How to Identify Patterns in DNA

Xiaohe Luo*[1], Wenhao Sheng[2], Guannan Shi[3], Shuli Ruan[4], and Hao Jiang[5]

[1] *3rd year undergrad: Math-Econ, A91116453*
[2] *1st year grad: Electrical Engineering, A99033876*
[3] *1st year grad: Electrical Engineering, A98051827*
[4] *4th year undergrad: Math-Econ, A13222376*
[5] *1st year grad: Electrical Engineering,A91426797*

February 24, 2017

1 Introduction

Virus is a severe factor that can endanger human's health. Cytomegalovirus (CMV) is one of the virus that can cause a lot of life-threatening diseases for people who have weak immune system. In order to having a good health, scientists determine to study the method of virus' replication to prevent the harm of virus. Since the DNA of virus contains all of the information about how virus grow, survive and replicate, investigating the construction of DNA is a key factor to find out the replication of virus. DNA sequence consists of four kinds of nitrogen-containing nucleobases—either cytosine (C), guanine (G), adenine (A), or thymine (T). DNA sequences also contain many patterns. One of the patterns, namely, complementary palindrome structure (for example, GGGCATGCCC) may flag important sites on the DNA. In addition, complimentary palindromes are most likely to be considered the origin of replication for two viruses. CMV, a member of the herpes family, is marked by complimentary palindromes. One of them, Herpes simplex, is marked by a long palindrome of 144 letters. However, for the CMV, the longest palindrome is 18 base pairs, and contains 296 palindromes between 10 and 18 base pairs long. In order to find the origin replication, biologists need to cut the DNA into segments and, test whether it can replicate or not. This process can be very expensive and time-consuming as well. Thus, in this report, we want to find unusual cluster of complementary palindromes to help the biologist reduce the amount of testing needed to find the origin of replication.

2 Methods

2.1 Reference Models: Uniform, Poisson, Exponential

As the goal of this project is to identify irregular patterns in the DNA genomes, it is necessary to first hypothesize reference models that the data is supposed to follow. To be more specific, according to the setting of DNA genomes, the appearance of palindromes should be random if there are no replication sites on this DNA. In other words, without palindromes that are replication sites, the starting location of all palindromes should follow a uniform distribution in the interval from 0 to 229354, since there are 229354 bases in our DNA.So the first step of this project is to generate 296 numbers from a uniform distribution in the interval [0,229354], and compare the simulated data sets with our data numerically and graphically. So, if we are given that the locations of palindrome sites are uniformly distributed, it is equivalent to say that the hits occur randomly independent with the past. Now, instead of looking at the number of hits, we look at the locations between each consecutive hits, it is an exponential distribution. Consequently, in order to determine whether a certain segment of the DNA carries CMV origin of replication, we would need to check the following three statements:

[*] 1-5 Corresponding author

- Is the location of hits follows a uniform distribution?
- Is the number of hits with certain interval follows a Poisson distribution?
- Is the spacing between two consecutive hits follows a exponential distribution?

If the starting locations of palindromes follow a uniform distribution, then on each genome position, the probability of having a palindromes is $\frac{296}{229354}$. By theory, in an interval consisting of many identically and independently distributed random variables, genome positions in this study, and each random variable follows a Bernoulli distribution with probability 296/229354, then in this interval, the number of palindromes will follow a Poisson distribution with parameter λ. In addition, the space between every two palindromes should follow an Exponential distribution with parameter λ.

2.2 Simulation

To do simulation, we want to first find an appropriate number of intervals. Hence, for each simulation, we separate the whole DNA two to three times according to different interval length.

- Uniform Distribution: As stated in last section, we divided the whole DNA into 40, 50, 60 and 70 intervals. Then, we want to generate a set of 296 pseudo random numbers from a sequence of 0 to 229354. Lastly, we compare the number of observed points in each interval and the pseudo random points generated by R in each interval. To make the comparison more straightforward, we plot a histogram with locations as the x-axis and counts of palindromes in each interval as the y-axis.

- Poisson Distribution: For the Poisson distribution, we use the same set of number generated when examining the Uniform distribution. However, since only the number of palindromes in a given interval follows a Poisson distribution, it is helpful to calculate the frequency of intervals according to the number of palindromes in an interval. For example, we observe the number of intervals that contain 5 palindromes and then we want to compare this Then we plot a histogram with the density(counts or frequency) as the y-axis and the number of palindromes as they x-axis.

- Exponential Distribution: For exponential distribution, we are interested in the spacing between two consecutive palindromes. To run the simulation, we simply divided the spacing into equal size intervals with length = 500. We also made sure that the number of counts in every interval varies.

2.3 Method of Moments

In order to compare the information encoded in our data to different probability models, one important task is to estimate parameters of the model. Here, we have two choices of estimation methods. One is the maximum likelihood estimate (MLE), the other is the Methods of Moments (MOM). For Poisson and exponential distributions, they both have only one parameter, namely, λ. For the Poisson distribution, The rate λ is the rate of hits per unit. It is also the expected number of hits per unit interval. A good estimate of this λ is the empirical average number of hits per unit interval. This method of estimating the parameters is called the Method of Moments. We can very well do the estimation using MLE, but for the Poisson and exponential distributions, MOM and MLE give the same results. Hence, our group decided to use MOM in our analysis.

3 Analyses

3.1 Patterns indicated by Uniform Distribution

In this part, it is important to test whether the location follows the uniform distribution or not. In order to search the specific place where the unusual clusters are, we generate 296 numbers in the interval from 0 to 229354 under a uniform distribution, and comparing the data we get with the simulated data sets numerically and graphically. We split our data as region = 40, region = 50, region = 60, and region = 70 and graph them respectively. In the histogram of region = 40 (Figure 1), there are two clusters of palindromes appear in [91741,97475] and [189217,194950]. While in the histogram of region = 50 (Figure 2), two clusters of palindromes show up in [91741,96328] and [197244,192657]. However, if we make our region larger, there are more and more clusters appear in the DNA. By viewing the histogram of region = 60 (Figure 3), we can deter that there are three clusters in [91741,95564], [164370,168192] and [194950,198773]. And in the graph of region 70 (Figure 4), there are five clusters in [52423,55700], [75359,78635], [88465,91741], [140888,144165] ,and [196589,199865]. Therefore, we can conclude that different length of region can cause different number of clusters appear in the DNA. Then, we compare the number of observed points in each interval and the pseudo random points generated by R in each interval. After that we do the uniform distribution test with corresponding regions with our observed and simulated data. According to the results in Table 1, we observe that when the interval length gets finer, the location of palindromes given by our data is becoming more and more non-uniform.This is a sign of the importance of setting the intervals length.As we can see in this table, if the interval length is too large, than the intervals containing more palindromes would become less abnoraml because every interval has many palindromes. On the other hand, if the interval length is too small, then having 1 or 2 more palindromes will make the interval look different from a unifrom distribution. Since the first 3 ways of deciding the number of intervals all yield the same result that we cannot reject the null hypothesis that our data appears to be uniform, we conlude that the location of palindromes provided by our data has approximately a Uniform distribution.

Table 1: Uniform Distribution Test Results

n.region	p-value
40	0.9970297
50	0.2345429
60	0.04214039
70	0.02593189

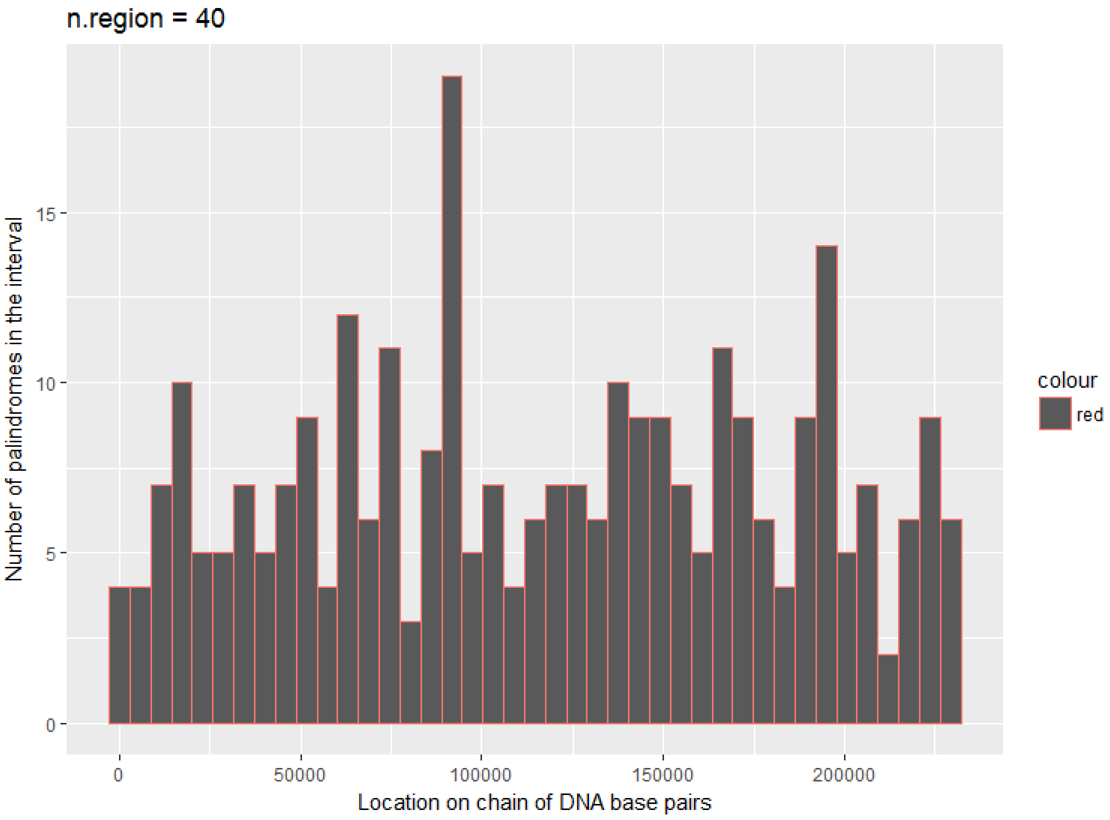

Figure 1: Histogram for Uniform test, n.region=40

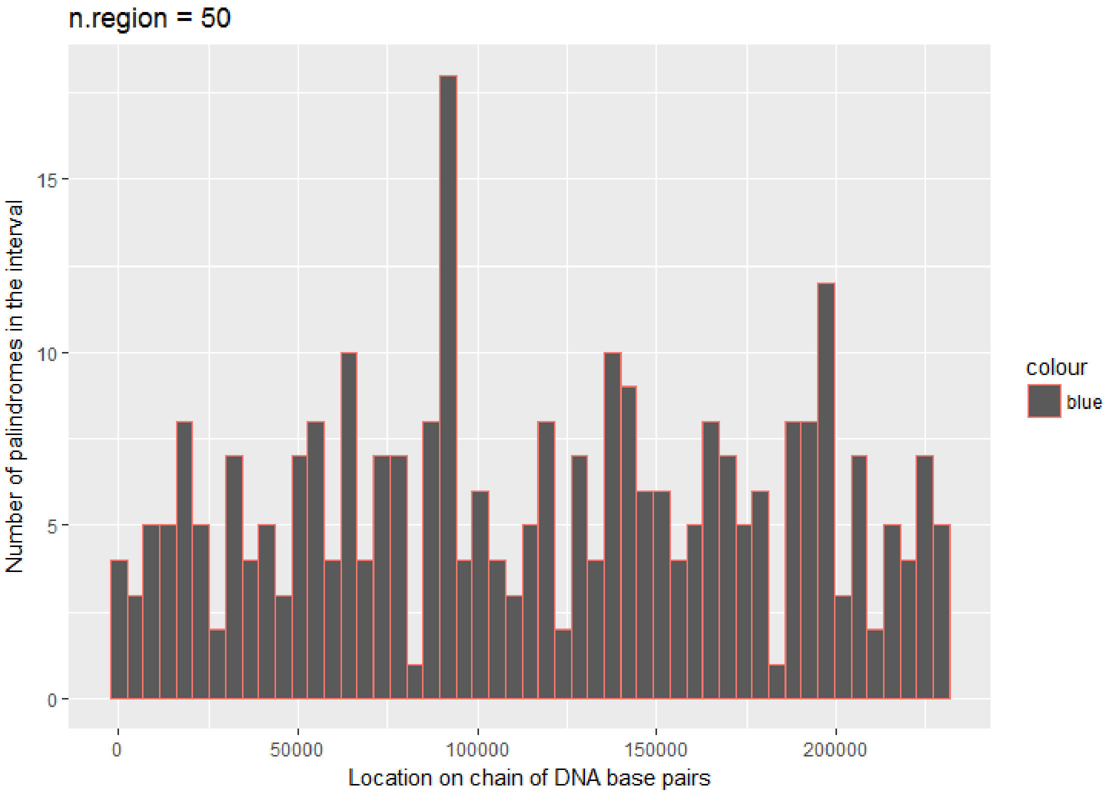

Figure 2: Histogram for Uniform test, n.region=50

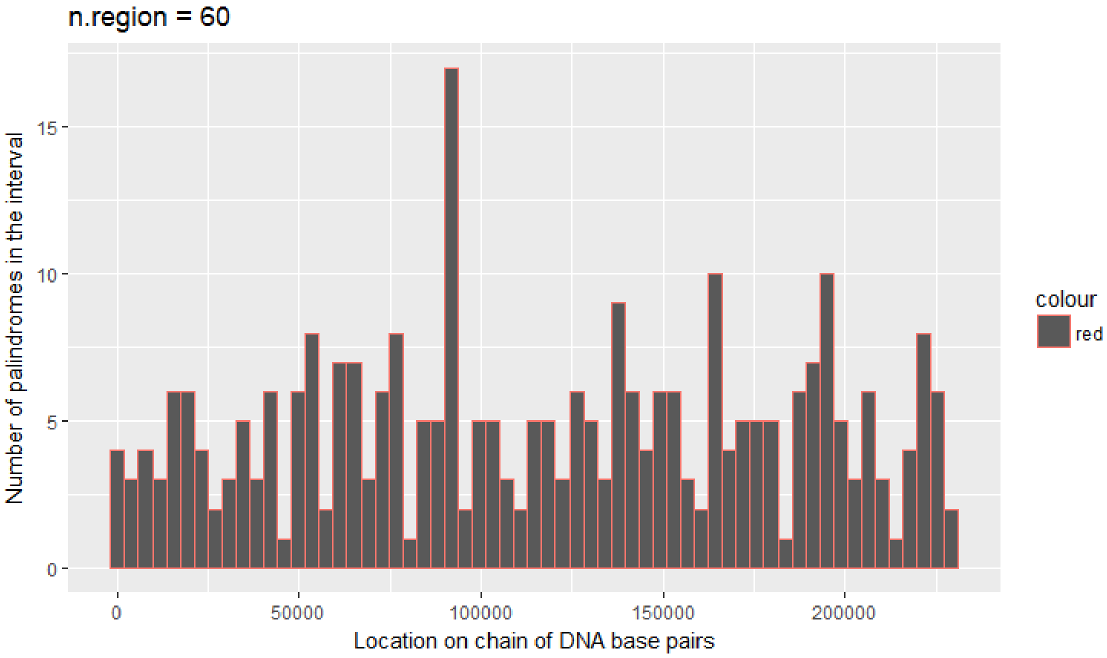

Figure 3: Histogram for Uniform test, n.region=60

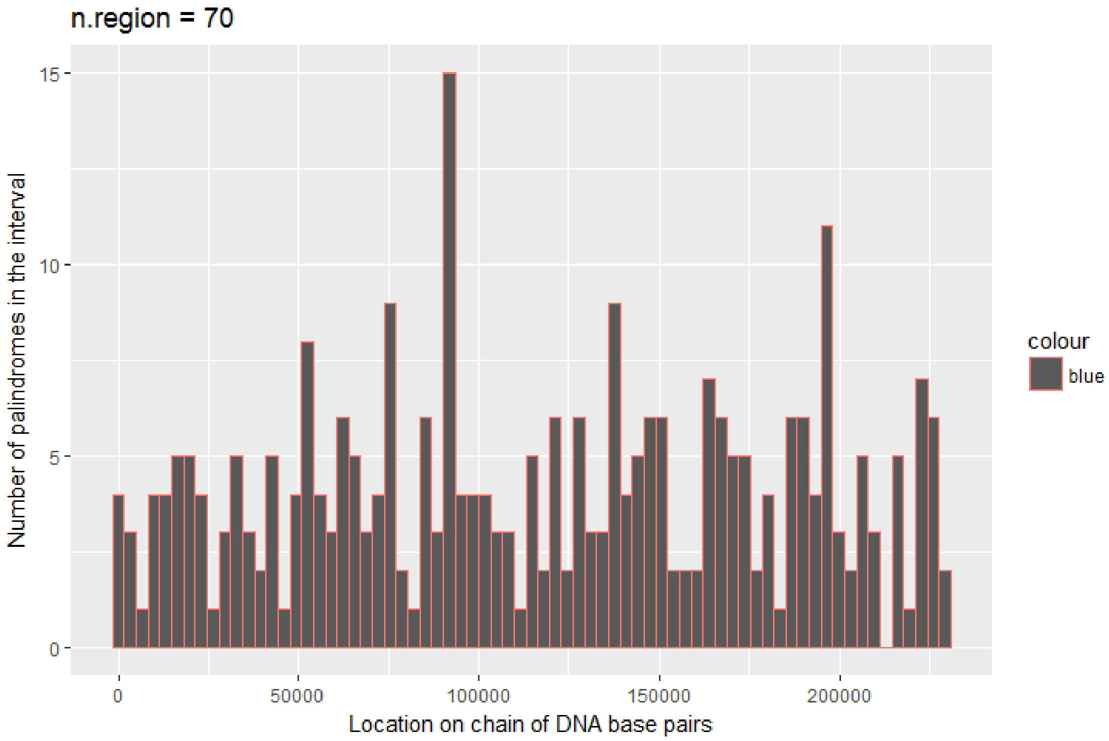

Figure 4: Histogram for Uniform test, n.region=70

3.2 Patterns indicated by Poisson Distribution

Here, we intend to examine whether the count of hits inside all segments of the DNA follows a Poisson distribution. The null hypothesis that our group has is that it does have a Poisson distribution. And, the interpretation of the null hypothesis being true would consequently be that there is no abnormal behavior in the DNA. In other words, it is unlikely to find an origin of replication site of CMV on this segment of DNA.

One thing to note here is that the size of the interval is of our choice. While a long interval would very likely gives us a distribution that looks very similar to a Poisson distribution. However, if the interval is chosen to be too large, we might not be able to capture potential abnormal behavior of the palindromes inside each interval. Alternatively, if we choose a shorter interval, every subtle deviation from a normal behavior will be seen, but chances are that we will end up with most intervals having very small amount of palindromes. Hence, the distribution will be extremely skewed to low palindrome occurrence, which will make us reject our null hypothesis, and cause a type II error. Considering the risks, we decided to try out multiple size of intervals, and look for the best size. If we stick to our null hypothesis, we would be looking for the interval size that make the distribution most likely Poisson.

The following plots graphically show how changing the number of intervals alters the distribution of hits within an interval. The first plot depicts the case when there is only 40 intervals, while the second and the third show the cases when there is 50 and 57, respectively. The blue bars is our simulated Poisson distribution, and the red bars comes from our data.

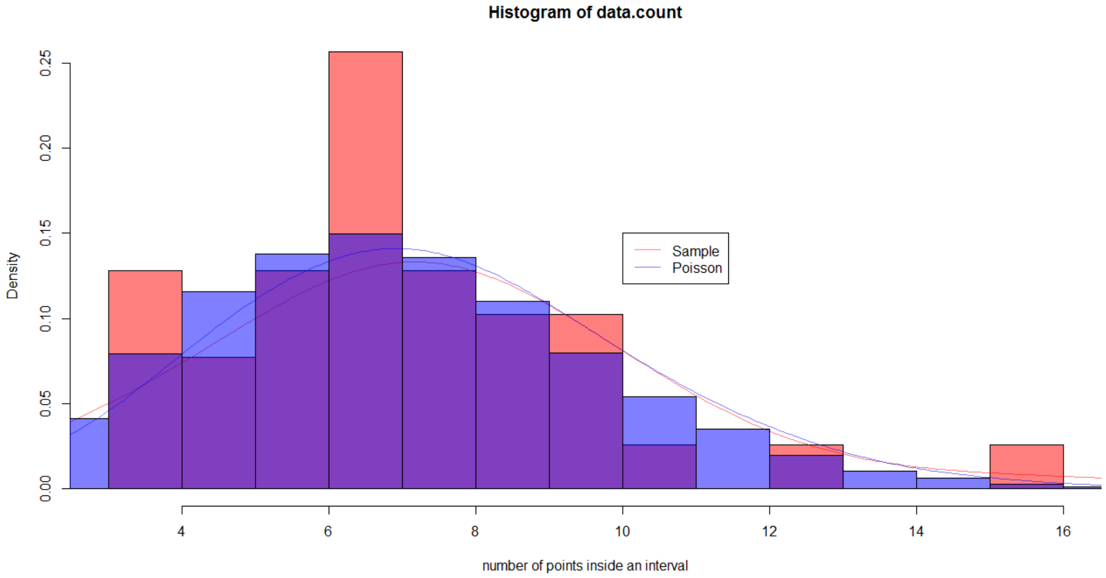

Figure 5: Histogram Poisson test, n.region=40

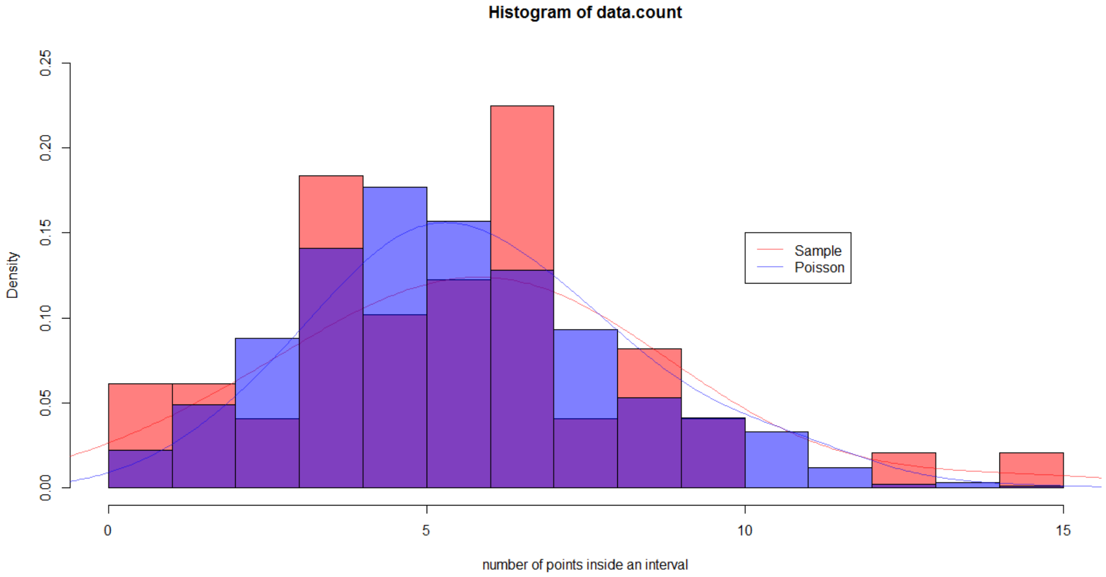

Figure 6: Histogram for Poisson test, n.region=50

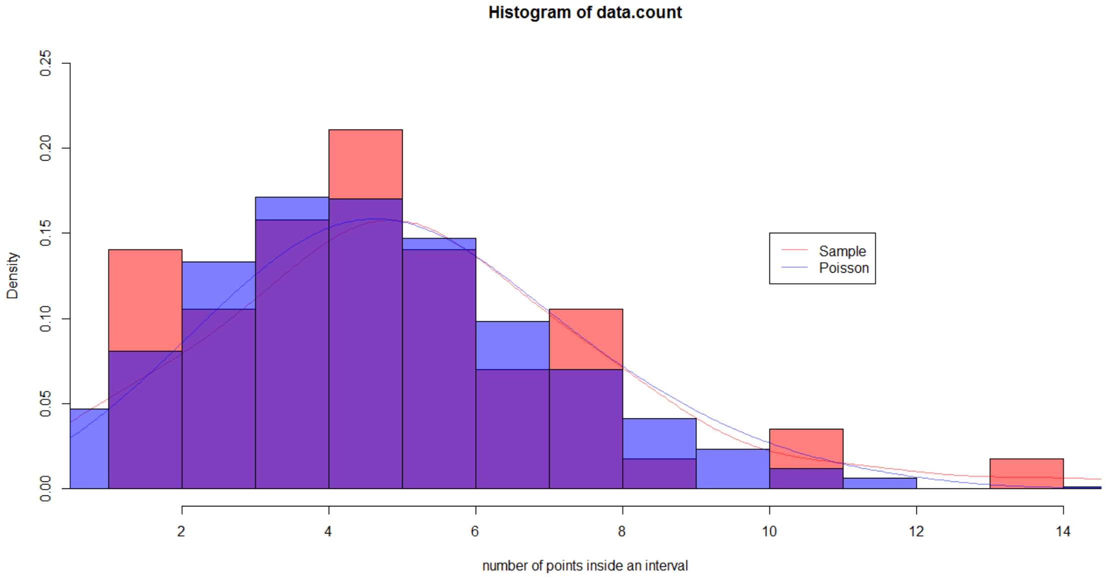

Figure 7: Histogram for Poisson test, n.region=57

Actually, these are not the only interval numbers our team experimented. However, they are representative in a sense that they are showing some trend that we can trace. To be specific, in the case when the number of interval is 40, the distribution looks quite like a Poisson except for the second, the fifth, and the very last bar being the outliers. A closer examination on the curves of our data distribution and the simulated Poisson distribution confirms that our null hypothesis is true. Same thing happens to the case when we pick a large interval number n.region = 57. However, as we argued earlier, choosing a interval number that is either too large or too small is problematic. When we choose an interval number which falls between 40 and 57, such as in the second figure above where the interval number was set to 50, the distribution of the data actually deviated quite a lot from the Poisson distribution. Notice that the distribution curve of our sample no longer collides with the Poisson curve. Hence, to make a conclusion, we then do a Pearson's chi-square goodness of fit to test these three conditions.
The detailed explanation of how Pearson's chi-square test works is moved to the 'Theory' section. The following tables shows the results of the chi-square test. And, the graphs are the respective residual plot for each of the three cases we discussed above.

Table 2: first test, interval length being 50, p value = 0.1364952

Interval with Counts	Number of Observed	Intervals Expected
0-2	4	3.281744
3	4	4.642592
4	8	6.871037
5	8	8.135307
6	5	8.026836
7	9	6.78841
[8,9]	9	8.32772
[10,15]	3	3.866017

The table 2 above is showing the number of observed intervals that have the specified counts, and it also includes the expected number of intervals under each condition. As the table 2 shows, the p-value is 0.1365 when the interval number is 50, therefore, we can't reject the null hypothesis that the distribution is following the Poisson distribution. Nevertheless, there are noticeably deviations from the expected interval. For instance, number of intervals observed with counts 6 is only 5, while its expected number of interval is over 8. Also, for the interval with counts 7, the expected number is 6.7, however, we had 9 observations. So, there is a potential that among the intervals that have either 6 or 7 hits, an origin of replication virus might be found.

As long as we think the Poisson is a good fit for this data (as the our calculated p-value is larger than 0.05), we can use the Figure 8, which is the standardized residual plot, to find the abnormal interval counts. With the interval number being 50, the abnormal counts are at count 6 and count 7. We are going to exam the other conditions before we make a conclusion for this part.

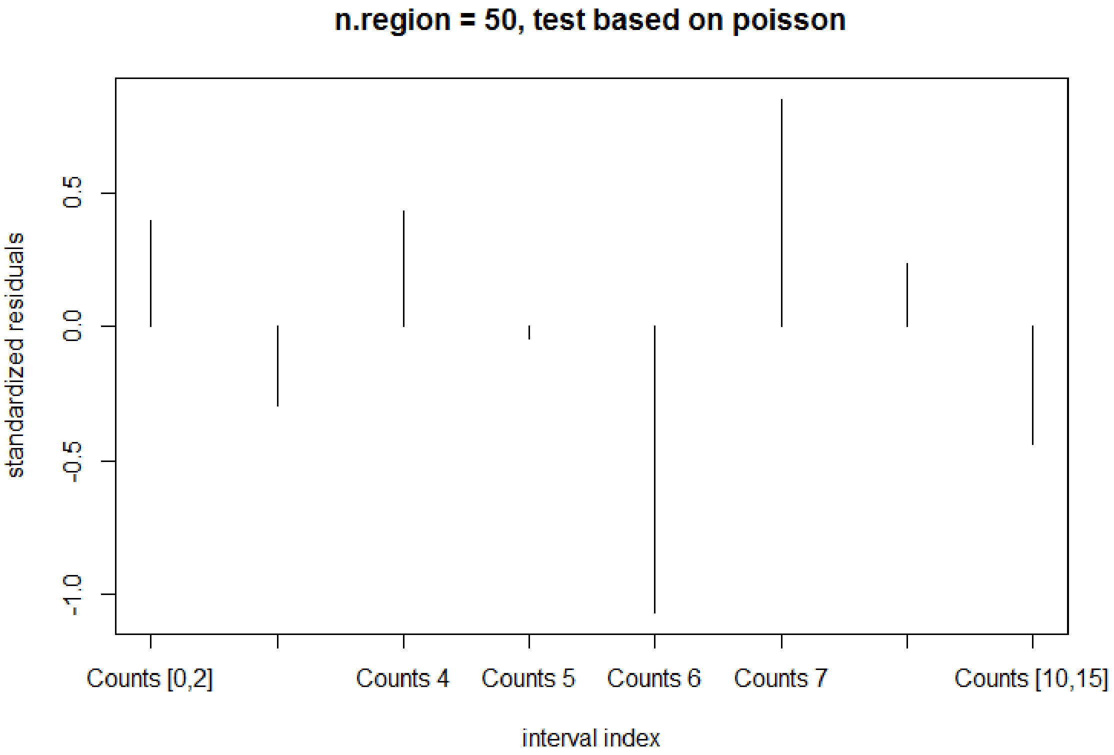

Figure 8: Residual Plot for Poisson test, n.region=50

Table 3: second test, interval length being 57, p value = 0.9540093

Interval with Counts	Number of Observed	Intervals Expected
count 1	6	1.961061
count 2	5	4.269676
count 3	5	7.390785
count 4	5	9.595054
count 5	10	9.965389
count 6	8	8.625015
count 7	8	6.398507
count 8	7	4.153417
count 9-13	3	4.485561

The table 3 is the second chi-square test for 57 intervals, and the P-value is as big as 0.95, which indicates that the sample distribution under this condition is almost following the Poisson distribution. Again, we could use the Poisson distribution with the specified lambda to find the abnormal counts, and the residual plot for Poisson fitting test is shown as Figure 9. In this figure, the count number 1 is abnormal, and the count number 3, 4, 7 and 8 are suspicious.

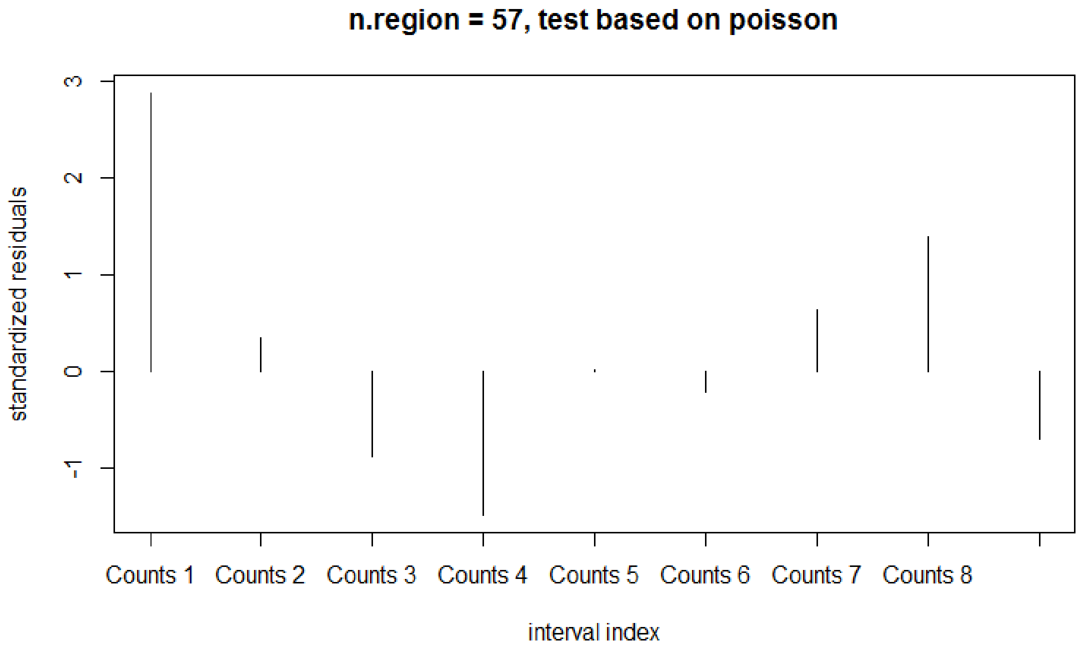

Figure 9: Residual Plot for Poisson test, n.region=57

Table 4: third test, interval length being 40, p value = 0.7969956

Interval with Counts	Number of Observed	Intervals Expected
count 1	6	1.961061
count 2	5	4.269676
count 3	5	7.390785
count 4	5	9.595054
count 5	10	9.965389
count 6	8	8.625015
count 7	8	6.398507
count 8	7	4.153417
count 9-13	3	4.485561

The table 4 is the third chi-square test for 40 intervals, and the P-value is as big as 0.7969956, which indicates that the sample distribution under this condition is pretty much following the Poisson distribution, and the residual plot for Poisson fitting test is shown as Figure 9. In this figure, the count number 7,8 and 9 is abnormal.

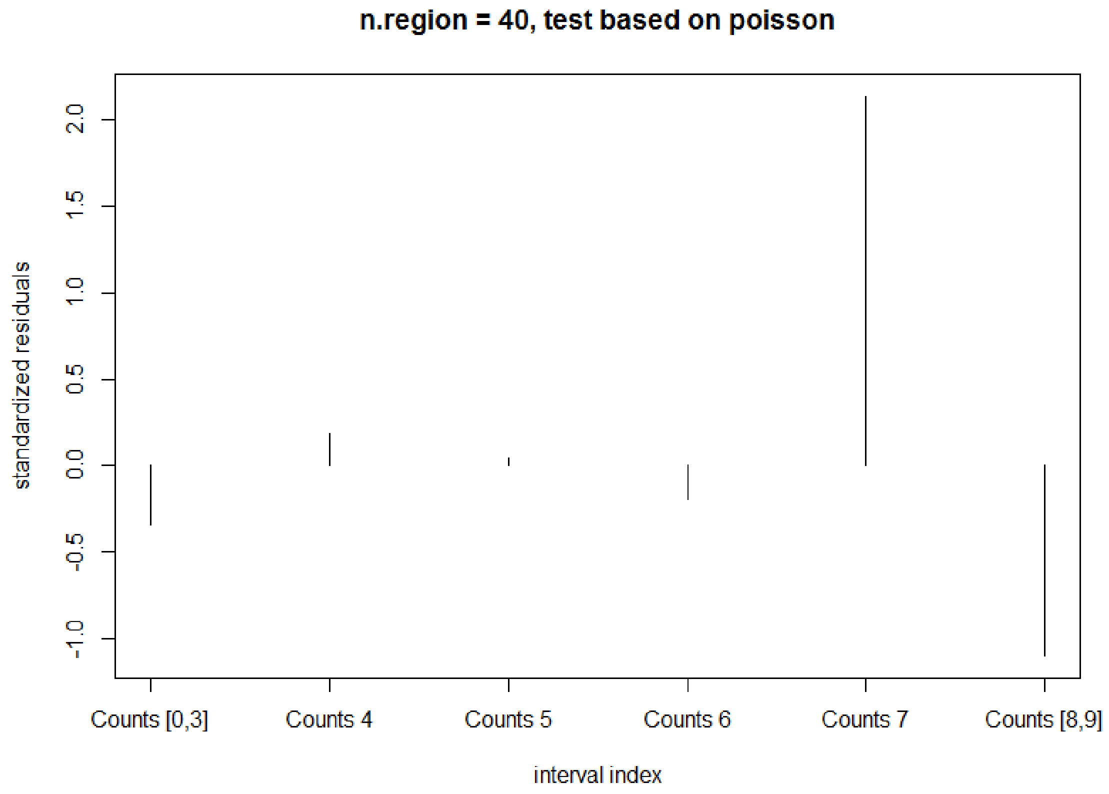

Figure 10: Residual Plot for Poisson test, n.region=40

As we notice that the sample distribution is following the Poisson distribution, we want to use the residual plots to find the abnormal counts to help biologist to locate the replication site of CMV. When we make the interval small which means the interval number is large, such as 57 intervals here, we notice that the standardized residual of count 1 is almost 3, and this should be a result that we separate a cluster of palindromes into small pieces. When we increase the interval size, such as 50 intervals, the count 1 and 2 is still bigger, but the count 7 and 4 immediately goes up, and the negative residual of count 6 also represents that there are should be several clusters with counts at least 7 when there are 50 intervals, and the count 4 is just the separated from the cluster by the intervals. When we make the interval much larger, and the interval number becomes 40. At this time, the count 7 is very large, and this support our thought that the cluster should have 7 counts in a region. We are going to do a test about the spacing to find the relationships between the palindrome cluster and the spacing.

3.3 Patterns indicated by Exponential Distribution

First of all, it is important to examine if the space between every consecutive pairs of palindromes follows an exponential distribution.If the space between every pairs of consecutive palindromes follows an exponential distribution, then it will be easier to identify the abnormal behaviors that don't confront with expectation of an exponential distribution. In order to achieve this goal, the first thing needed to be done is to determine a set of ranges of values of space,in which we hope to compare the observed frequency to the expected frequency according to the Exponential distribution.As mentioned in the section 2.3, we calculate the expected frequency by the method of moment. In order words, we provide an estimation of the parameter λ with the following formula: $\lambda = \frac{\sum_{i=1}^{295} X_i}{295}$, where X_i represents the data point of spacing. Following the formula of the Exponential distribution with our estimated λ, we are able to calculate the probability that the random variable X:the space between 2 consecutive palindromes, falls into the corresponding range. The result is provided in Table 5.

Table 5: Comparison of Observed Counts of Spacing and Expected Counts

Spacing Interval	Observed counts	Expected counts
[0,500]	146	140.183601
(500,1000]	56	73.568543
(1000,1500]	47	38.608871
(1500,2000]	25	20.261988
(2000,2500]	7	10.633519
(2500,3000]	5	5.580485
(3000,3500]	6	2.928646
(3500,6000]	3	3.105593

After the numerical comparison, it will be also helpful to see if the exponential model really fits our data graphically. The histogram generated by our data indicates that the density of the space between two consecutive palindromes does appear to be similar to the density distribution of an exponential distribution. Although eyebowling gives us a general perception, it is not precise and rigorous enough to conclude that the distribution of our data is approximately an exponential distribution. To be rigorous, it is necessary to run a Chi-square test based on Table 5.

The null hypothesis of this Chi-square test is that the space between two consecutive palindromes follows an exponential distribution. The formula of Chi-square is $X^2 = \sum_{i=1}^{n} (Oi - Ei)^2 / Ei$, in which Ei corresponds to Expected Counts in the table and Oi corresponds to Observed Counts in Table 5.The results based on our data are $X^2 = 11.89499$ and p-value=0.06435285. Hence, at a 5% significance level, we can not reject the null hypothesis that the space between a pair of consecutive palindromes is approximately exponentially distributed. With this result, it is meaningful to find out in which intervals the observed frequency deviates from the expected frequency a lot.

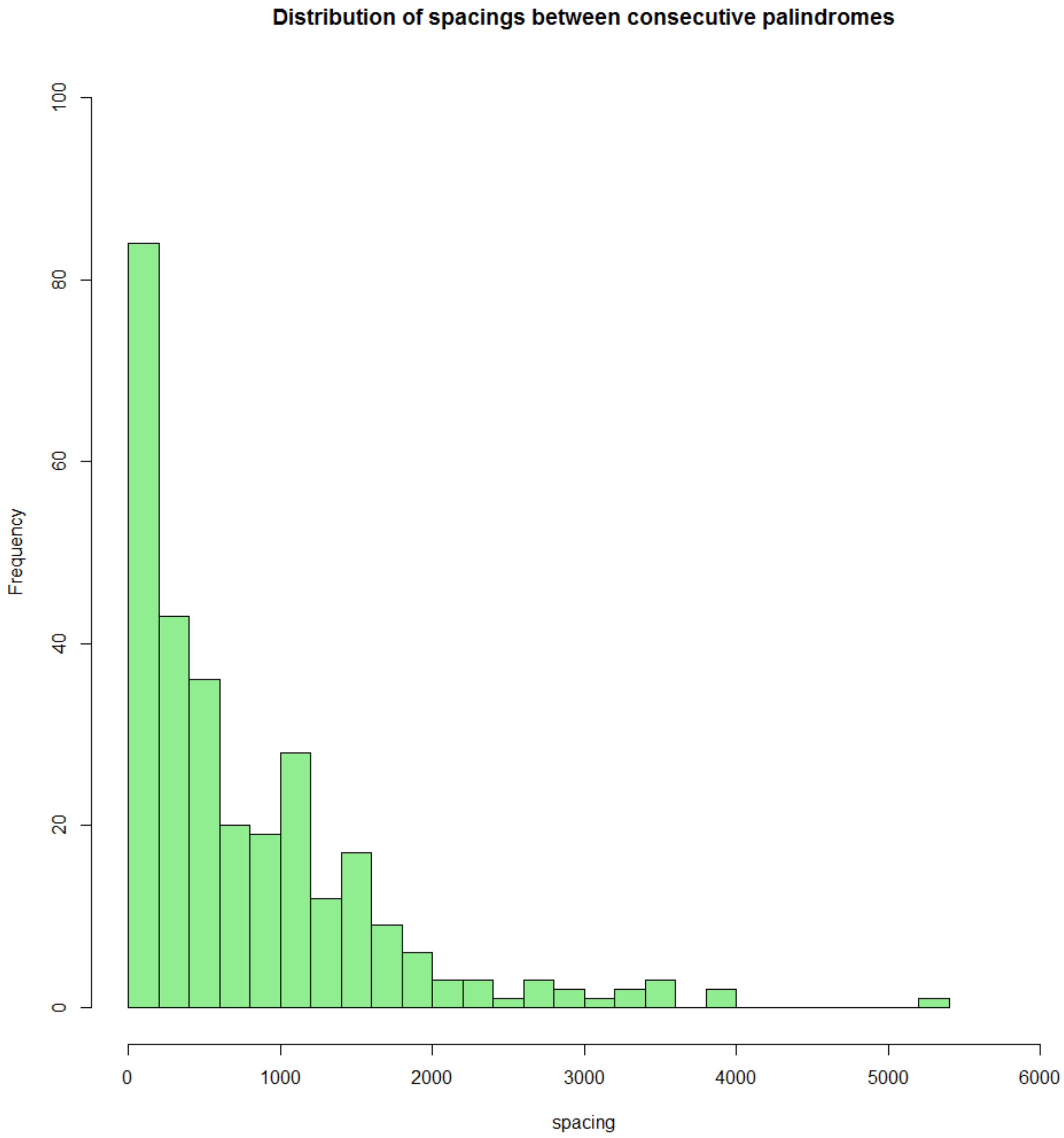

Figure 11: Histogram for Exponential test

The discrepancy between observed values and theoretical values can be illustrated by the standardized residual plot.

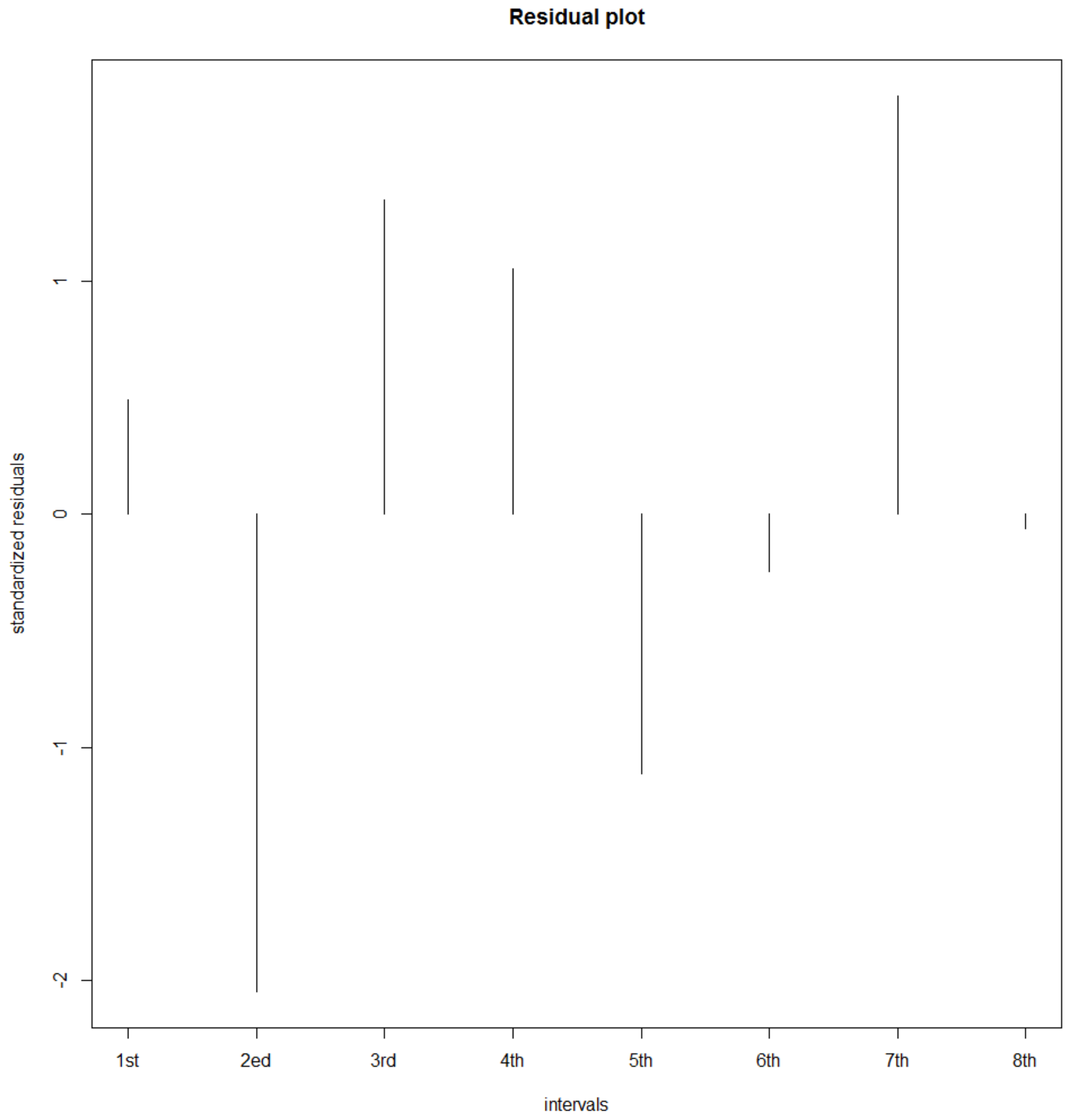

Figure 12: Residual plot for Testing Consecutive Palindromes spacing

Based on the above residual plot, now we are able to determine which palindromes are more likely to be abnormal. In the residual plot, it indicates that the residuals are relatively larger in the 2nd, 3rd and 7th intervals, which are corresponding to spacing interval 500-1000, 1000-1500 and 3000-3500. In these three intervals, the observed counts of palindromes deviate from the expected counts calculated based on exponential distribution. Therefore, we can conclude that the 500-1000, 1000-1500 and 3000-3500 spacing intervals are more likely to be the spacing between abnormal palindromes.

3.4 Patterns indicated by Gamma Distribution

We know that distances between the hits that are two apparatus follows a Gamma distribution with parameters 2 and λ where λ is the same λ in the previous Exponential distribution. If the space between the hits that are two apparatus is a Gamma distribution, it will also be easier to identify the abnormal occurrence of palindromes. In order to examine whether or not it is a Gamma distribution, we repeat the steps that we did in 3.3 and recalculate the spacing. In Table 6, it shows the new observed counts and expected counts.

Table 6: Comparison of Observed Counts of Spacing and Expected Counts

Spacing Interval	Observed counts	Expected counts
[0,500]	221	54.299150
(500,1000]	25	52.783878
(1000,1500]	23	24.959801
(1500,2000]	17	9.745770
(2000,2500]	2	3.475252
(2500,3000]	3	1.175099
(3000,6600]	4	0.383691

The histogram generated by our data indicates that the density of the space between two consecutive palindromes doesn't appear to be a Gamma distribution. To verify this, we run a Chi-square test based on Table 6. The null hypothesis of this Chi-square test is that the space between two consecutive pairs of palindromes follows a Gamma distribution. The formula of Chi-square is $X^2 = \sum_{i=1}^{n} (Oi - Ei)^2 / Ei$, in which Ei corresponds to Expected Counts in the table and Oi corresponds to Observed Counts in Table 6. The results based on our data are $X^2 = 569.50149$ and p-value=8.8159e-120. Hence, at a 5% significance level, we can reject the null hypothesis that the space between a pair of consecutive palindromes is a Gamma distribution.

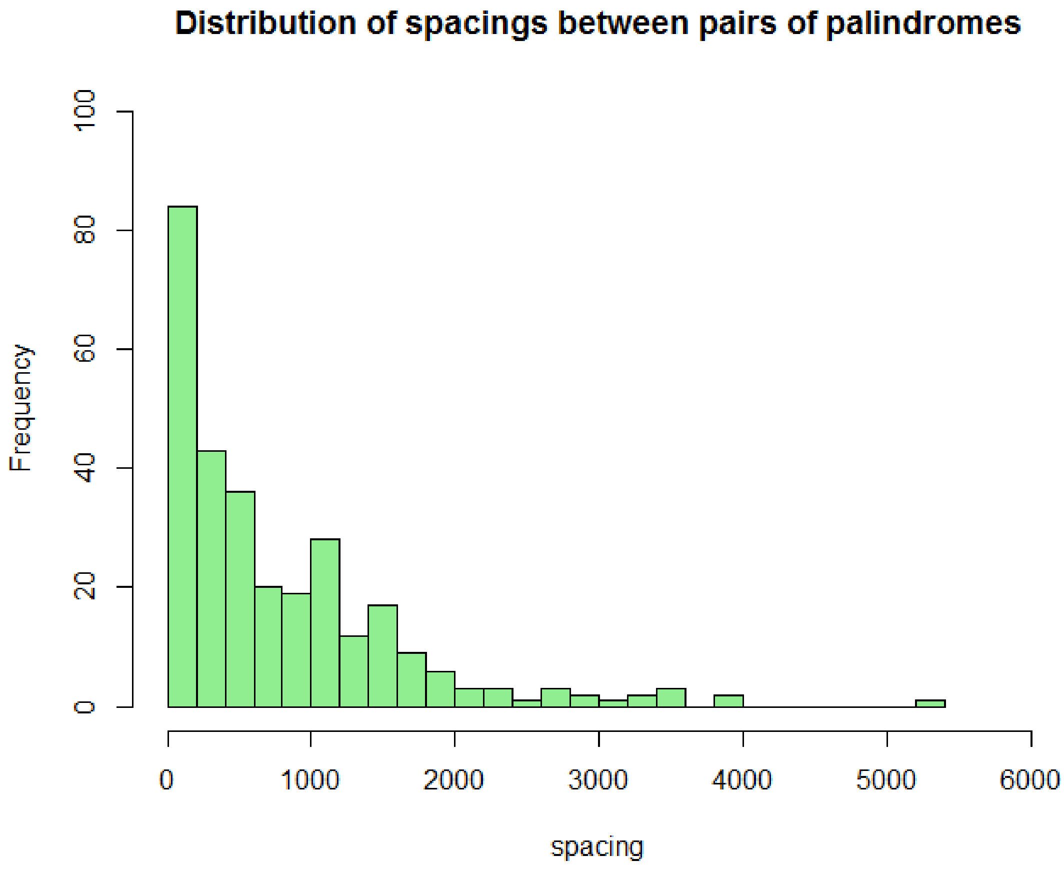

Figure 13: Histogram for Testing consecutive Pair Palindromes spacing

The discrepancy between observed values and theoretical values can be illustrated by the standardized residual plot.

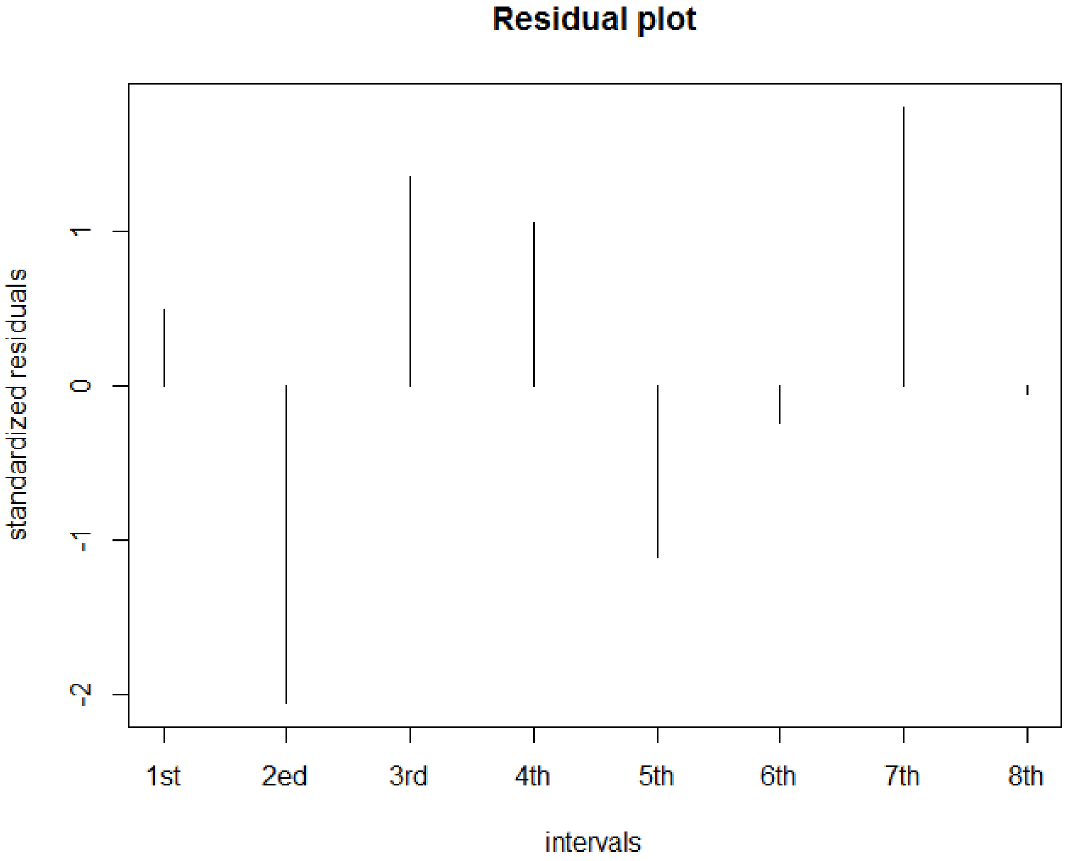

Figure 14: Residual plot for Testing Consecutive Pair Palindromes spacing

Since we rejected our null hypothesis that the distribution is a Gamma distribution, the expected counts we calculated are unreliable. Therefore, we decided to just use the residual plot of the exponential distribution to analyze the occurrence of abnormal palindromes.

4 Conclusion

4.1 How to Identify Clusters of Palindromes?

First of all, we divided the whole DNA into 40, 50, 60 and 70 region under an uniform distribution. Then we generate a set of 296 pseudo random number from a interval of 0 to 229354. After that, we plot several histograms to compare the difference between these different regions. By viewing the histograms, we find that the number of clusters changes depends on the length of the region of our data changes.However, there are a few intervals that always contain significantly more palindromes than other intervals no matter how we changed the interval length. These irregular intervals are: [52423,55700],[75359,78635],[91741,97475],[164370,168192]and [189217,194950]. We suggest biologists put more attention on the palindromes within these five base pair intervals, since the number of palindromes in these five intervals are apparently more than should it follow a Uniform distribution.

4.2 Eliminate Random Chance of Occurrence of Palindromes

According to the analyses in part 3, we make a conclusion that we should expect the palindrome cluster containing 7 palindromes, and the palindromes with 500-1000, 1000-1500 or 3000-3500 spacing distance are more likely to be the palindrome in the site of the cluster. Therefore, the cluster is not a chance occurrence site but a potential replication site, and the cluster has a property that it containing about 7 palindromes. In order to find the cluster, we can do two ways. First, we can cut the DNA into several equal size pieces and just find the region that have the biggest number of palindromes. The second way is to test the distance between two consecutive palindromes, and if they are fit in 500-1500 or 3000-3500 distance intervals, we should then take a look at the palindromes closed to them. If there is a big cluster, then we just find the replication site.

5 Theory

5.1 Poisson Distribution

Poisson distribution is used to estimate the probability of an event happen within a certain interval. Before adopting this theory to calculate the probability, we need to make sure that the process is homogeneous Poisson process. Thus, the process should satisfy two conditions. First of all, each event must be independent. Second, the probability of an event occurs in a short interval must equal to the probability of an event occurs in a long interval.
The formula of Poisson distribution is given by:
$Pr(\text{k points in a unit interval}) = \frac{\lambda^k}{k!} * e^{-\lambda}$
We use the formula to calculate the probability that counts of the number of points in a unit interval. By using the method of moments or maximum likelihood method, we estimate the rate λ. By adopting some values in formula, we could get several probabilities of each region we set.

5.2 Uniform Distribution

Uniform distribution is a statistical distribution that every possible outcome has an equal probability to occur. The formula of uniform distribution is
$f(x) = \frac{1}{b-a}$ for $a \leq x \leq b$
Under the uniform random scatter, the 296 palindromes in CMV DNA are similar as 296 independent observations from a uniform distribution. For example, if we split DNA as 10 equal sub-intervals,we would expect 1/10 of the palindromes exist in each interval.

5.3 Exponential Distribution

The definition of exponential distribution with parameter λ is: $f(x) = \begin{cases} \lambda * e^{-\lambda * x} \\ 0 \end{cases}$
In our research, we find that distances between successive hits follows an Exponential distribution.

5.4 Gamma Distribution

By the definition of gamma distribution, we get
$f(x) = \frac{(\frac{x-\mu}{\beta})^{\gamma-1} * exp(-\frac{x-\mu}{\beta})}{\beta \tau(\gamma)}$

5.5 Chi-Squared Goodness of Fit Test

Chi-squared test is a statistic measure that to test how well the observed data fits the expected distribution of data. The test is applied to analyze categorical data. Thus, we have to counted our data and transferred them into categories. Then, we get a data table with meaningful information to apply Chi-square test. The Chi-square formula is given by
$X^2 = \sum_{i=1}^{n} (Oi - Ei)^2 / Ei$
where O represents the Observed frequency and E represents the Expected frequency.
By using this formula, we can calculate the Chi-square value.If the test statistic we find is large,then we can conclude that the observed and expected value are not close and the model we apply is a poor fit to the data.