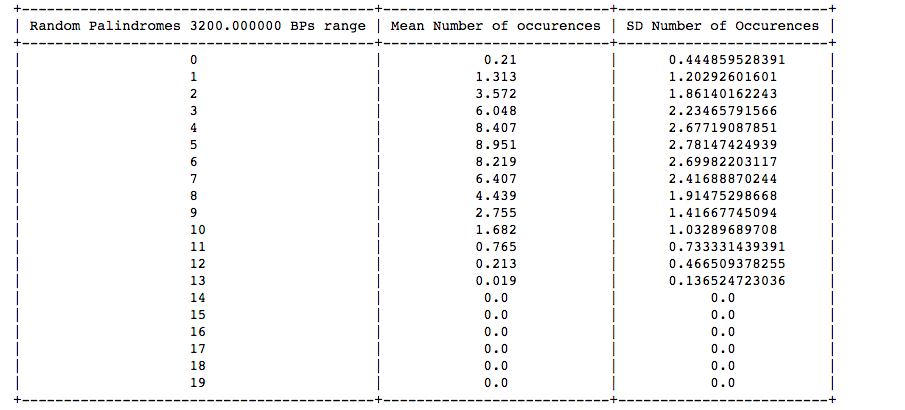
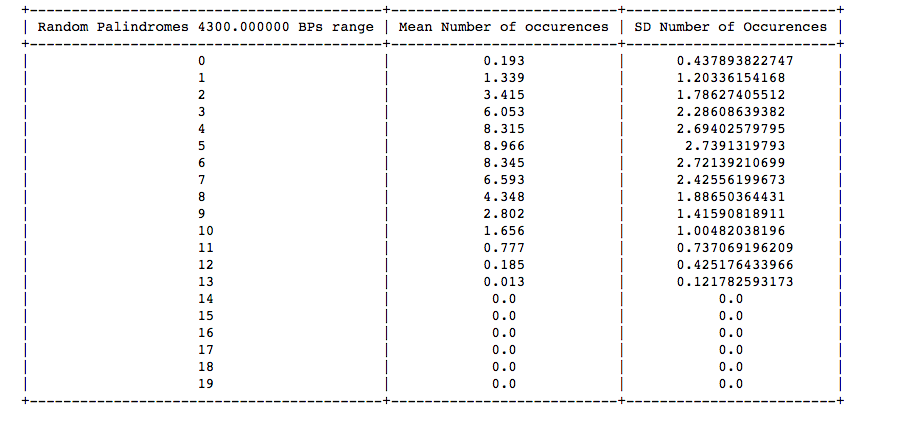




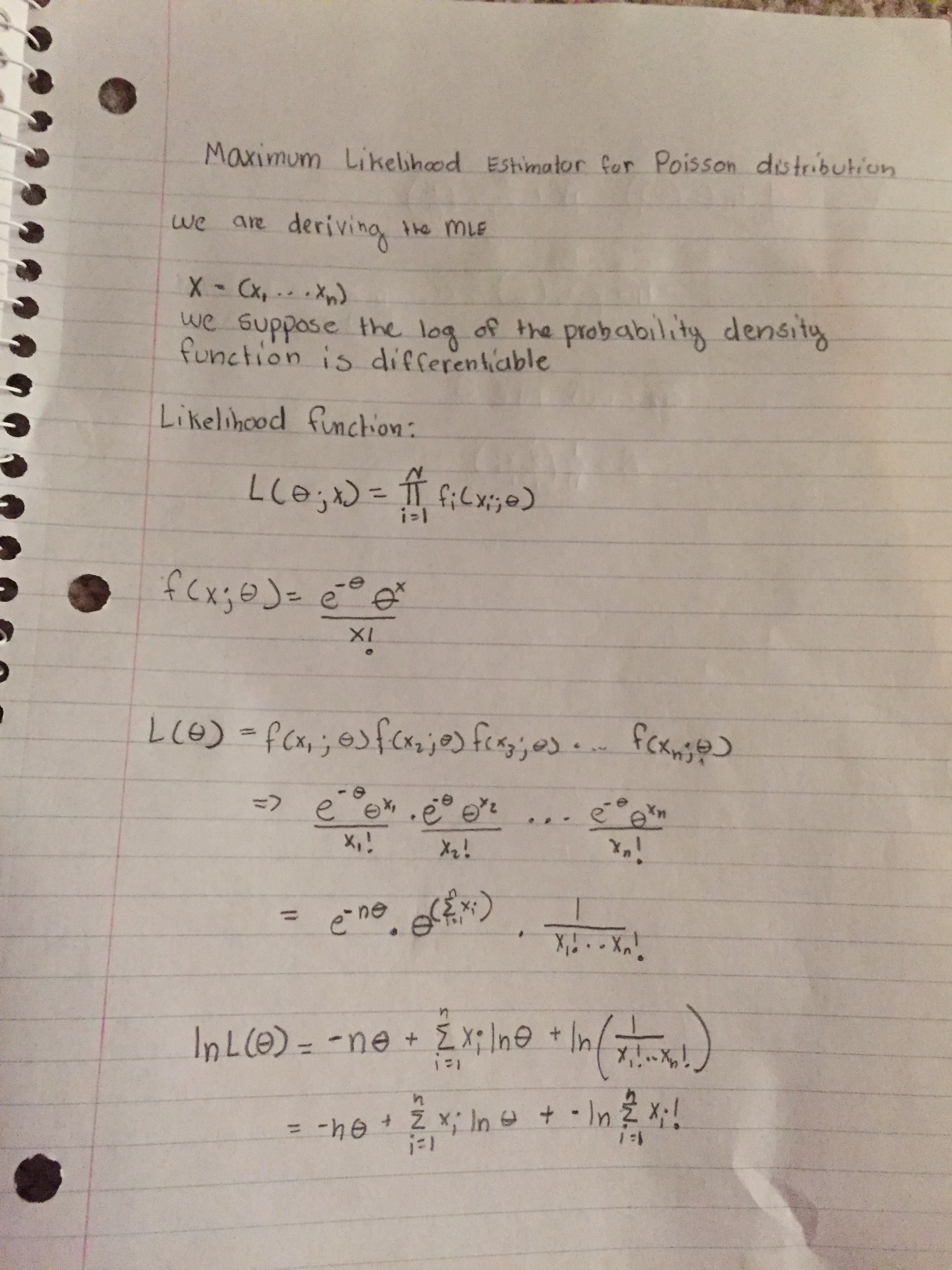
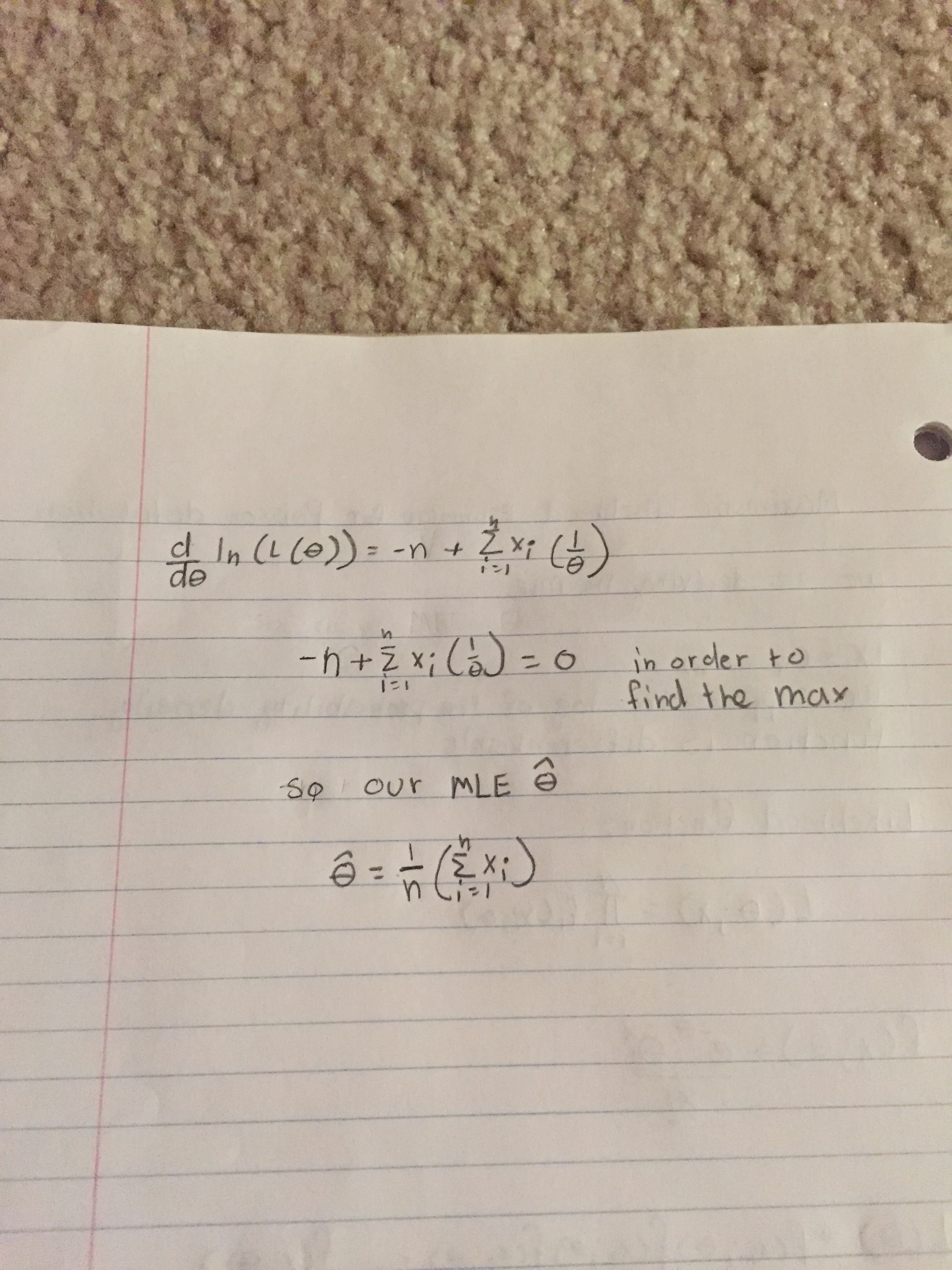
We performed a bootstrap procedure, with 1000 iterations, to find the expected number of palindrome occurrences in both the 3200 and 4300 base pair divisions







7) Proof that our Maximum Likelihood Estimator is unbiased.



Works Cited

Masse, Marie. "Human Cytomegalovirus Origin of DNA Replication (oriLyt) Resides within a Highly Complex Repetitive Region." *Proceedings of the National Academy of Sciences of the United States of America*. Web.