**Case Study 3: Patterns in DNA**

Yu-Chun Chen, A13356506
Yanyu Tao, A13961185
Bolin Yang, A92111272
Shuibenyang Yuan, A14031016
Haoming Zhang, A14012520
Mingxuan Zhang, A13796895

## Introduction:

The human cytomegalovirus, or CMV, is a potentially life-threatening disease for people with suppressed or deficient immune system. In order to develop strategies for combating the virus, scientists study the way in which the virus replicates. In particular, they are in search of a special place on the virus' DNA that contains instructions for its reproduction: origin of replication. A virus' DNA contains all of the information necessary for it to grow, survive, and replicate. DNA can be thought of as a long and coded message made from four-letter alphabet: A, C, G, or T. Its sequences contain many patterns, and have the potential to flag important sites on the DNA, such as the origin of replication. Complementary palindrome is one type of the patter. In DNA, the letter A is complementary to T, and G is complementary to C. Complementary palindromes are sequences of letter that read in reverse as the complement of the forward sequence, such as GGGCATGCCC.

The origin of replication for two viruses from the same family as CMV, the herpes family, are marked by complimentary palindromes. One of them, Herpes simplex, is marked by a long palindrome of 144 letters. The other, the Epstein-Barr virus, has several short palindromes and close repeats clustered at the origin of replication. For the CMV, the longest palindrome is 18 basepairs, and altogether, contains 296 palindromes between 10 and 18 base pairs long. Biologist conjectured that clusters of palindromes in CMV may serve the same role as the single long palindrome in Herpes simplex, or the cluster of palindromes and short repeats in the Epstein-Barr virus' DNA. To find the origin of replication, DNA is cut into segments and each segment is tested to determine whether it can replicate. If it does not replicate, then the origin of replication must not be contained in the segment.

**In this case study, we are going to search for unusual clusters of complementary palindromes**. This is because the process of testing and determining if every segment of DNA can replicate is time consuming and expensive without clue on where to begin the search. Thus, a statistical investigation of the DNA to identify unusually dense clusters of palindromes can help narrow the search and potentially reduce the amount of testing needed to find the origin of replication.

Before we go into our own investigation, some literatures are reviewed. David G. Anders and Suzanne M. Punturieri introduce that the global virologists are investigating the origin of CMV replication to find the most effective antiviral strategy killing CMV since CMV is a significant cause of morbidity and mortality after the infection. Because CMV has two categories, Herpes simplex and Epstein-Barr virus, the origin of these two viruses varies in their structure and function. Herpe simplex's origin consists of single symmetry sequence while Epstein-Barr virus's origin formed by the multiple elements (David G. Anders and Suzanne M. Punturieri). "Human Cytomegalovirus Origin of DNA Replication (oriLyt) Resides within A Highly Complex

Repetitive Region" futher investigates the condition of origin. oriLyt, the minimal replication origin, is living in the region between 92,210 and 93,715 bp on the viral genome. Human CMV oriLyt exhibits an elaborate and repeated sequence structure, dividing the region into two domains for the important function (Marie J. Masse, Karlin, Samuel, Gabriel A. Schachtel, and Edward S. Mocarski). Two origins are marked by palindromes. D. Nolan & T. P. Speed introduce that the characteristics of palindromes. They found 296 palindromes that were at least 10 letters long and the longest one has 18 letters. Palindromes with shorter than 10 letters occur too frequently by chance, which is hard to detect. Virologists speculate that the clusters of palindromes in CMV have the same function with the single long palindrome in Herpes simplex, or the cluster of palindromes and short repeats in the Epstein-Barr virus' DNA (Bradic Jelena).

## Data:

The DNA sequence of CMV, which is 229, 354 letters long, was published in 1990 (Chee et al.). Leung et al. implemented search algorithms to screen the sequence for many types of pattern. Altogether, 296 palindromes were found that were at least 10 letters long, and palindromes shorter than 10 letters were ignored. The longest ones found were 18 letters long and occurred in locations 14719, 75812, 90763, and 173893 along the sequence.

We begin to group the data of the 296 palindromes found by segmenting the DNA chain into intervals of base pairs and count the number of palindromes found in each interval. The distribution is shown below in *Figure 1*. It is easy it is fairly easy to see that there appears to be clusters of palindromes in at least two locations: around the 93,000th and 195,000th pairs of DNA. This is enough to formulate a hypothesis which claims that the clusters at these two locations are exceptions within the typical structure of the DNA chain, i.e. that the clusters are not due to chance. In the investigation section, we will compare histograms of actual palindromes to histograms to simulated distribution to look for pattern of clusters.

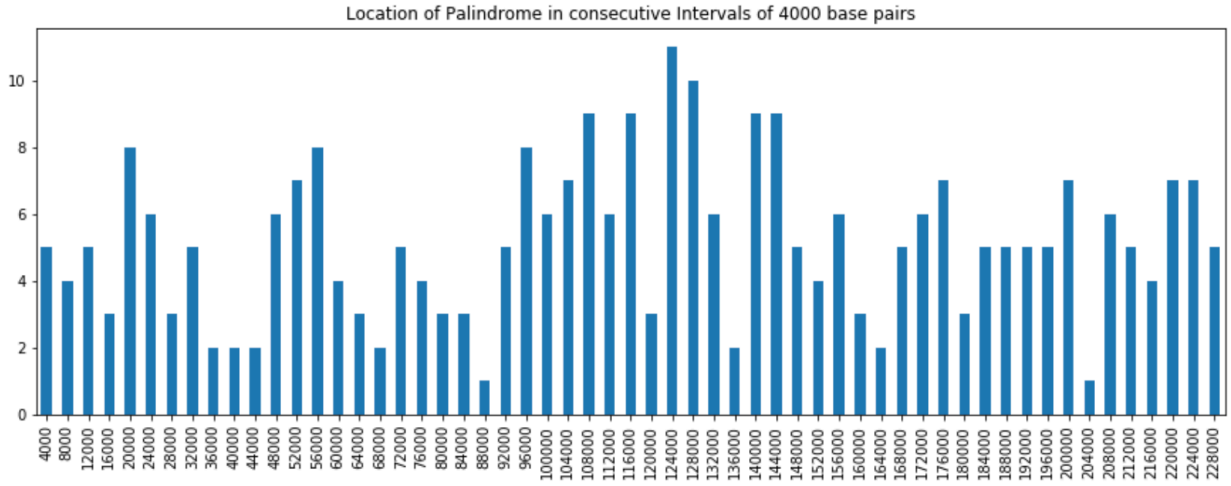



*Figure 1. Location of Palindrome in Consecutive Intervals of 4000 Base Pairs*

By computing number of palindromes in each interval, we can also generate plot as shown below in *Figure 2*. We can see that the observed palindromes present higher spikes of number of palindromes per intervals. There seems to be one or two outliers of intervals containing a higher number of palindromes.

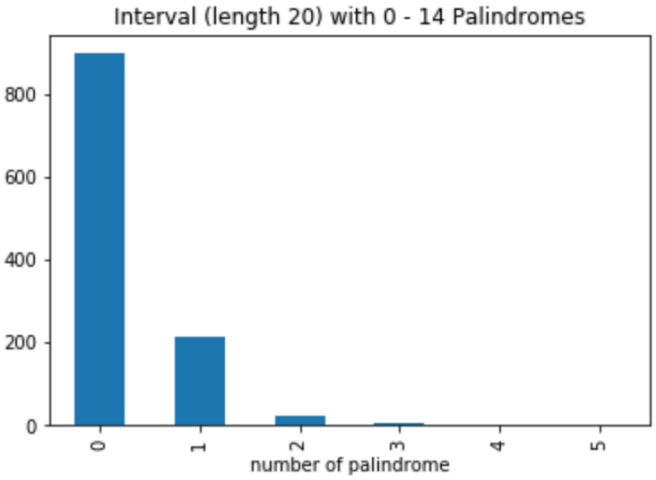



*Figure 2. Interval (length 20) with 0-14 Palindromes*

We will also generate the plot to look at the spacing between palindromes for investigation. The plot is shown below in *Figure 3*. From the figure, it does not seem to show any apparent useful pattern.

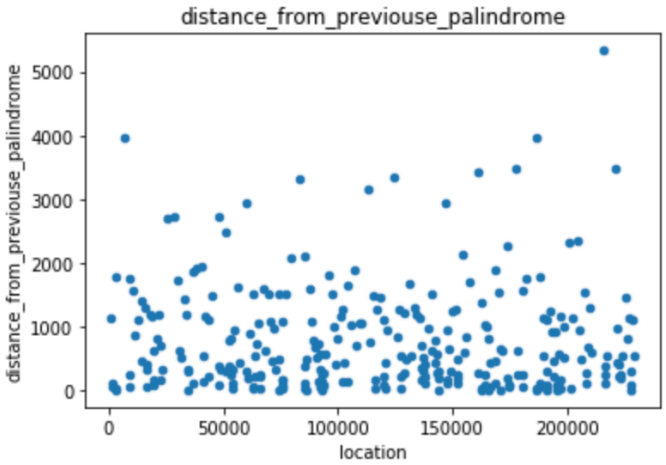



*Figure 3. Distance From Previous Palindrome*

## Background:

Deoxyribonucleic acid (DNA) is the basis to carry the genetic instructions for human and other living organisms' growth, development, and important functioning. In 1944 Avery, MacLeod and McCarty discovered that DNA carries the hereditary information (Bradic Jelena). Then in 1953, Franklin, Watson and Crick found that DNA is a double helical structure with two long chains of nucleotides. One single nucleotide has three parts: a sugar, a phosphate and a base. Four types of adenine (A), cytosine (C), guanine(G), and thymine (T) form the basis for the information storage. The basis can vary from one nucleotide to another, giving the long, coded, different message. Two strange of the nucleotides connecting at the bases create the complementary pairs. And there are four pairs: A to T, C to G, G to C and T to A. The CMV DNA molecule contains

229,354 complementary pairs of letters or base pairs, while human DNA has more than 3 billion base pairs.

Viruses have the simple structure with two main parts: a DNA molecule and a protein shell called a capsid. The viruses' DNA, storing all the information for controlling life process, typically ranges up to several hundred thousand base pairs in length. For example, when a "snipping" enzyme cuts the DNA strand apart at a small region called origin, it begins the E coli. replication. The free nucleotide only sticks the complementary base on DNA and bounces away from the "wrong" base. More nucleotides are added when the snipping enzyme opens the DNA further. and then a clipping enzyme puts them together.

CMV is a member of the herpes virus family, which threatens the people's immune system. It is a common virus for children that 10%-15% of children are infected with CMV before the age of 5 typically. This infection will level off in the young adulthood. CMV will be dominant after the infection but it is not life threatening. It only becomes harmful when the virus enters a productive and quickly replicates tens of thousands of copies. In this cycle it poses a major risk for people in immune-depressed states: transplant patients, AIDS patients, etc. Therefore, virologist can find an effective vaccine against the virus after locating CMV's origin of the replication.

# Investigation:

### Scenario 1: Random scatter

To begin, we need to find clusters of palindromes in our dataset and also determine whether these cluster occurred by chance or not. If it is not occurred by chance, it then can be further used to analyze if this is a potential replication site. We start our investigation by analyzing the structure of the data by departures from a uniform scatter of palindromes across the DNA. Of course, a random uniform scatter, does not mean that the palindromes will be equally spaced as milestones on a freeway. There will be some gaps on the DNA where no palindromes occur, and there will be some clumping together of palindromes. In order to visualize palindromes clusters in our data, we use Monte Carlo to simulate random scatters from the uniform distribution. We randomly gather 296 palindromes from a DNA sequence of 229,354 base pairs for three time. And then, we use different graph and method to compare the distribution and spacing for origin data and our three new generated sample.

First of all, we compared the location of where the palindromes occur. From the dot-plot graphs shown in *Figure 4* (density of the palindromes vs the location in the sequences), we cannot tell a difference between the original data and the data we generated. However, we do observe that the spacing of the palindromes are not equally appeared. There is some pattern, which we will discuss in the latter scenario.

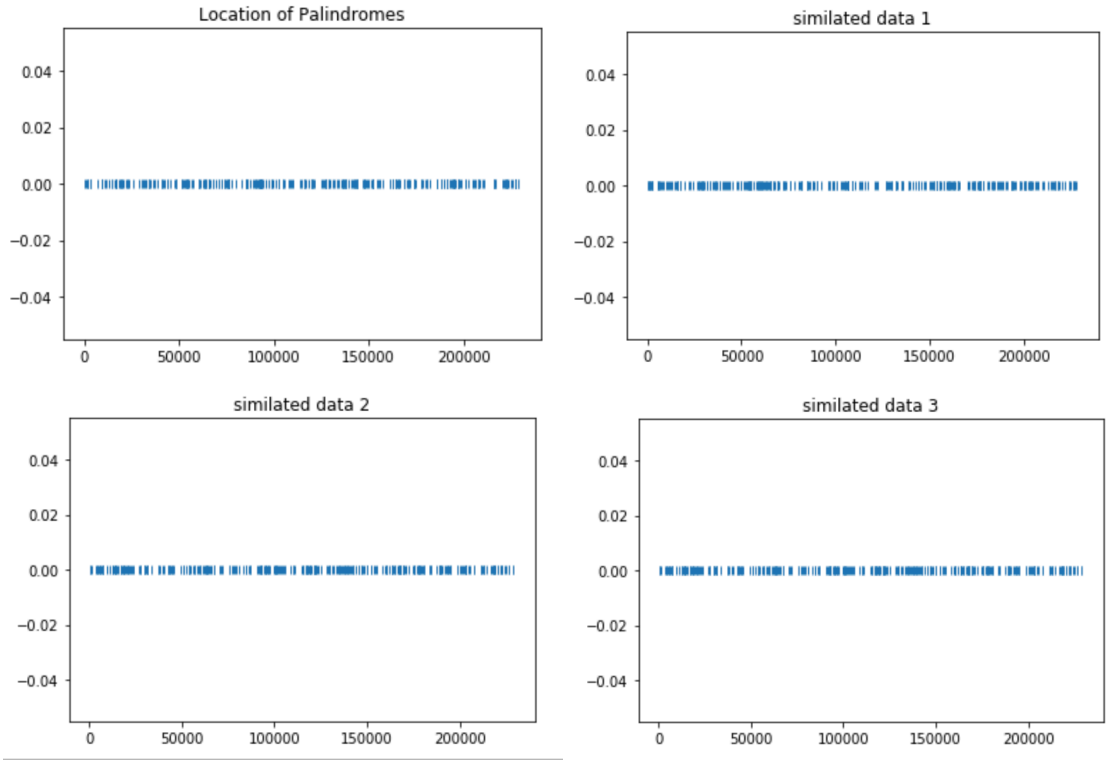
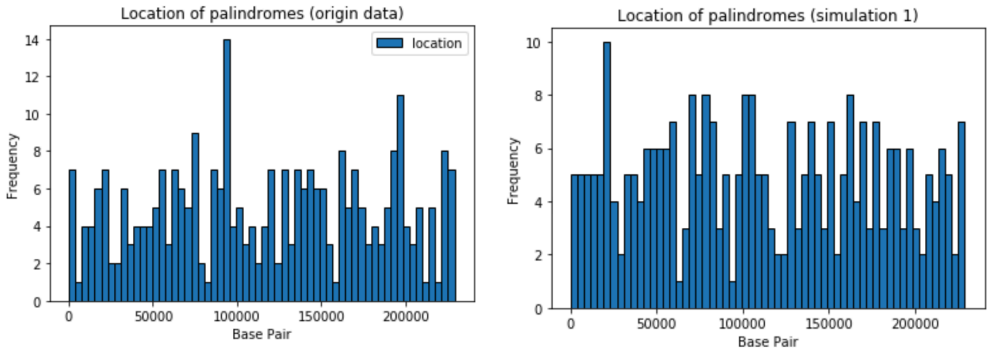



*Figure 4. Dot Plot of Palindrome Locations.*

We want to visualize the cluster instead of counting the density, thus we plot a histogram of the palindromes for the original data and the simulated data. (we choose 4000 location for each bin). From the comparison of the *Figure 5*, if it is a random generated sample, there is no pattern in the clusters. However, for the original data, we can clearly observe that there is a cluster around 90000$^{th}$ base pair with more than 15 frequency in the data unlike other clusters in the sample. Therefore, we confirm that there is an unusual cluster that depart from a random scatter that we need to pay attention to in order to determinate the "the origin of replication"



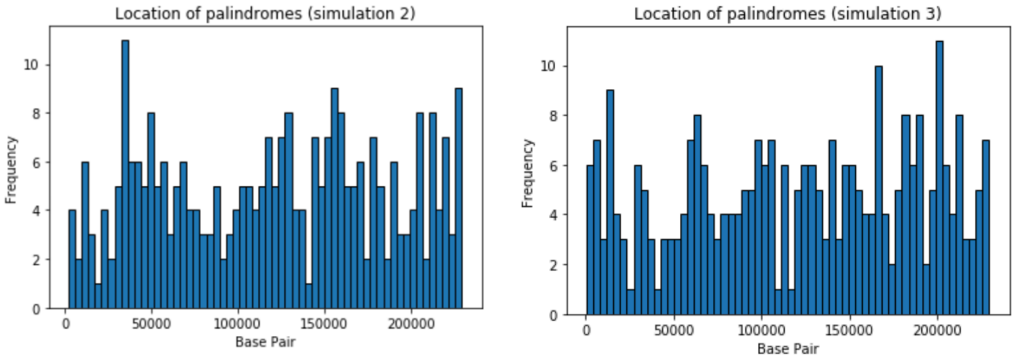



*Figure 5. Histogram of Palindrome Locations*

Next, we compare the spacing between each of the palindroms. We have done a scatterplot of the spacing of the palindroms. Similar to the dotplot we did before, it is difficult to see any pattern in the spacing between the orginal data set and the simulated data. Therefore, we will furthur discusse the spacing between palidromes in the latter scinearios.

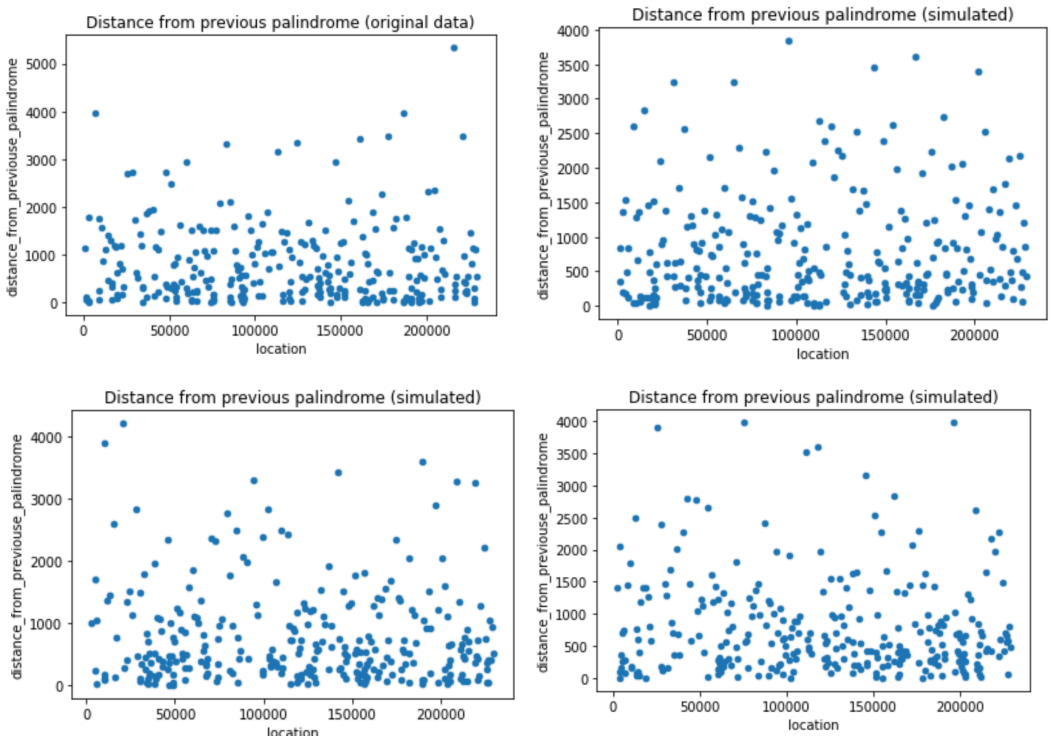



*Figure 6. Scatterplot of Spacing Between Consecutive Palindromes*

Finally, we want to see for each non-overlapping region of the DNA sequence, how many palindromes can there be. We made a bar plot on the count of each interval (2000) for the original data and the simulated data. In the *Figure 7*, we observe that the frequency of counts of the palindromes in each intervals is skewed to the right for all four samples, but the original data has a outlier way furhter than the normal counts. (12 palindromes in a interval) Thus, we suspect that there could be some unusual cluster in the DNA sequence needed to be investiage further to be

sure of. It could be just the count of the palidromes in the 90000th base pair that we found using histogram above.

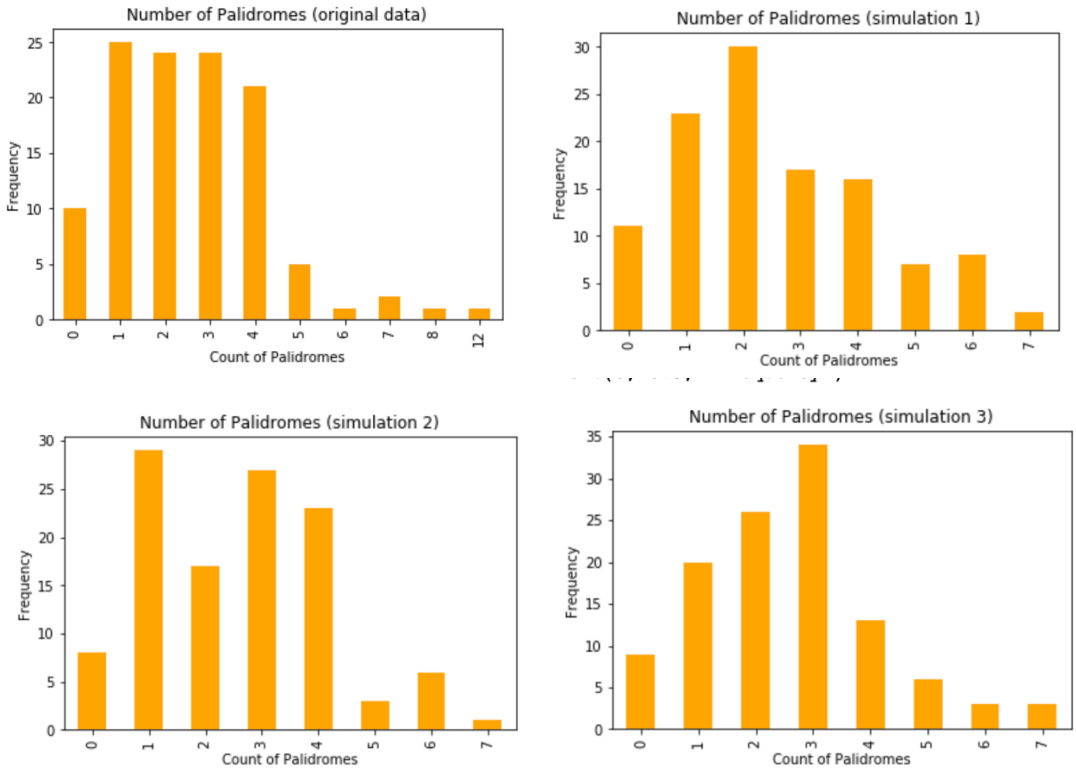



*Figure 7. Histogram of Number of Palindromes in Non-Overlapping Region*

**Scenario 2: Locations and Spacing**

We first divide our data into 30, 40, and 50 intervals and compare locations of palindromes of our data to simulated uniform distributions by poisson process. We will be using graphical comparison, Chi-Square goodness of fit test, and residual plots.

From *Figure 8* to *Figure 10* below, we can see that distributions of actual data deviate from simulated uniform data by Poisson process. Moreover, at location close to 100000, all three distributions of orginal dataset show an especially high frequecy. Thus, it is reasonable for us to consider unusual cluster of palindromes around this location close to 100000.

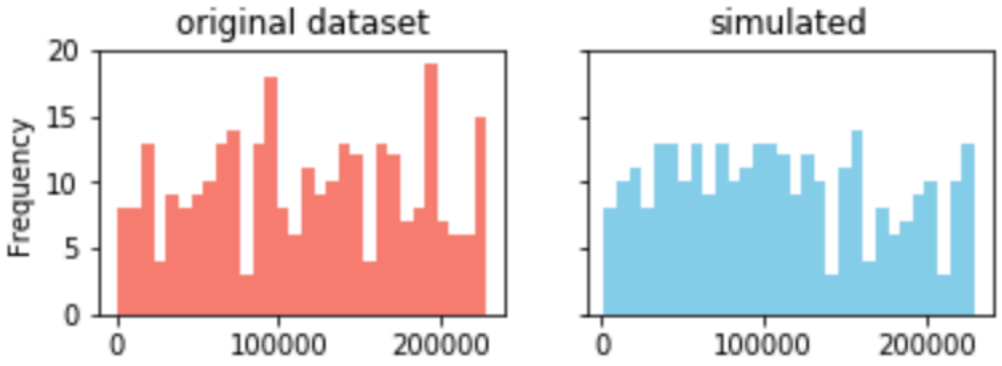



*Figure 8. Distribution of palindromes in data v.s. simulated uniform data for 30 intervals*

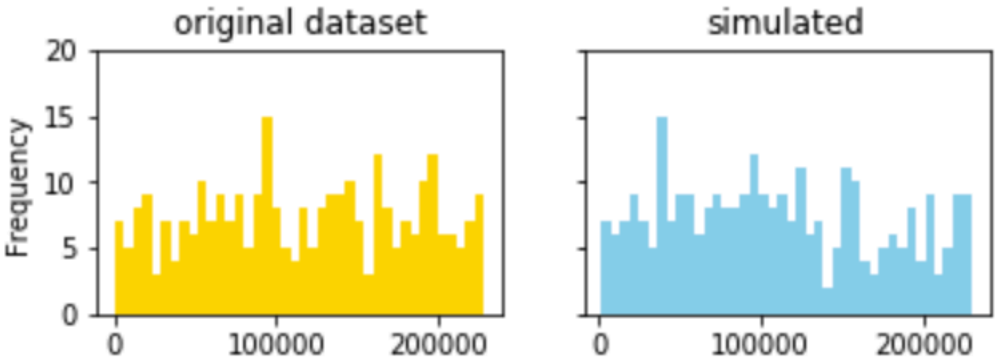



*Figure 9. Distribution of palindromes in data v.s. simulated uniform data for 40 intervals*

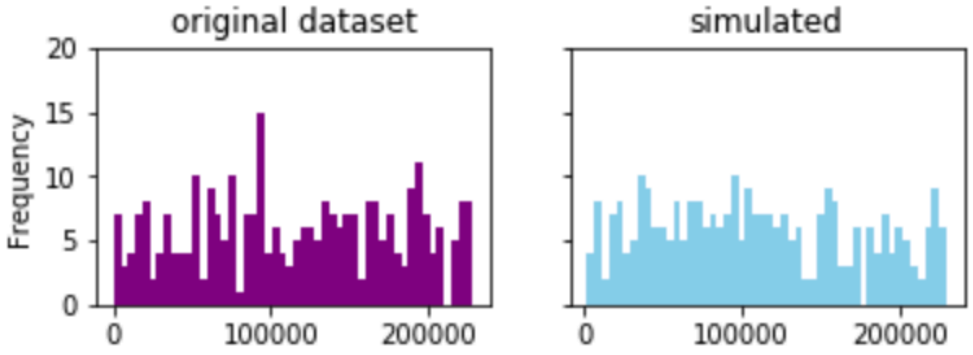



*Figure 10. Distribution of palindromes in data v.s. simulated uniform data for 50 intervals*

We then uses Chi-Square fitness test to see if distributions of actual data fit the simulated uniform data. The test statistics and p-values are shown in *Table 1* below. The null hypothesis is that the actual distribution fits the simulated uniform data by Poison process, and the alternative vice versa. By observing p-value for all different number of intervals, they are all extremely small,

and we reject the null hypothesis. Thus, we conclude that the distribution of location of palindromes does not fit the simulated uniform distribution by Poisson process with any of the number of interval. As a result, we doubt if the palindromes are located randomly.

| Number of Interval | 30 | 40 | 50 |
|---|---|---|---|
| **Test Statistic** | 103.73 | 125.00 | 175.77 |
| **P-Value** | 2.45e-10 | 6.11e-11 | 1.82e-16 |

*Table 1. Results of Chi-Square Goodness of Fit Test of Actual Distribution and Simulated Uniform Distribution by Different Number of Intervals*

Since the p-values are small according to *Table 1* above, we doubt the fit of distribution, and residual plots can help determine where the lack of fit occurs. From *Figure 11* shown below, we can see that for all 30, 40, and 50 number of intervals, the residuals exceed 3, which indicates lack of fit of actual distribution and simulated data.

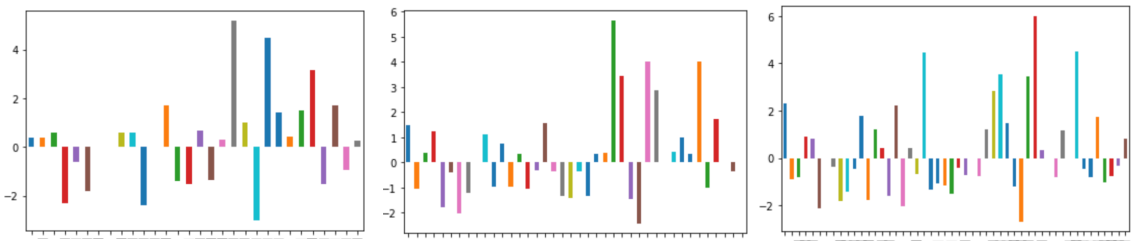


*Figure 11. Residual Plots of Actual Distributions and Simulated Uniform Data by Different Number of Intervals*

We then start to investigate the spacing between consecutive palindromes, pairs of palindromes, and triplets of palindromes. We will do this by graphical comparison and Chi-Square Goodness of Fit test. For graphical comparison, we will compare spacing between consecutive palindromes to exponential distribution, as spacing between uniform data by Poisson process is exponential. Then, we will compare spacing between pairs to Gamma distribution with parameter 2, and spacing between triplets to Gamma distribution with parameter 3.

From *Figure 12* to *Figure 14* shown below, we can see that the actual distributions of spacing between consecutive palindromes, pairs of palindromes, and triplets of palindromes all deviate from exponential distribution, Gamma 2 distribution, and Gamma 3 distribution respectively.

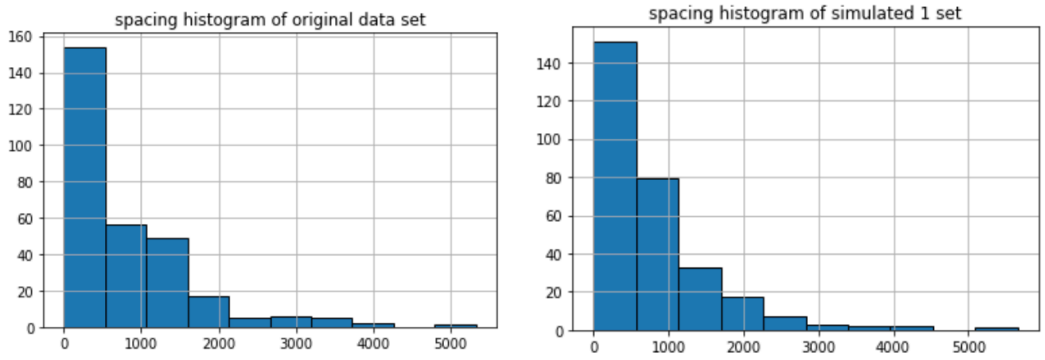



*Figure 12. Distribution of Spacing Between Consecutive Palindromes in Actual Data v.s. Simulated Uniform data*

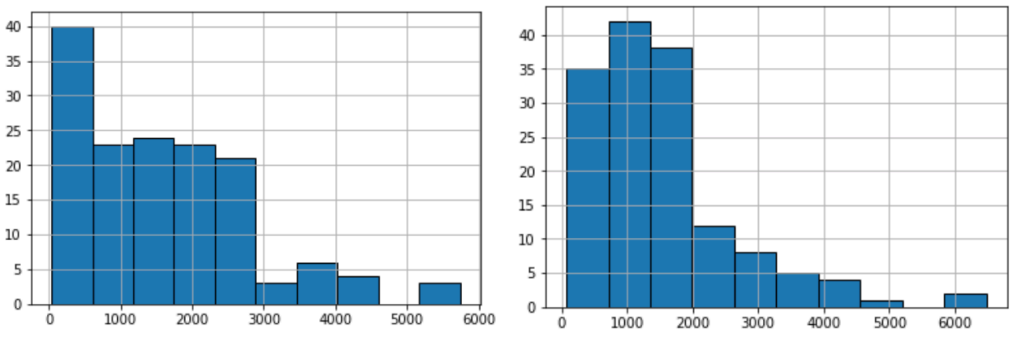


*Figure 13. Distribution of Spacing Between Pairs of Palindromes in Actual Data v.s. Simulated Uniform data*

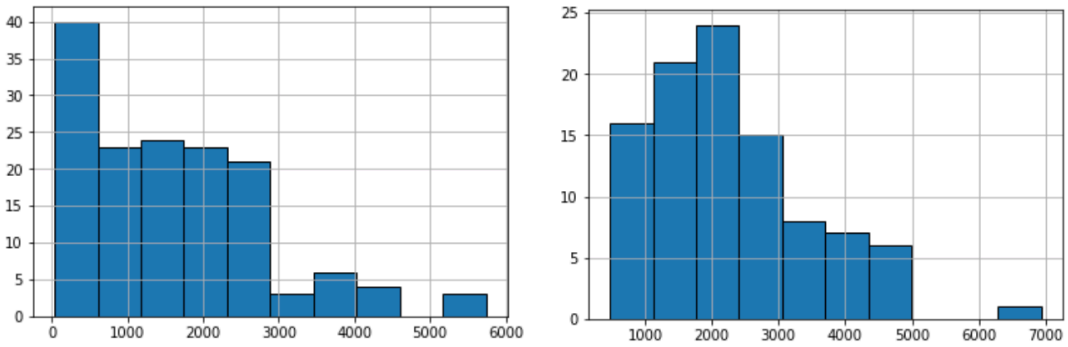


*Figure 14. Distribution of Spacing Between Triplets of Palindromes in Actual Data v.s. Simulated Uniform data*

We then confirm this lack of fit of actual distributions of spacing and exponential, Gamma 2, and Gamma 3 distribution by Chi-Square Goodness of Fit test. Before testing, we smooth the data of spacing by applying linear transformation of division by 10. The test statistics and p-values are shown below in *Table 2*. The null hypothesis is that both distributions are similar, and the distribution of spacing is no different than spacing of simulated uniform data by Poison process. With extremely small p-values shown below, we reject null hypothesis and suspect that our data includes unusual cluster of palindrome.

| Spacing | Singles | Pairs | Triplets |
|---|---|---|---|
| **Test Statistic** | 384.24 | 401.44 | 149.94 |
| **P-Value** | 0.00031 | 1.05e-25 | 0.00046 |

*Table 2. Results of Chi-Square Goodness of Fit Test of Actual Distribution of Spacing and Exponential, Gamma 2, and Gamma 3 Distributions*

**Scenario 3: Counts**

In this scenario, we will examine the counts of palindromes in various regions of the DNA and determine whether Poisson distribution fits our data using graphical methods, goodness of fit test, and standardized residual plot.

Original and simulated data are segmented into 2000, 3000, 4000 and 5000 non-overlapping regions in the four plots below respectively. We display the distribution of the counts of palindrome with respect to the non-overlapping intervals. In each plot, we compare the distribution of counts in the original data with that in our simulated data. All blue bars represent the original data while all orange bars represent our simulated data.

Through visualization, we can see that even though the shape of the distribution of our simulated data roughly align with the shape of the distribution of the original data, the frequency of counts varies. The simulated data has a high concentration at relatively small counts around 4-6 while the original distribution is rather smooth compare to the simulated. More importantly, in all four plots, the original data has tail to the right of the distribution, whereas there is no significant outlier in the simulated distribution. Overall, the original data deviates from random scatter. Since we cannot come to a conclusion whether the counts fit Poisson distribution just by graphical visualization, we continue exploring using chi-square goodness of fit test.

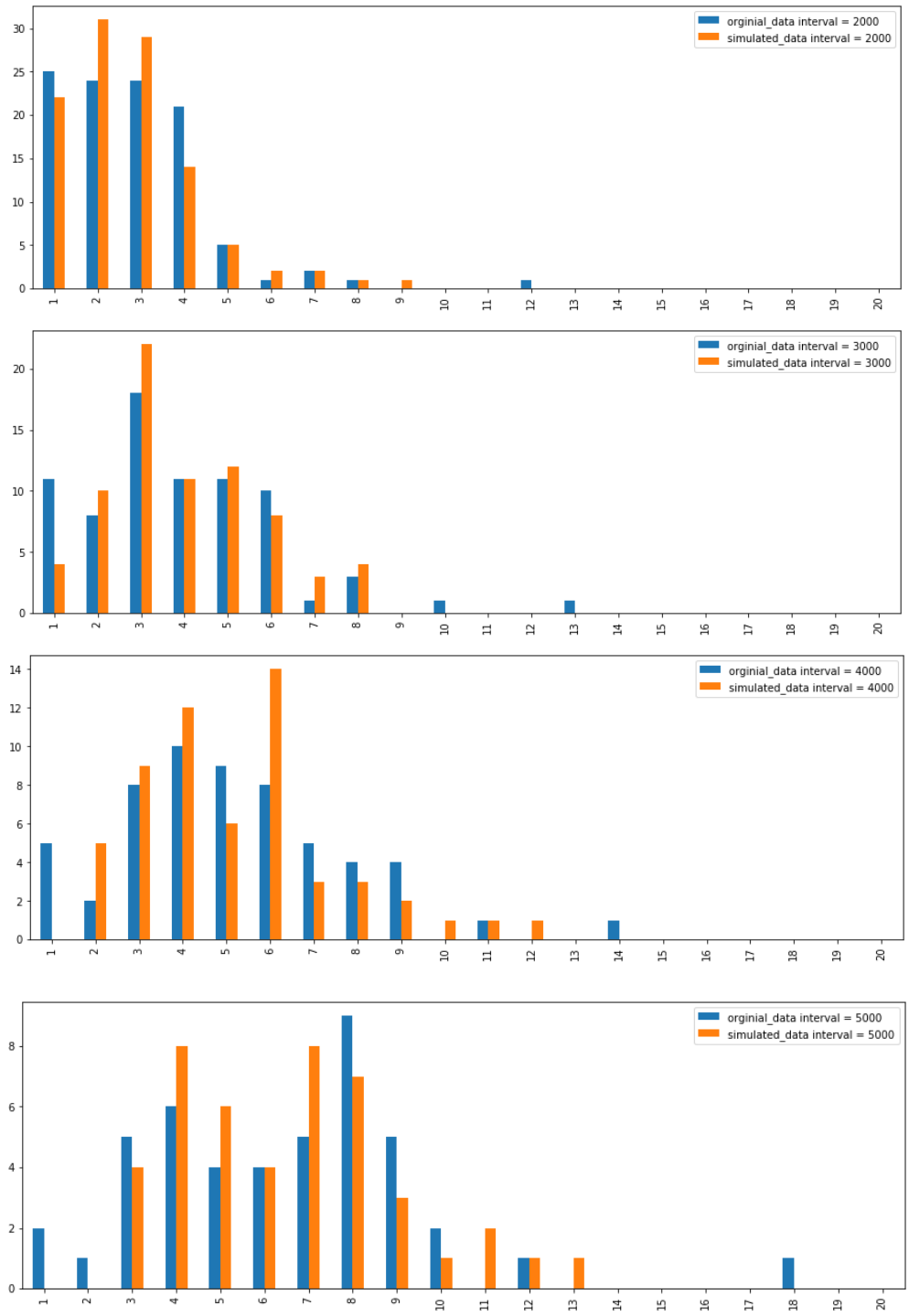



*Figure 15: Distribution of Counts of Palindrome in Non-Overlapping Regions*

Goodness of Fit Test

Our Null Hypothesis ($H_0$) is: Counts are realizations of independent variables from Poisson Distribution and Alternative Hypothesis ($H_1$): Counts are not realizations of independent variables from Poisson Distribution. Assuming the null hypothesis is true, we apply the chi-square

goodness of fit test to the data with a significant level $\alpha = 0.05$, and we expect the counts follows Poisson distribution. We again segment our data into 2000, 3000, 4000 and 5000 non-overlapping intervals. We first calculate the expected lambda which is the chance of containing palindrome in each interval, then we figure out the expected value according to Poisson distribution, finally we perform chi-square goodness of fit test to the counts of actual data and simulated uniform data four times accordingly. The table below include all the p-values calculated in each test corresponding to the number of intervals. We can see that all four p-values are relatively large. Since all p-values are greater than $\alpha = 0.05$, we cannot reject the null hypothesis thus keep believing that counts are consistent with Poisson distribution.

| K (# of non-overlapping regions) | 2000 | 3000 | 4000 | 5000 |
|---|---|---|---|---|
| P Value | 0.3754643358182173 | 0.7711521463799177 | 0.2393852869355991 | 0.4893186745810869 |

*Table 3: P-Values of Counts of Palindrome in Non-Overlapping Regions from Goodness of Fit*

Standardized Residual Plot

We also generated the standardized residual plot, and we observe that the residual plot has a fan shape and most of the residual data points are smaller that 3 regardless of couple outliers. This observation somewhat confirms the conclusion we draw from the goodness of fit test that counts are consistent with Poisson distribution.

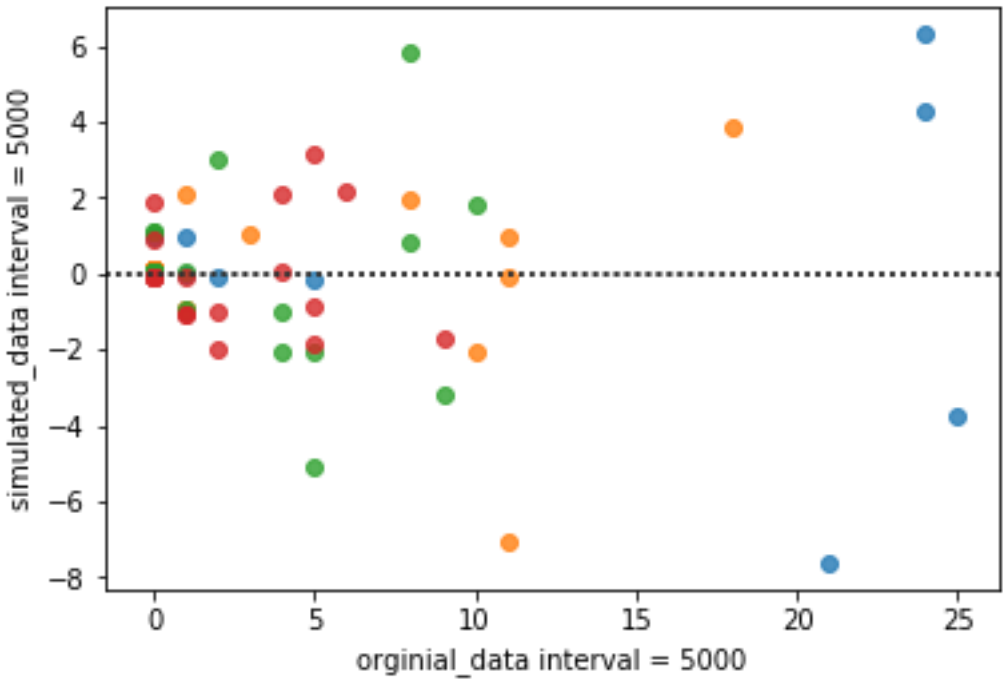



*Figure 16: Residual Plot of Counts of Palindrome in Non-Overlapping Regions*

Graphically, we observe that simulated distribution has a similar shape as the original, but the original data also has a signal of deviation from random scatter. Applying the chi-square goodness of fit test, we keep the null hypothesis that counts are consistent with Poisson distribution. By visualizing the residual plot, we somewhat confirm and have confidence in the conclusion from the chi-square test. In all, we tend to believe that counts are realizations of independent variables from Poisson Distribution.

**Scenario 4: The biggest cluster**

In order to find the biggest cluster and not letting it get away with the interval width, we use five different bins of equal length of intervals to find the cluster. (40, 60, 80, 100, 120) In the *Table 4*, the index is the different bins, and the interval width is calculated by dividing the length of DNA sequence (229,354) by the number of bins. During the previous investigations, we analyze that the distribution of the palindromes follows a Poisson distribution, thus we use the Poisson distribution to calculate the probability of getting am extreme number of palindromes based on the lambda. The lambda for the passion process is calculated by the mean count in each interval for the same bin. (This is the MLE estimation of the parameter lambda for the Poisson process, the mean of the sample is equal to lambda, see more about this in the Theory section of this project). We then find the maximum number of palindromes in the interval for each bin, and then calculate the probability of getting such maximum number based on the Poisson model and the lambda parameter we have. After computing all the test statistics, we found that under 0.05 significance level, all of the p-value is less than the significance level. Thus, there must be one or more unusual cluster appear in the DNA sequence, which does not occur by chance. And as we observed in the probability, the likelihood of getting maximum count of the palindromes decreases if we increase

our number of bins, which equal to decrease the length of the interval. After searching for the biggest cluster in each different bin cut, we found that it is the same cluster in the interval range (91700, 94000). And now we know that the maximum cluster is this one and has a very high probability of being the "origin of the replication".

| | lambda | interval_width | probability | maximum | prediction_interval |
|---|---|---|---|---|---|
| **40** | 7.4 | 5733 | 0.00510737 | 15 | (91728, 97461] |
| **60** | 4.93333 | 3822 | 0.000417868 | 14 | (91728, 95550] |
| **80** | 3.7 | 2866 | 2.5558e-05 | 14 | (91712, 94578] |
| **100** | 2.96 | 2293 | 1.1143e-05 | 13 | (91720, 94013] |
| **120** | 2.46667 | 1911 | 1.70567e-06 | 13 | (91728, 93639] |

*Table 4. The probability of getting the maximum in different length of intervals*

## Additional Hypothesis:

From investigation done above, we see that there is a great chance that an unusual cluster of palindromes exists in the interval of (91728, 93639). We are interested in whether if after we eliminate the palindromes in this interval, there will be another unusual cluster of palindromes. Our hypothesis is that there might be another possible unusual cluster of palindromes, indicating the "origin of the replication". To do this investigation, we eliminate all palindromes in this interval, and generate graphs for comparison and perform chi-square goodness of fit test with simulated Poisson distribution to test if there is any unusual cluster of palindromes left.

From *Figure 17* and *Figure 18* below, we can see that distributions of actual data does not deviate much from simulated uniform data by Poisson process. At least, we cannot make clear comparison just by graphs. Thus, we will conduct chi-square goodness of fit test of actual data after elimination of the suspect unusual cluster and simulated Poisson uniform distribution.

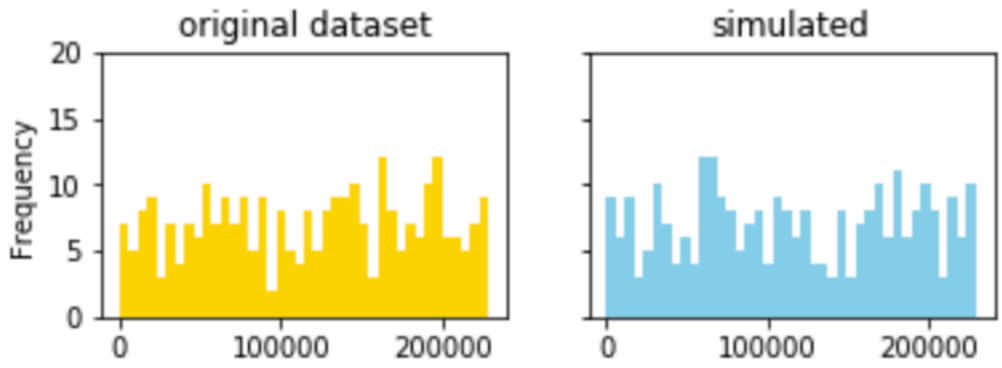



*Figure 17. Distribution of palindromes After Elimination of the possible unusual cluster detected above in data v.s. simulated uniform data for 40 intervals*

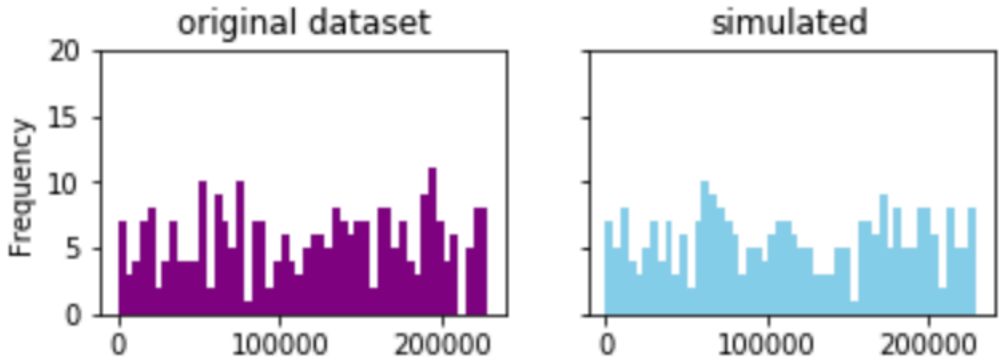



*Figure 18. Distribution of palindromes After Elimination of the possible unusual cluster detected above in data v.s. simulated uniform data for 50 intervals*

After performing chi-square goodness of fit test, we see that for both 40 and 50 intervals, test-statistics are large and we obtain extremely small p-values. This means that we are rejecting the null hypothesis that and that we can suspect another unusual cluster of palindromes besides the one we eliminate above.

| **Number of Interval** | **40** | **50** |
|---|---|---|
| **Test Statistic** | 99.24 | 119.53 |
| **P-Value** | 3.74e-07 | 7.96e-08 |

*Table 5. Results of Chi-Square Goodness of Fit Test of Actual Distribution and Simulated Uniform Distribution by Different Number of Intervals*

Since the p-values are small, we also generate residual plots. From the plots shown below, we see that several residuals are above 3, indicating again that our actual data has a chance of having another unusual cluster.

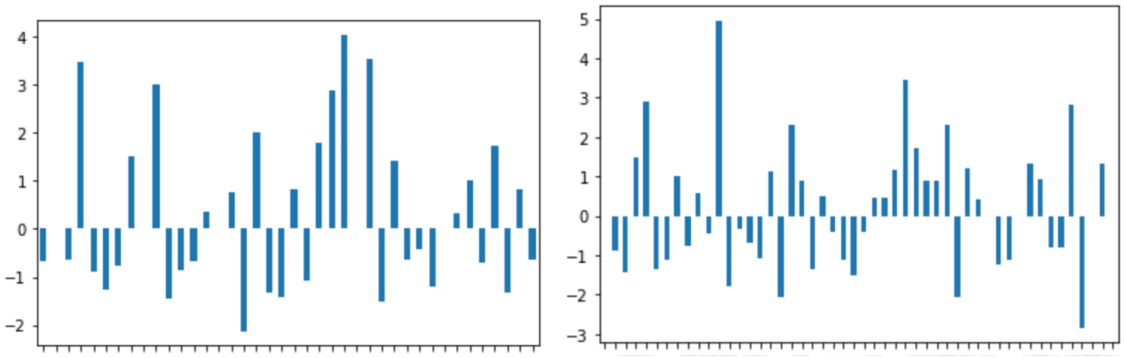


*Figure 18. Residual Plots of Actual Distributions and Simulated Uniform Data by Different Number of Intervals After Elimination of the possible unusual cluster detected above*

As there is a chance of another unusual cluster, we then find the maximum number of palindromes in the interval for each bin, and then calculate the probability of getting such maximum number based on the Poisson model and the lambda parameter we have. After computing all the test statistics, we found that under 0.05 significance level, all of the p-value is less than the significance level. By observing the result table shown below, we see that this possible second unusual cluster will be in the interval of (194905, 197198). Thus, we conclude that the second maximum cluster is this one and also has a very high probability of being the "origin of the replication" besides the one we eliminate above.

| | lambda | interval_width | probability | maximum | prediction_interval |
|---|---|---|---|---|---|
| **40** | 7.075 | 5733 | 0.0277809 | 12 | (194922, 200655] |
| **60** | 4.71667 | 3822 | 0.00226393 | 12 | (194922, 198744] |
| **80** | 3.5375 | 2866 | 0.00245971 | 10 | (194888, 197754] |
| **100** | 2.83 | 2293 | 0.000535856 | 10 | (194905, 197198] |
| **120** | 2.35833 | 1911 | 0.00761368 | 7 | (74529, 76440] |

*Table 5. The probability of getting the maximum in different length of intervals*

## Theory:

- Goals:
    - Understand a random model that describes the behavior of "counts" of the number of palindromes and for a "uniform" aka random scatter of palindromes.
- Homogeneous Poisson Process
    - The Homogeneous Poisson Process is a process arises naturally from the notion of points haphazardly distributed on a line with no obvious regularity. Most of time,

it is used as a model for random phenomena such as arrival times of telephone calls at an exchange, the decay times of radioactive particles, and the position of stars in parts of the sky.

- There are three characteristic features of the process:
  - Homogeneity: the rate λ at which hits occur will never change with location.
  - Independence: the number of hits falling in different intervals are independent.
  - No two hits can occur at the exactly same location.
- The counts of number of points in different intervals follow Poisson distribution with rate λ, which represents the rate of hits per unit. We have

  P( k points in a unit interval) $= \frac{\lambda^k}{k!}e^{-\lambda}, for\ k = 0, 1, \ldots$ and

  P( k points in an interval of length t) $= \frac{\lambda t^k}{k!}e^{-\lambda t}$. The expected value of the number of hits per unit interval is λ. In most cases, we don't know the exact λ. A good estimate would be the empirical average number of hits per unit interval, which could be get by either method of moments or maximum likelihood method.
- In this study case, we treat the strand of the DNA as a line, and the location of a palindrome can as a point on the line. According to the uniform random scatter model, palindromes are scattered randomly and uniformly across the DNA, which meets the homogeneity. Also, the number of palindromes in any small piece of DNA is independent of the number of palindromes in another and none of any two hits occur at the same point on the DNA.

- Chi-Square Goodness of Fit Test
  - In the study, we want to use the Homogeneous Poisson Process as a reference model to seek of some unusual clusters of palindromes. However, we need to estimate how likely the Poisson distribution fits the data. A technique for accessing how well the reference model fits to the data is to apply the chi-square goodness of fit test.
  - Sometimes a parameter of the distribution needs to be estimated in order to compute the probabilities. In this case, we use our data to estimate unknown parameter(s). The measure of discrepancy between the sample counts and the expected counts is $\sum_{j=1}^{m} \frac{(j^{th}\ sample\ count - j^{th}\ expected\ count)^2}{j^{th}\ expected\ count} = \sum_{j=1}^{m} \frac{(N_j - \mu_j)^2}{\mu_j}$, where m represents the number of categories( intervals), $N_j$ stands for the number of observations that appear in category j, j = 1, … , m and $\mu_j = n * P(an\ observation\ is\ in\ category\ j)$. The discrepancy we get follows approximate chi-square distribution with m-k-1 degree of freedom, where k is the number of parameters we estimated to obtain expected values. We use $X_{m-k-1}$ to get the p-value and if the p-value is less than significance level α, we need to doubt the fit of distribution.
- Location and Uniform Distribution
  - Under the Poisson process model for random scatter, if the total number of hits in an interval is known, then the positions of the hits are uniformly scattered across the interval. In other words, the Poisson process on a region can be viewed as a process that first randomly generates the number of hits, and then generated locations for the hits according to the uniform distribution. Here, the positions of

these palindromes are like 296 independent observations from a uniform distribution, so we can apply another Chi-Square Goodness of Fit Test.

- Spacing and the Exponential and Gamma Distributions
  - Distances between successive hits should follow an exponential distribution.
    P (the distance between the first and second hits $>$ t) = P ( no hits in an interval of length t) = $e^{-\lambda t}$.
  - Distances between the hits that are two apparatus, follows a Gamma distribution with parameters 2, λ.
- Maximum Number of Hits
  - Under the Poisson process model, the numbers of hits in a set of non-overlapping intervals of the same length are independent observations from a Poisson distribution. This implies that the greatest number of hits in a collection of intervals behaves as the maximum of independent Poisson random variables.
    - If we suppose that there are m such intervals then
      P ( maximum count over m intervals ≥ k)
      = 1 − P ( maximum count over m intervals < k)
      = 1 − P( all interval counts < k)
      = 1 − P ( first interval counts < k)$^m$
      $= 1 - [\lambda\text{^}0\ e\text{^}(-\lambda) + \cdots + \lambda\text{^}(k-1)/(k-1)!\ e\text{^}(-\lambda)\ ]\text{^}m$
  - For a given estimate of λ, from the above expression, we can find the approximate chance that the greatest number of hits is at least k. If this chance is unusually small, then it provides evidence for a cluster that is larger than the expected from the Poisson process. We can use the maximum palindrome counts as a test statistic, and the computation above provides the p-value for the test statistic.
- Method of Moments
  - The method of moments is a way to estimate population parameters, like the population mean or the population standard deviation. The basic idea is that you take known facts about the population, and extend those ideas to a sample. For example, it's a fact that within a population:
    - Expected value E(x) = μ
    - For a sample, the estimator (method of moments) is just the sample mean, x̄. The formula for the sample mean is:
    - sample mean formula:
      - $\overline{X} = \frac{1}{N}\sum X_i$
  - For a Poisson distribution with unknown rate parameter λ.
    - Method of Moments is one estimation technique that proceeds as follows:
      - Find E(X) where X has Poisson distribution with rate λ.
      - Express λ in terms of E(X)
      - Replace E(X) with $\overline{X}$ to produce an estimate of λ, called $\overline{\lambda}$
    - $E(X) = \lambda \ \ggg \overline{X} = \hat{\lambda}$
    - If higher moments need to be computed then $E(X^2)$ is replaced with $\sum_i {x^2}_i/n$
- Maximum Likelihood
  - Maximum likelihood, also called the maximum likelihood method, is the procedure of finding the value of one or more parameters for a given statistic which makes the known likelihood distribution a maximum.
  - from a Poisson distribution with unknown rate parameter λ.

Maximum Likelihood method searches among all Poisson distributions to find the one that places the highest chance on the observed data.

- For Poisson distribution, the chance of observing $X_1, \dots, X_n$ is

$$L_n(\lambda) = \prod_{i=1}^{n} f(X_i; \lambda)$$

- The log likelihood will be:

$$l(\lambda) = \sum_{i=1}^{n} (X_i log\lambda - \lambda - logX_i!) = log\lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n} \log X_i!$$

- We find the maximum by finding the derivative:

$$l'(\lambda) = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n = 0$$

and we check that the function $l$ is concave down in which the second derivative is less than 0.

- The estimate is $\hat{\lambda} = \overline{X}$
- The MLE estimator agrees with the MME estimator.

o Maximum-likelihood for continuous distributions is the same. Suppose we have an independent sample

$$X_1, \dots, X_n$$

o from an Exponential distribution with the unknown parameter θ. Now, the Likelihood function, given the data is

$$L(\theta) = \theta^n e^{-\theta \sum_i X_i}$$

and the log-likelihood function

$$l(\theta) = nlog(\theta) - \theta \sum_i X_i$$

By solving the last equation for $\theta$ we obtain:

$$\hat{\theta} = \frac{1}{\overline{X}}$$

- Mean Square Error
  - o The mean square error of an estimator $\hat{\theta}$ for a parameter $\theta$ is

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

  - o It measures the average of the squares of the errors – that is, the average squared difference between the estimated values and what is estimated.
  - o The MSE is a measure of the quality of an estimator – it is always non-negative, and values closer to zero are better.
  - o Many people memorize MSE as Var + Bias squared.
  - o Many of the estimators we use are UNBISED, but sometimes an estimator with a small bias will have a small MSE.
  - o Theorem: Under certain regularity conditions, as the sample size increases, the Maximum-likelihood estimator, $\hat{\lambda}$ satisfies

$$\hat{\lambda} \sim N(\lambda, \frac{1}{nI(\lambda)})$$

Where $I(\lambda)$ is called the Fisher's Information Matrix.

- Asymptotic Distribution

- Fisher's Information matrix is defined as

$$I(\lambda) = \mathbb{E}\left(\frac{\partial}{\partial\lambda} log f_\lambda(X)\right)^2 = -\mathbb{E}\left(\frac{\partial^2}{\partial\lambda^2} log f_\lambda(X)\right)$$

  Hence, as n increases

$$\sqrt{nI(\lambda)}\left(\hat{\lambda} - \lambda\right) \sim N(0,1).$$

- The approximate normal distribution can be used to build the 95% confidence interval for the unknown $\lambda$ as

$$\hat{\lambda} \pm 1.96\sqrt{nI(\lambda)}$$

- Hypothesis Test
  - Hypothesis test is a method of inference which refers to the formal procedures used by statisticians to accept or reject statistical hypotheses. In our study, the Chi-Square goodness-of-fit test and the test for the maximum number of palindromes in an interval, are two examples of hypothesis tests.
  - During the test, we first propose a null hypothesis, denoted by $H_0$, which is usually the hypothesis that sample observations result purely from chance and an alternative hypothesis, denoted by $H_a$, which is the hypothesis that sample observations are influenced by some non-random cause. Secondly, we select an appropriate test and state the relevant test statistic T. Thirdly, we derive the distribution of the test statistic under the null hypothesis from the assumptions and compute from the observations the observed value $\mathrm{t_{obs}}$ of the test statistic T. With the observed value $\mathrm{t_{obs}}$ and the distribution, we can get the p-value for our observations. Finally, we compare the p-value to the significance level ($\alpha$) and decide to either reject the null hypothesis in favor of the alternative or not reject it.
  - When we reject the null hypothesis, we don't know if we have been unlucky with our sampling and observed a rare event or if we are making the correct decision. Thus, we have this table to define the 2 types of error we could possibly make in the hypothesis tests.

| | Decision | | |
|---|---|---|---|
| Truth | | Fail to reject $H_0$ | Reject $H_0$ |
| | $H_0$ True | No error | Type 1 Error |
| | $H_a$ True | Type 2 Error | No error |

## Discussion & Analysis:

Since the human cytomegalovirus (CMV) is a life-threatening disease that we want to avoid of, the finding of the "origin of the replication" is very important to cut off the spread of the disease in order to further eliminate it. Since the DNA sequence is so long, it is very difficult for the biologist to cut the piece one by one to find the correct part in the DNA that has the function of replicating, thus, our job is to find the unusual big cluster of palindromes that has the very high potential containing such function. After carefully designing and investigating, we have made four scenarios in finding such cluster.

In scenario 1, we get a feel about our data through visualization of the clusters. Most importantly, we found that there is a cluster around $90000^{th}$ base pair with more than 15 frequency of palindromes in the DNA sequence unlike other clusters in the data. Moreover, the general shape of the scatter does not look like normally distribution. In order to conclude whether this unusual cluster is the protentional "origin of the replication". We have made some investigations in the following scenarios.

In scenario 2, we first divide our data into 30, 40, and 50 intervals and compare locations of palindromes of our data to simulated uniform distributions by poisson process. The graphs show that no matter how we divide our data, there seems to have a interval that has a high frequncey compare to our simulated data. Thus, we suspect that this cluster is the same cluster we found in the scenario 1. We than use Chi-Square fitness test to test if the data distribution fits uniformal distribution. After computing the p-value for each division, we observe that all of the p-value is smaller than 0.05 significance level, indicating that the distribution of location of palindromes does not fit the simulated uniform distribution. We can get the same result from the residual plots. Since this is a poisson process, we expect to have exponetial distribution for the spacing between each location of the palindromes. Moreover, Gamma 2 distribution for pairs of palindromes and Gamma 3 distribution for triplets of palindromes. However, after using graphical and Chi-Square Goodness of Fit test, we found that the spacing does not follow all of the mentioned distribution (exponential, Gamma 2, Gamma 3). Thus, this is the evidence that support our initial guess that there is a unusual cluster in our DNA sequence that could not happen by chance.

In scenario 3, we will examine the counts of palindromes in various regions of the DNA and determine whether Poisson distribution fits our data using graphical methods, goodness of fit test, and standardized residual plot. We first divide our data and simulated data into 2000, 3000, 4000 and 5000 non-overlapping regions. And then we count the number of the palindromes in each region for each interval. From all the plots we have, we observe that the count range of 1-9 is around same for both data and simulated data for each interval, and the distributions of the counts are all skewed to the right. But in our original data, there is a outlies around 14 counts for each interval. This is may be due to the unusual cluster we found in the previous clusters. From the graphs, we conclude that the original data deviates from random scatter. Moreover, we want to find if Counts are realizations of independent variables from Poisson Distribution. If the palindromes occur randomly, it should follow a Poisson process. Therefore, we apply the chi-square goodness of fit test to the data for all the non-overlapping intervals we have, we can see that all the p-value is larger than the 0.05 significance level, and thus, we can conclude that the distribution of counts is consistent with Poisson distribution. We also can get the same conclusion from the residual plot that we have done.

In scenario 4, for five different numbers of equal cuts in the DNA sequence (40, 60, 80, 100, 120), we get the lambda value by averaging the counts of palindromes in each interval, and then use it to calculate the probability of getting the maximum count in the intervals for the same cut. All of the p-values are below 0.05 significance level, thus indicating that there are one or more unusual clusters in the DNA sequence. After searching for it in each cut, we found that it is the same cluster in the interval (91700, 94000). And now we know that the maximum cluster is this one and has a very high probability of being the "origin of the replication". Therefore, the next step of understanding the human cytomegalovirus (CMV) is to hand over to the biologist for them to cut into segments.

In the additional hypothesis, we want to find another potential cluster beside the biggest cluster we found, just in case our biologist could not find the "origin of the replication" in the

biggest cluster. From the graphical comparison, we cannot tell the difference of the original data with the simulated data after getting rid of the biggest cluster that we found. Thus, we make a chi-square goodness of fit test on whether the clusters in our dataset occurs at random. However, no matter how we choose the bins (40, 50), the p-value are both smaller than the 0.05 significance level, thus indicating that there are one or more unusual clusters in the DNA sequence. And the residual plots also support that. After using MLE to test the probability of getting the cluster with such extreme frequency (12) in the DNA sequence, we found that the probability is so small that it is almost impossible to get such cluster at random. After searching for it in the sequence, we found the second largest unusual cluster, it occurs at (194905, 197198).

From all the scenarios above, we can conclude that there existing an unusual cluster in the DNA sequence, which cannot occur by chance. Since the whole DNA sequence is so long, it is very hard to cut the DNA sequence one by one. Therefore, in order to find the "origin of replication" for the human cytomegalovirus (CMV), we suggest that the biologist should experiment in the (91700, 94000) base pair on the DNA sequence, which is highly likely to be the location of the "origin of replication". From additional hypothesis, if the biologist could not find the "origin of replication" in the biggest cluster, we suggest doing experiment on (194905, 197198) base pair on the DNA sequence, as it is the second largest unusual cluster we found.

## Work Cited:

Anders, DG, Punturieri, SM. 1991. "Multicomponent origin of cytomegalovirus lytic-phase DNA replication." J Virol 65:931–937.

Bradic, Jelena. "Chapter 4: Patterns in Data." MATH 189 Lecture, UC San Diego. Lecture.

D. Nolan & T. P. Speed (1999) "Teaching Statistics Theory through Applications." The American Statistician, 53:4, 370-375, DOI: 10.1080/00031305.1999.10474492

Masse, M. J., Karlin, S., Schachtel, G. A., & Mocarski, E. S. (1992). "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*, *89*(12), 5246-50.