Math 289C HW3: Suggestions for Searching for the Origin of Replication of CMV

Mitesh Gadgil, Saurabh Kulkarni, Kyle Kole

February 26, 2016

Contents

1 Introduction

1.1 Background

The Human Cytomegalovirus (CMV) is a potentially life threatening disease. Those with suppressed or a deficient immune system are at most risk of fatality with CMV. Infection rates vay by geography, typically ranging around 30%-80% of the population. Children under the age of 5 are particularly susceptible, with infection rates around 10%-15%. CMV is difficult to detect because the virus remains dormant until a critical mass is achieved through reproduction. In order to combat the virus, virologists have decided that if the origin of replication may be isolated, the reproductive cycle of the virus may be interrupted, thereby the virus will remain perpetually dormant. In this report, we assume basic knowledge of DNA.

1.2 Data

Two viruses, Herpes and Epstein-Barr virus, of the same family have been studied in order to determine the best method to isolate the origin of replication. Both viruses contain one or more palindromes in its DNA sequence. Understanding the process by which DNA is replicated, these special sequences may indicate the origin of replication for these viruses. In particular, complementary palindromes are interesting because the latter half of the palindrome is exactly written in the complementary base pairs to the first half of the palindrome. Data published by Chee et. al (1990) on CMV indicates the locations of palindromes in the CMV DNA sequence. The CMV DNA is 229,354 base pairs long, and a study published by Leung et. al (1991), which implemented algorithms to search for different patterns, indicate the longest palindromes of 18 base pairs occur at locations 14719, 75812, 90763, 173893. Only palindromes of length 10-18 base pairs were counted, totaling 296.

2 Objective

We are seeking to confirm that certain clusters of palindromes are statistically different from the other clusters. In other words, certain clusters of palindromes that do not occur by random chance given the distribution of the clusters may signify the location of the origin of replication. The researchers working on the CMV should begin their search within these clusters. Doing this analysis will save researchers time and money in the search for the origin of replication.

3 Statistical Theory

We look to a few statistical tools in order to perform the data analysis. First, it is important to distinguish if the distribution of clusters of palindromes is indeed statistically significant. In other words, we must ascertain that clusters occur in a fashion not purely out of chance. We first have to generate random data to determine which distribution our data best fits. We start by using the uniform distribution and the homogenous Poisson process. The advantages of the homogenous Poisson process include an easy estimate for parameter lambda, a distribution which counts natural random processes, and a parameter which does not depend on the location of the distribution. The expected value of the poisson process is lambda, so lambda captures the underlying rate of number of hits, or underlying rate of number of points along an interval. Lambda may be approximated using the sample data by calculating the sample average; this estimator is the same using method of moments and maximum likelihood estimation. Note that the probability for k points in an interval is

$$P(\text{k points in a unit interval}) = \frac{\lambda^k e^{-\lambda}}{k!}, \text{ for k = 0,1,2,...}$$

For spacing between palindromes, we look to see if an alternate spacing model, $(x[i+2] - x[i])$, may be modeled using the Gamma Distribution. The gamma distribution has two positive real numbered parameters that may be parametrized in three different manners:

1. With shape parameter k and scale parameter θ
2. With shape parameter $\alpha = k$ and an inverse scale parameter $\beta = \frac{1}{\theta}$, called the rate parameter
3. With shape parameter k and mean parameter $\mu = \frac{k}{\beta}$

Specifying in terms of k and θ, we have the density function to be

$$\frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$$

Notice how the exponential distiribution and the chi-squared distribution are special cases of the gamma distribution, with $k = 1$ and $\theta = 1$, and $k = \frac{N}{2}$ and $\theta = 2\sigma^2$, respectively. The gamma function is great for modeling random waiting times. Hence, we want to see if spacing between palindromes may be modeled using the gamma distribution.

After simulating data, we must determine which distribution the data best fits. Hence, the chi-squared goodness of fit test. Upon partitioning the simulated data into non-overlapping intervals, the chi-squared goodness of fit test calculates the standardized deviation of the expected value of that interval based on a distribution to the actual average value of the simulated data in that interval. This calculation is performed for all intervals and summed, then compared to the quantiles of a chi-squared distribution with degrees of freedom equal to

$$(\text{Number of intervals}) - 1 - (\text{Number of estimators})$$

The null hypothesis is the data fits a specific distribution versus the alternative that the data does not. For large values of the test statistic, we can expect to reject the null hypothesis as the data deviates from the distribution with a large magnitude.

4 Part A: Organizing and Evaluating the Data

We know the following about the data:

1. **Total length of DNA (N):** 229,354 base pairs
2. **Number of Palindromes detected (k):** 296
3. Data contains location of each complementary palindrome (1 to 229354). Data is numerical and discrete in type (discrete in nature as you cannot have a fraction of base pair)

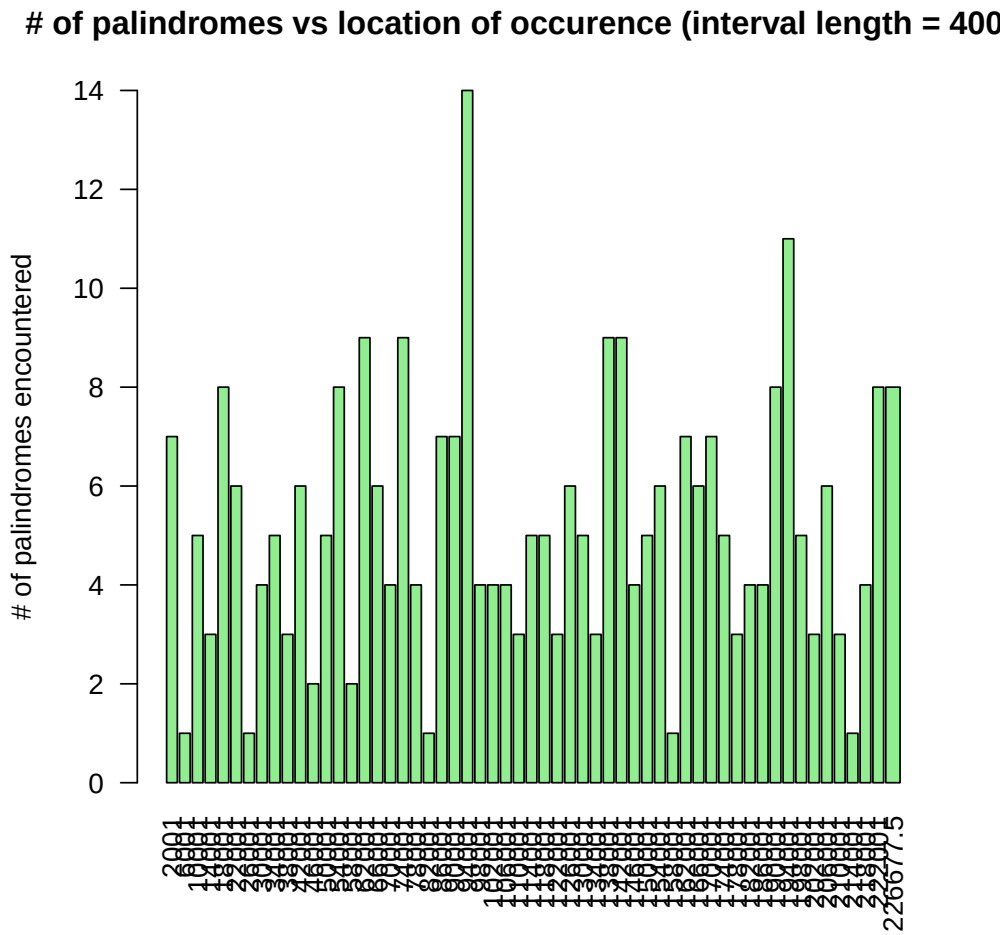
4.1 Test A1: Histogram plot of cluster locations

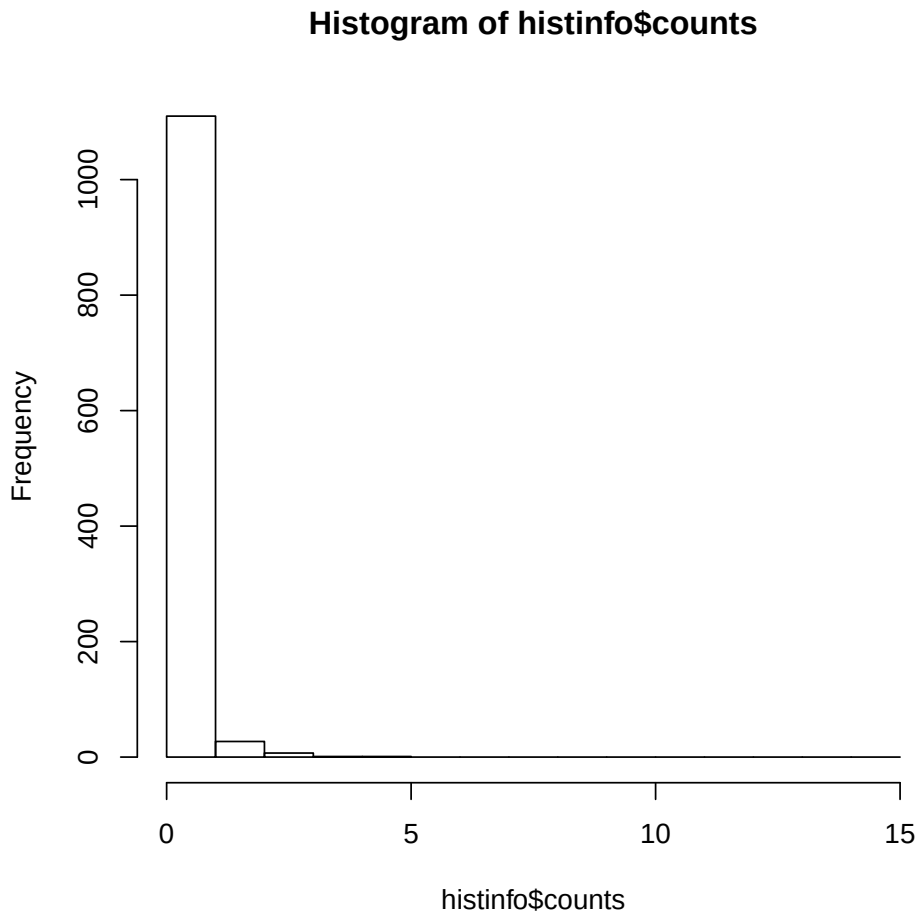
1. To count the number of palindromes per interval
2. To count the number intervals that contain a certain number of palindrome

We segment the DNA chain into intervals of base pairs and count the number of palindromes found in each interval. This can be done by taking a histogram of the data to see around what locations palindrome clusters are located. We can also count the number of intervals containing 0 palindromes, 1 palindrome etc upto 14 palindromes to study outliers

4.1.1 Objective of Test:

This test will tell us where the clusters of DNA are prominent and at these locations the DNA code to replicate may be located


# of palindromes vs location of occurence (interval length = 400
# of palindromes encountered
0
2
4
6
8
10
12
14



4.2 Analysis

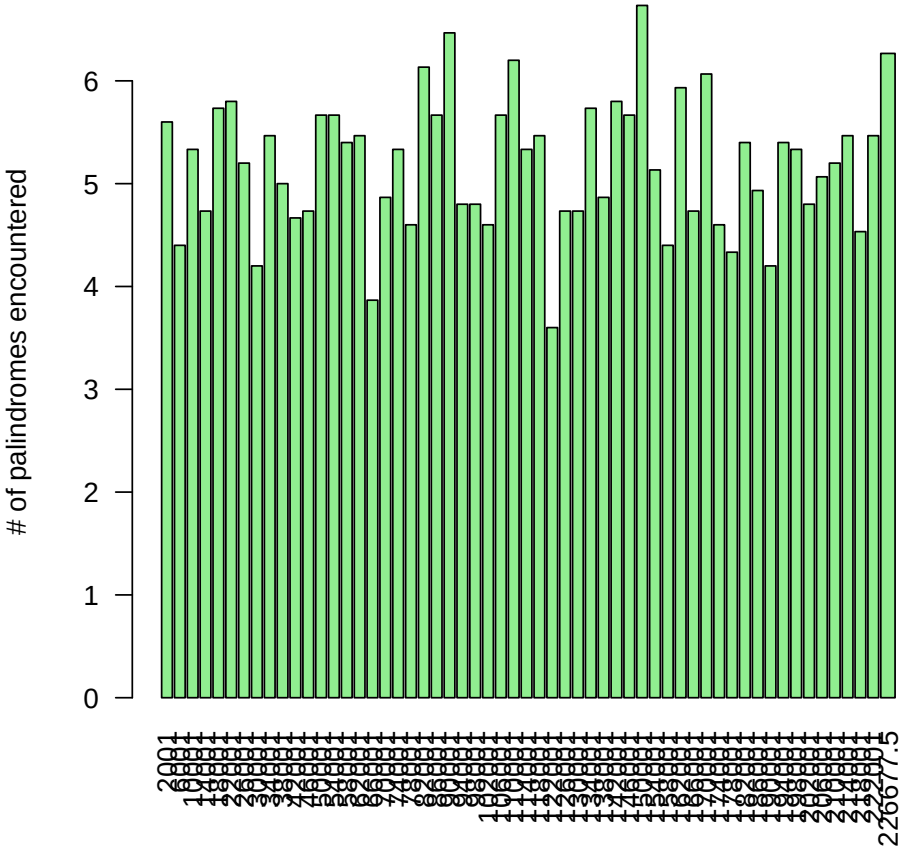
- There are prominent peaks at the $93,000^{th}$ and $195,000^{th}$ pairs of DNA. Hence, what we definitely say is that there is certain information in the $93,000^{th}$ and $195,000^{th}$ locations to further explore.
- We vary interval size to see the change in the nature of the plot. We observe that choosing a smaller intervals yields many intervals containing no palindromes, while intervals that are very large contain so many palindromes as to infer any meaningful or logical interpretations of the occurrences of palindromes.
- This cluster can be either a chance occurrence or may contain information about a potential replication site. The tests we have performed later will tell us more information on this. In order to conclusively say the location of palindrome clusters is not random, we do the following test.

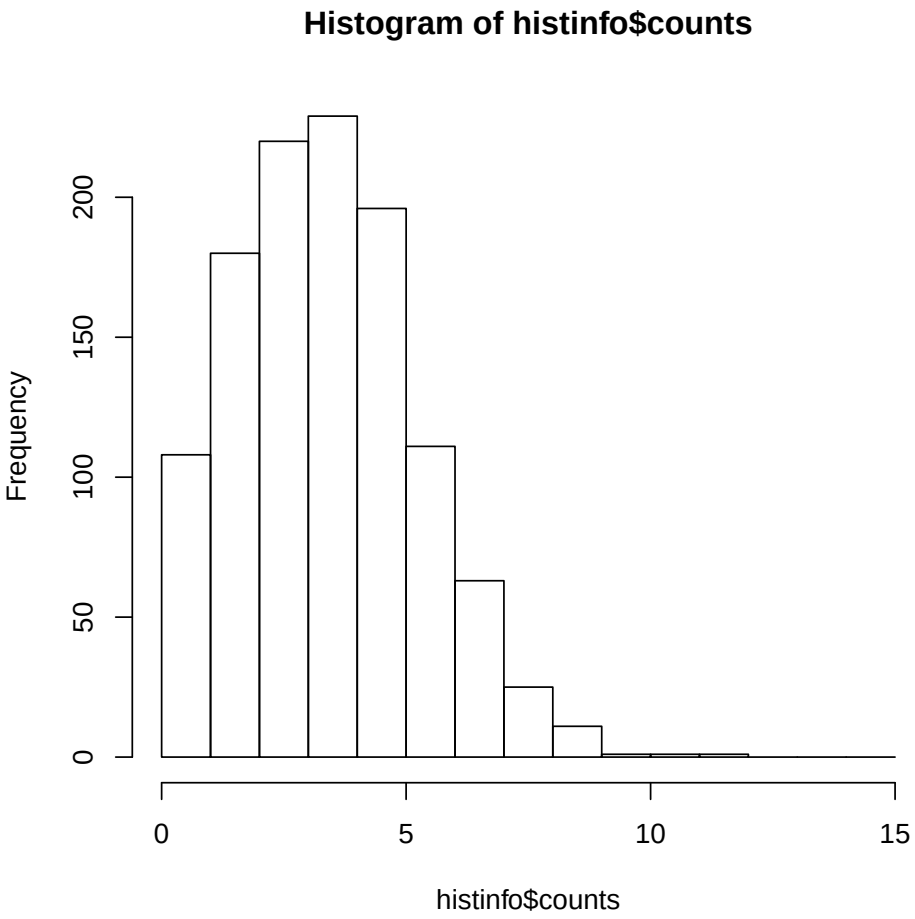
4.3 Test A2

We first generate 296 pseudorandom numbers as locations for palindromes from 1 to 229354. We study the histogram for random data.

4.3.1 Objective of Test:

This is to observe and compare, the kind of clusters generated from Random Data and from given data.


# of palindromes vs location of occurence (interval length = 400
# of palindromes encountered
0
1
2
3
4
5
6



4.4 Analysis

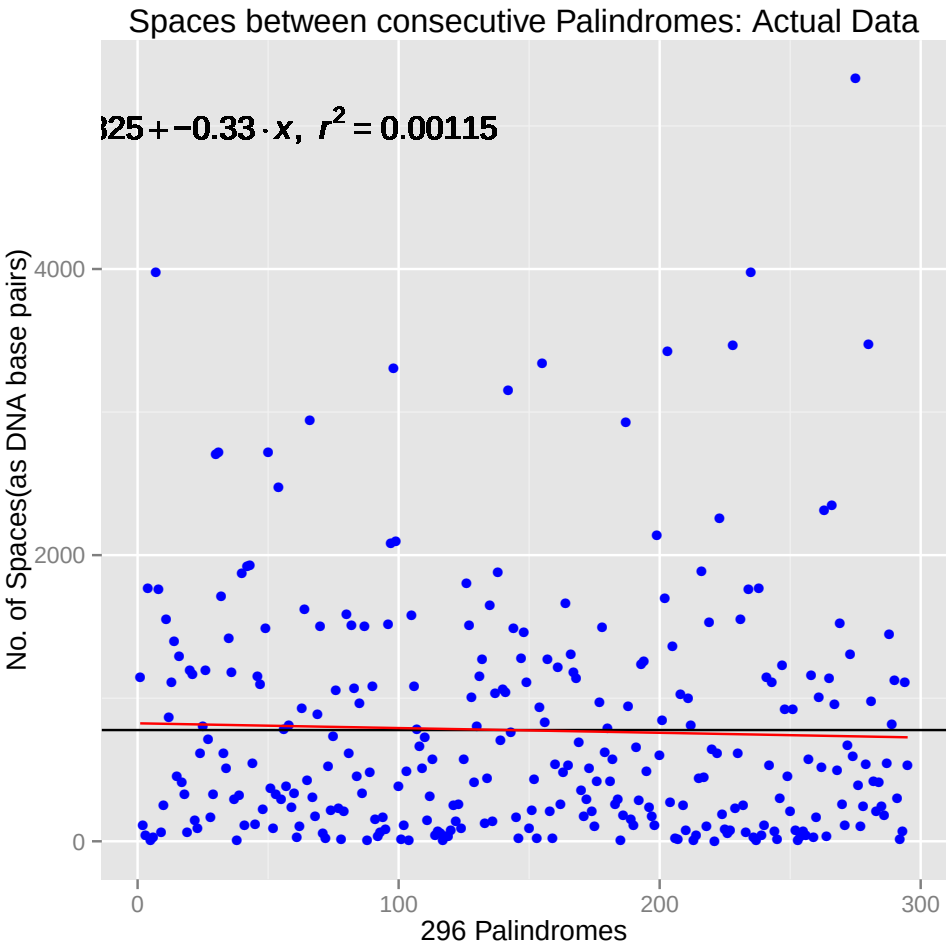
- By comparing histograms of the actual palindromes to histograms based on randomly generated numbers we can see that the random sets of numbers present no pattern of clusters at any given point, no matter what size intervals we use.
- The plot that shows the number of interval that contain a certain number of palindrome also appears random and not like the one we generated from our own data
- Therefore it would seem logical to deduce that the peaks we observe on the DNA are atypical and may be a site to further study for replication code.

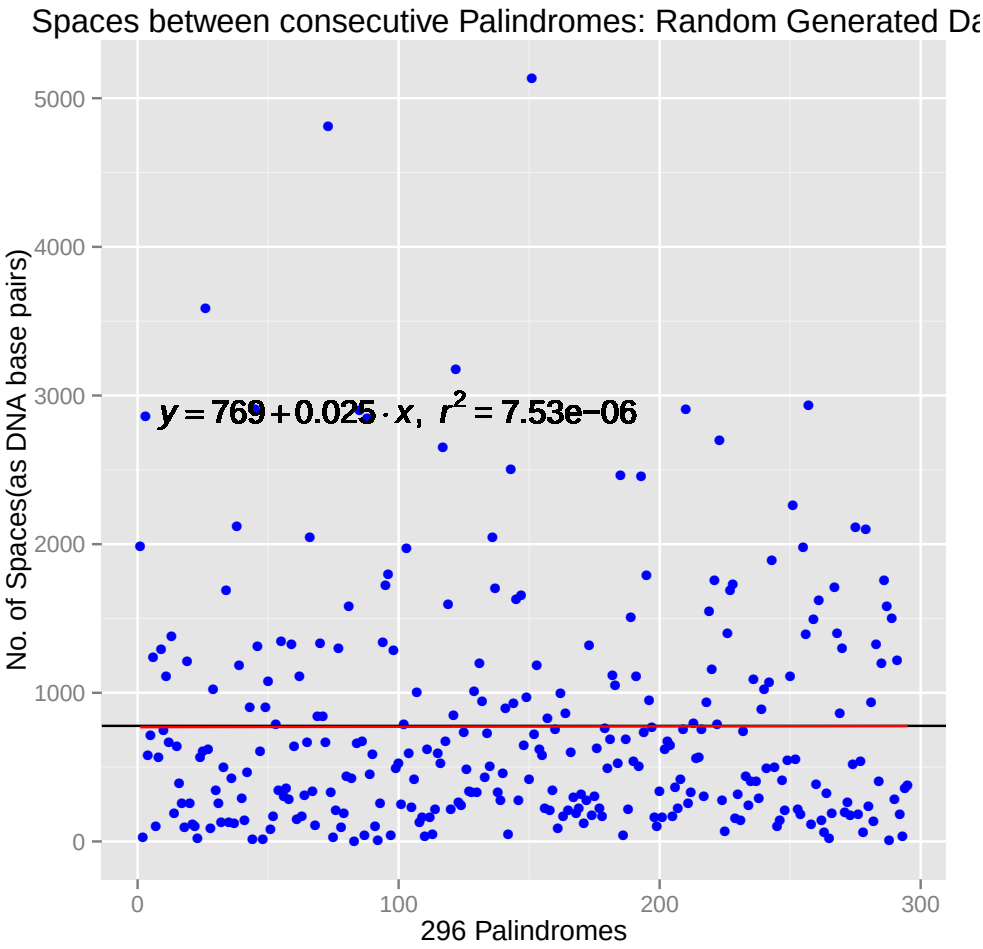
4.5 Test A3

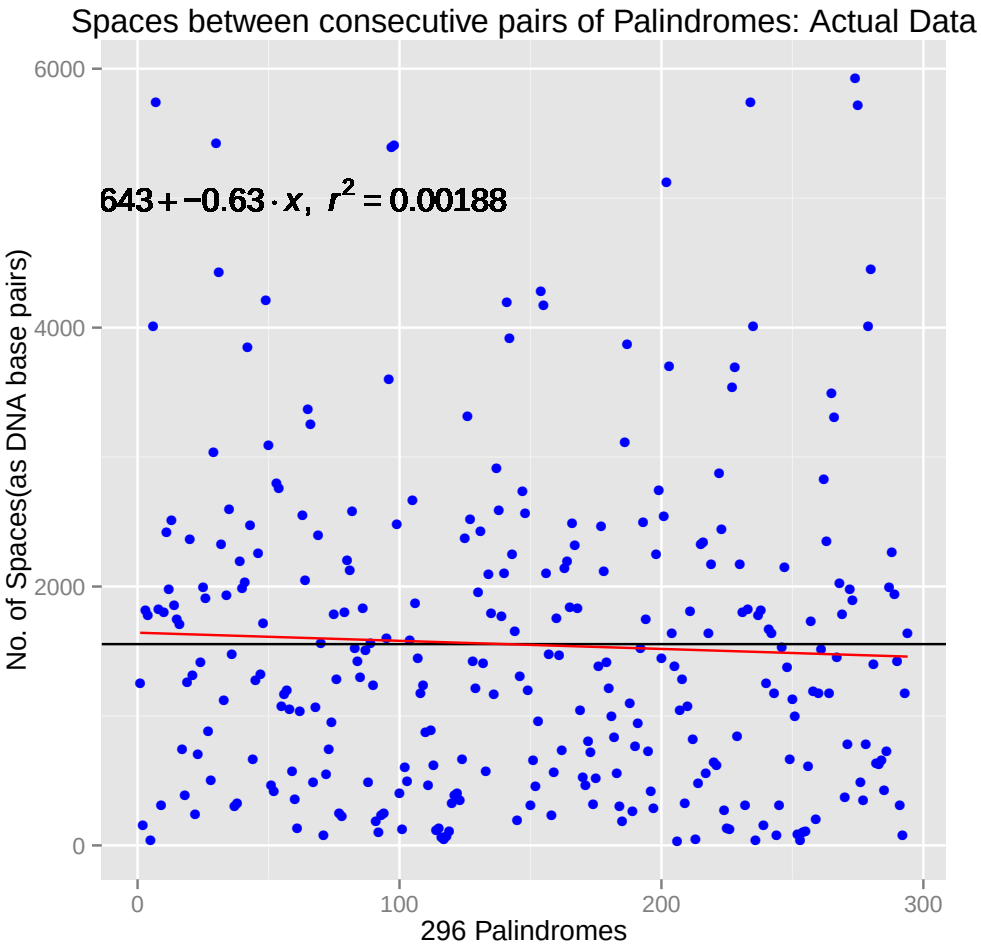
We now examine the spacing between consecutive palindromes, every other palindrome, and every other other palindrome for the given data and a randomly generated data.

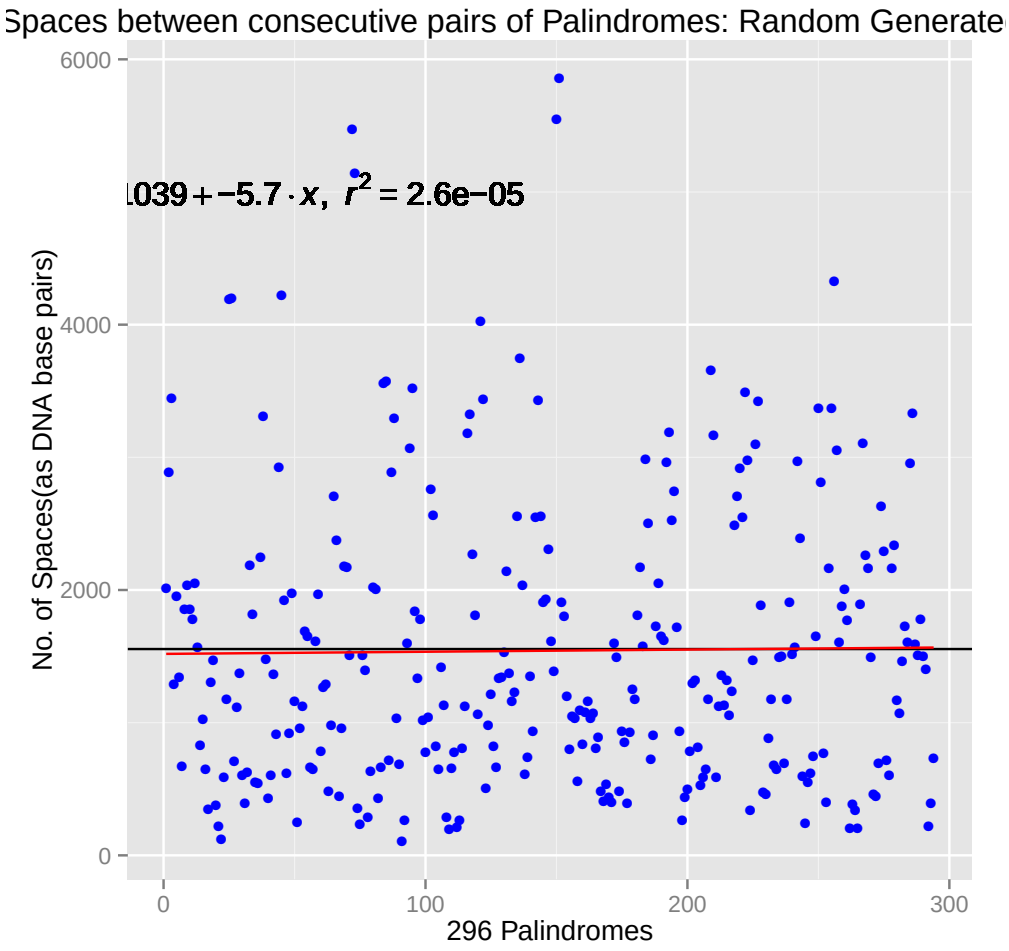
4.5.1 Objective of Test:

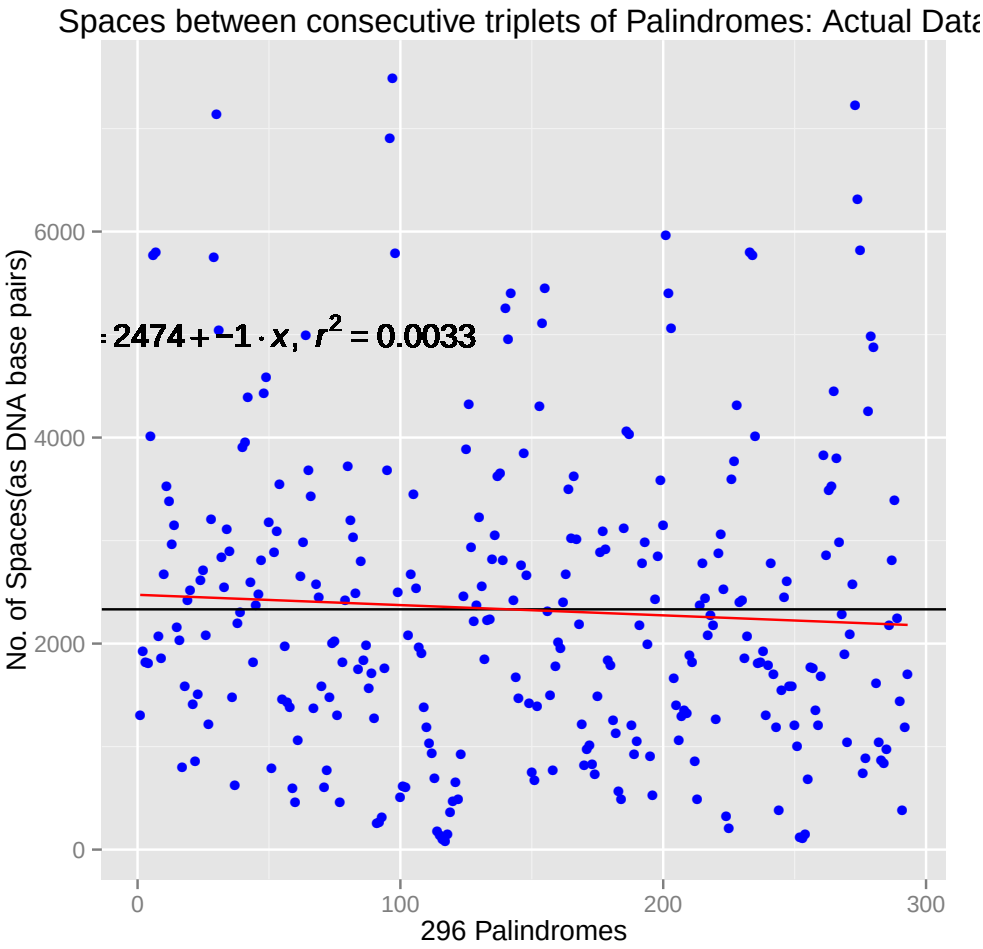
To observe if there spacing of palindromes follow a certain predictable distribution or not.

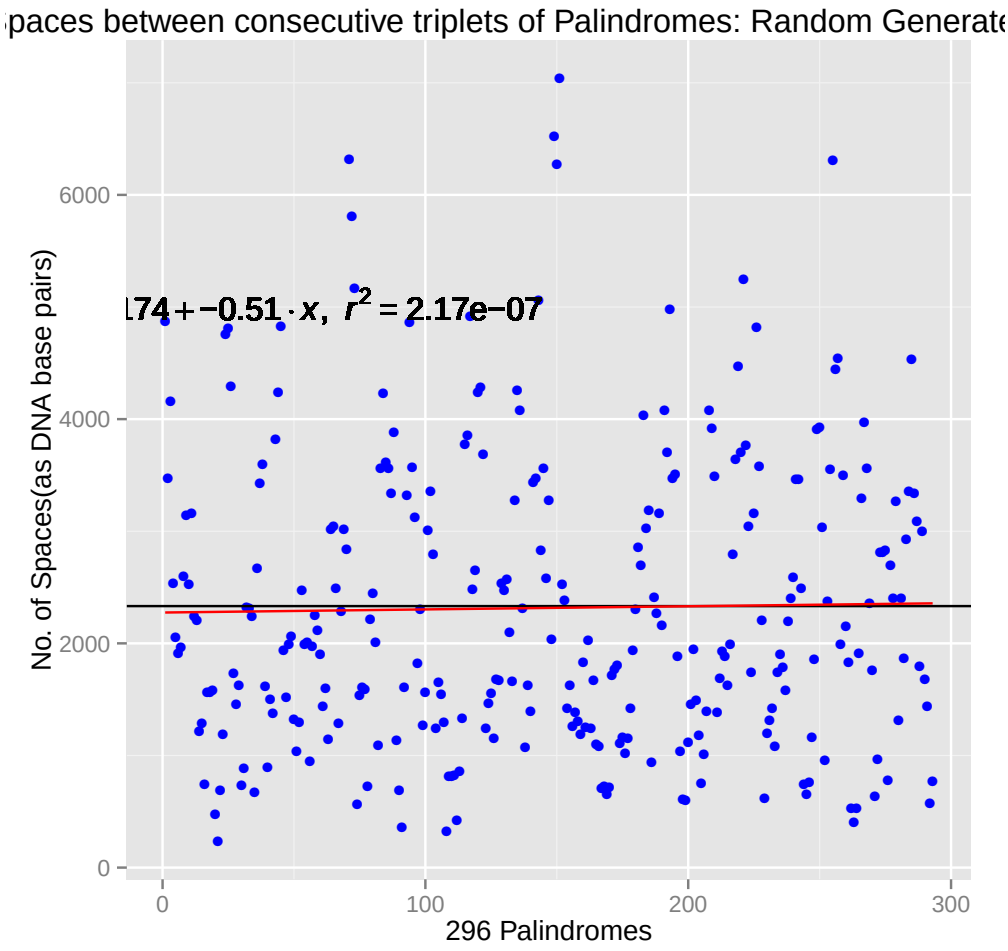

Spaces between consecutive Palindromes: Actual Data
325 + −0.33 · x, r² = 0.00115
No. of Spaces(as DNA base pairs)
4000
2000
0
0
100
200
300
296 Palindromes


Spaces between consecutive Palindromes: Random Generated Da
No. of Spaces(as DNA base pairs)
5000
4000
3000
2000
1000
0
y = 769 + 0.025 · x, r² = 7.53e−06
0
100
200
300
296 Palindromes


Spaces between consecutive pairs of Palindromes: Actual Data
643 + −0.63 · x, r² = 0.00188
No. of Spaces(as DNA base pairs)
6000
4000
2000
0
0
100
200
300
296 Palindromes


Spaces between consecutive pairs of Palindromes: Random Generate
1039 + −5.7 · x, r² = 2.6e−05
No. of Spaces(as DNA base pairs)
6000
4000
2000
0
0
100
200
300
296 Palindromes


Spaces between consecutive triplets of Palindromes: Actual Data
No. of Spaces(as DNA base pairs)
6000
4000
2000
0
2474 + −1 · x, r² = 0.0033
0
100
200
300
296 Palindromes



5 Part B: Statistical Testing

We have inferred the following from Part A analysis: the behavior of number of palindromes per interval follows a certain deterministic distribution and is not truly random in nature. In order to see whether this behavior can be represented by a certain distribution or not, we use hypothesis testing. We will check if the Poisson Distribution or the Uniform Distribution are good models for estimating the number of palindromes per interval. However, we could not infer any conclusive remark about the spacing between consecutive palindromes of the given data. We will check if the Exponential Distribution is able to model the spacings or not. We will also check if the every other other spacing $(x[i+2]-x[i])$ can be modeled by the Gamma Distribution.

5.1 Chi-Squared Goodness of Fit Test

The chi-square goodness of fit test is appropriate when the following conditions are met:

- Random sampling method
- Variable is categorical
- The expected value of the number of samples in each level of the variable is at least 5

This approach consists of four steps:

1. Hypothesis statement (hypothesize which distribution data fits)
2. Formulate analysis plan
3. Analyze sample data
4. Interpretation of results

5.2 Hypothesis Testing

Every hypothesis test requires the analyst to state a null hypothesis (H_0) and an alternative hypothesis (H_a). The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false; and vice versa. For a chi-square goodness of fit test, the hypotheses takes the following form:

H_0: Data is consistent with a specific distribution vs. H_a: Data is inconsistent

Typically, the null hypothesis (H_0) specifies the proportion of observations at each level of the categorical variable. The alternative hypothesis (H_a) is that at least one of the specified proportions is not true.

5.3 Analysis

The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements:

- Significance level: Researchers often choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 may be used.
- Test method: using the chi-square goodness of fit test to determine whether observed sample frequencies differ significantly from expected frequencies specified in the null hypothesis.

5.3.1 Analysis of Sample Data

Degrees of freedom: The degrees of freedom $DoF = k - e - 1$: where k is the number of intervals, e is the number of parameters estimated for the distribution.

Test statistic: The test statistic is a chi-squared random variable (χ^2) defined by the following equation: $\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$

where O_i is the observed frequency count for the i^{th} level of the categorical variable, and E_i is the expected frequency count for the i^{th} level of the categorical variable.

P-value: The P-value is the probability of observing a sample statistic as the test statistic.

Now we will use this technique to estimate the Chi-Squared Goodness of Fit for the following: the **objective** of each test is to develop a low error model for the specified graphs

1. Uniform Distribution
    - Number of palindromes per interval for our given data
2. Poisson Distribution
    - Number of palindromes per interval for the given data
    - Number of palindromes per interval for random 296 samples generated from the interval 1 to 229,354.
3. Exponential Distrubtion
    - Distances between successive spacing of palindromes for given data
4. Gamma distribution
    - Distances between every other other palindrome for the given data.

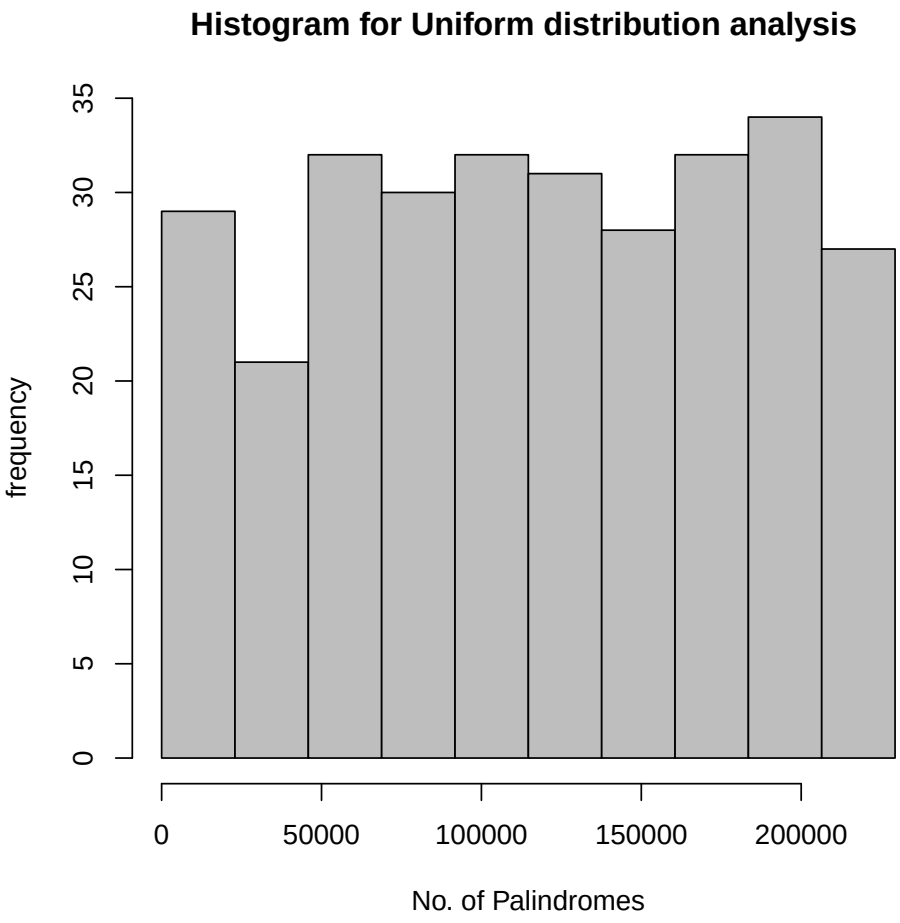
5.3.2 Test B1

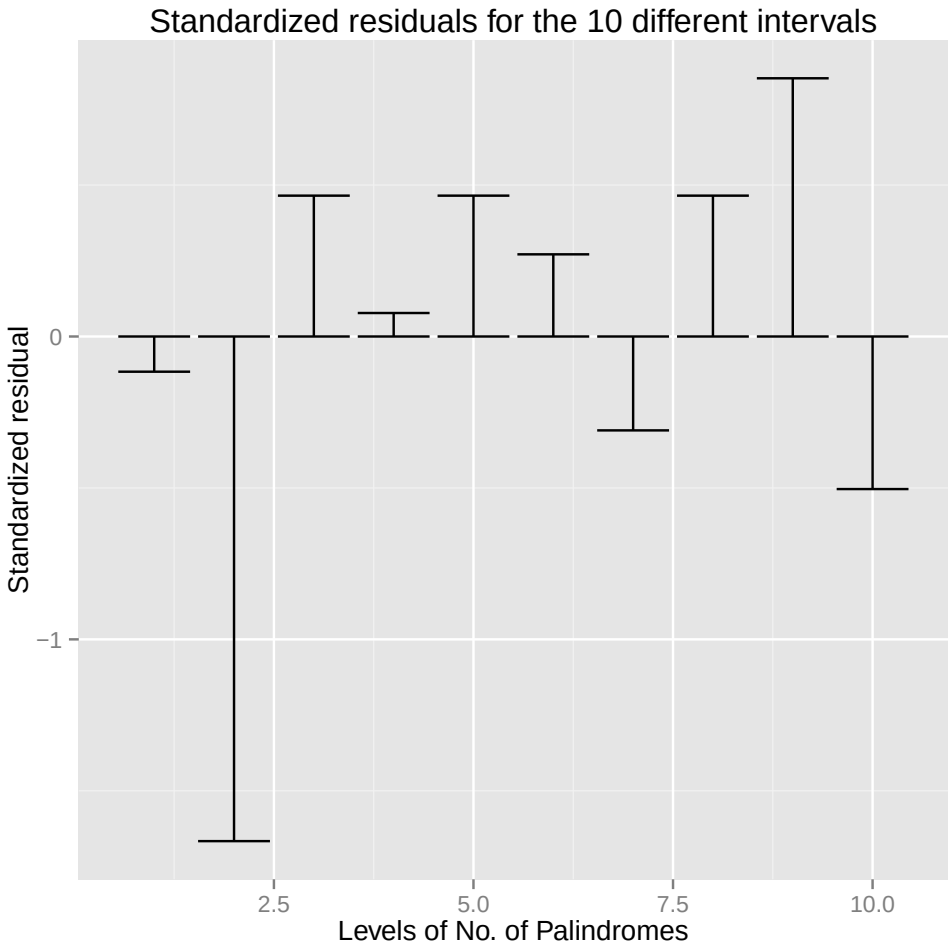
To test if the Uniform Distribution is a good fit for the number of palindromes per interval for our given data.

1. Number of Intervals: 10
2. Length of each interval: $\frac{229354}{10} \approx 22935$
3. Expected number of palindromes per interval: $\frac{296}{10} = 29.6$

We split the given data into 10 successive intervals of 22935 length each and find out the frequency of palindromes occurring in the given interval. We then find the test statistic χ^2 for the test. We also plot the standard residuals for this test using error-bars.

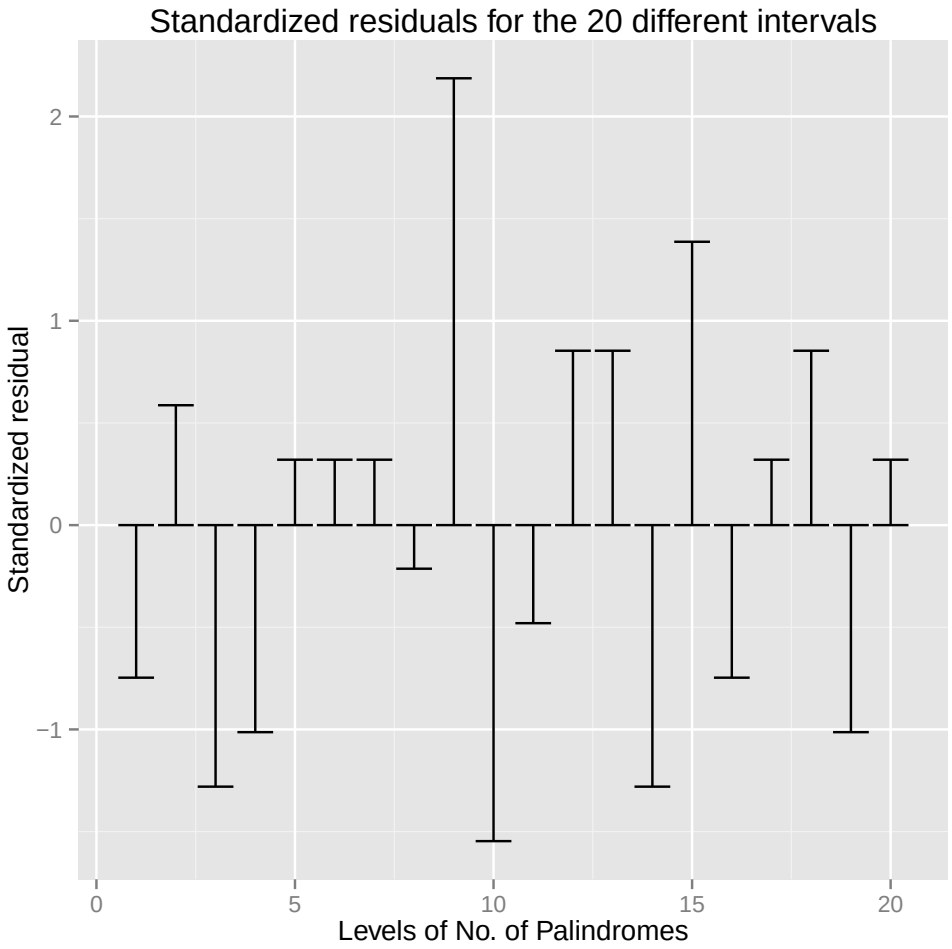
We further perform the same analysis for 20 intervals and notice that the chi square test is about 53%,which is very low.





For the Uniform Distribution over 10 intervals Goodness-of-Fit test: The value of the chi-squared statistic is 4.135, and the p-value is 0.902. The standard residuals are as follows:

```
 [1] -0.110 -1.581  0.441  0.074  0.441  0.257 -0.294  0.441  0.809
[10] -0.478
```



For the Uniform Distribution over 20 intervals Goodness-of-Fit test: The value of the chi-squared statistic is 17.919, and the p-value is 0.528. The standard residuals are as follows:

```
 [1] -0.728  0.572 -1.248 -0.988  0.312  0.312  0.312 -0.208  2.131
[10] -1.508 -0.468  0.832  0.832 -1.248  1.352 -0.728  0.312  0.832
[19] -0.988  0.312
```

Analysis We notice that the p value of the uniform distribution is 90%. We want a more accurate model which will fit the distribution better.

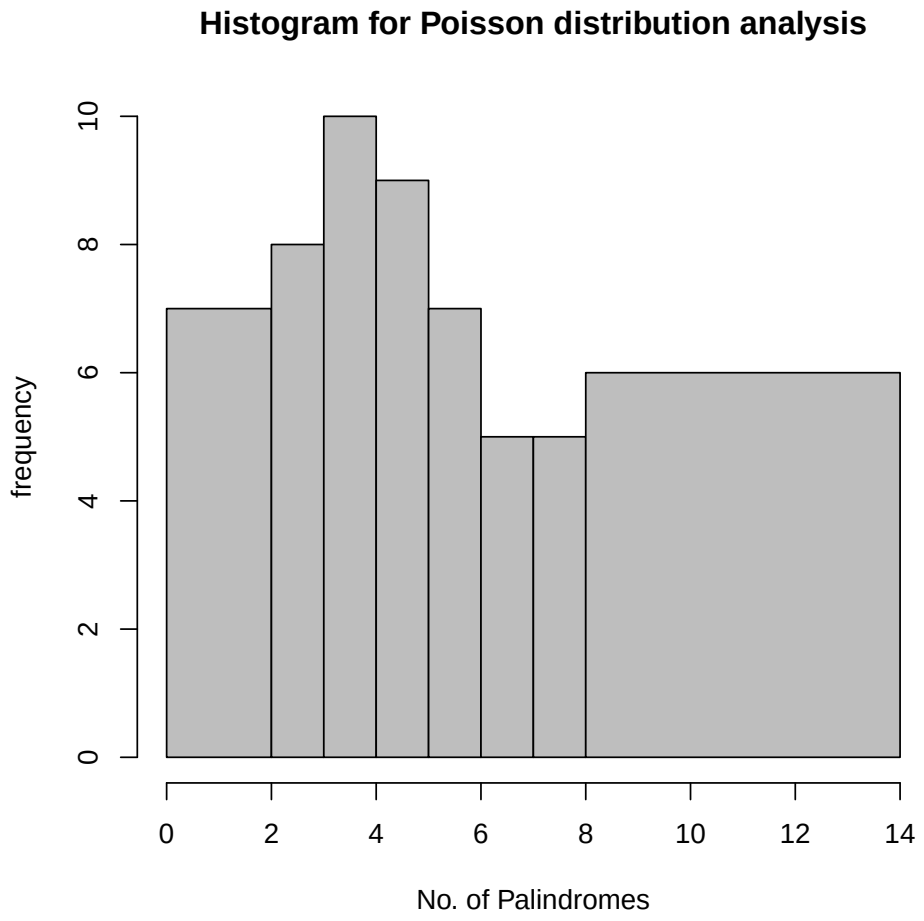
5.3.3 Test B2

To test if Poisson Distribution is a good fit for number of Palindromes per Interval for our given data.

- Data is binned into following intervals: 0-2, 3, 4, 5, 6, 7, 8, 9-14 (8 intervals total). This binning is done to ensure that counts exceed 5 for *chisq.test()*, a built-in R function
- Choice of maximum likelihood estimator $\lambda = 296/8 = 37$
- *chisq.test()* function is used to find the p-value

We chose the best possible estimate for λ : the sample mean. The sample mean is an unbiased estimator for the Poisson Distribution. We have made the following assumptions before testing for Poisson:

1. *Homogeneity*: the rate at which palindromes occur does not change with location.
2. The number of points falling in separate regions are *independent*.
3. No two points can land in exactly the same point.



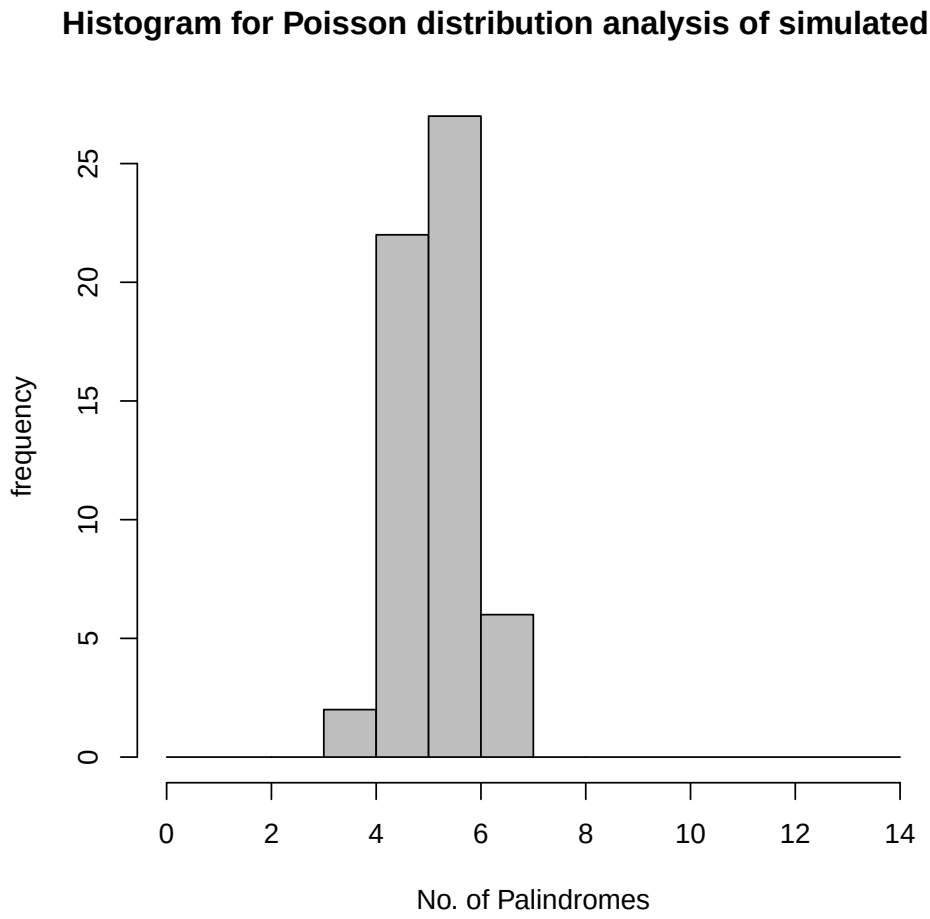
The Poisson Goodness-of-Fit test: Value of the chi-squared statistic is 3.211. The P-value is 0.865.

Analysis: Before we begin analyzing we must see an important point. There are 8 intervals here and we also estimated the mean using the data. Hence the degree of freedom for the Chi-squared test is actually $DoF = 8 - 1 - 1 = 6$. We used the estimated mean of the data to generate the probability vector which was then used as an input parameter for the *chisq.test()* function. But R still throws out $DoF = 7$. This is because R does not know that you estimated some parameter of the distribution. Hence, *Actual* $DoF = 6$ and not 7.

We see that we get a very high p-value for this distribution. The Chi-squared test essentially tells us that Poisson Distribution is a good fit for the number of intervals having a certain number of palindromes. Hence this further implies that the occurrence of palindromes was not random at all and follow a certain distribution, thus implying they play a role in origin of replication in the DNA. Hence we accept hypothesis H_0: the data follows the poisson distribution.

5.3.4 Test B3

To check for poisson distribution for a randomly generated Sample. We perform this test just to ensure that any random sample does not follow poisson.

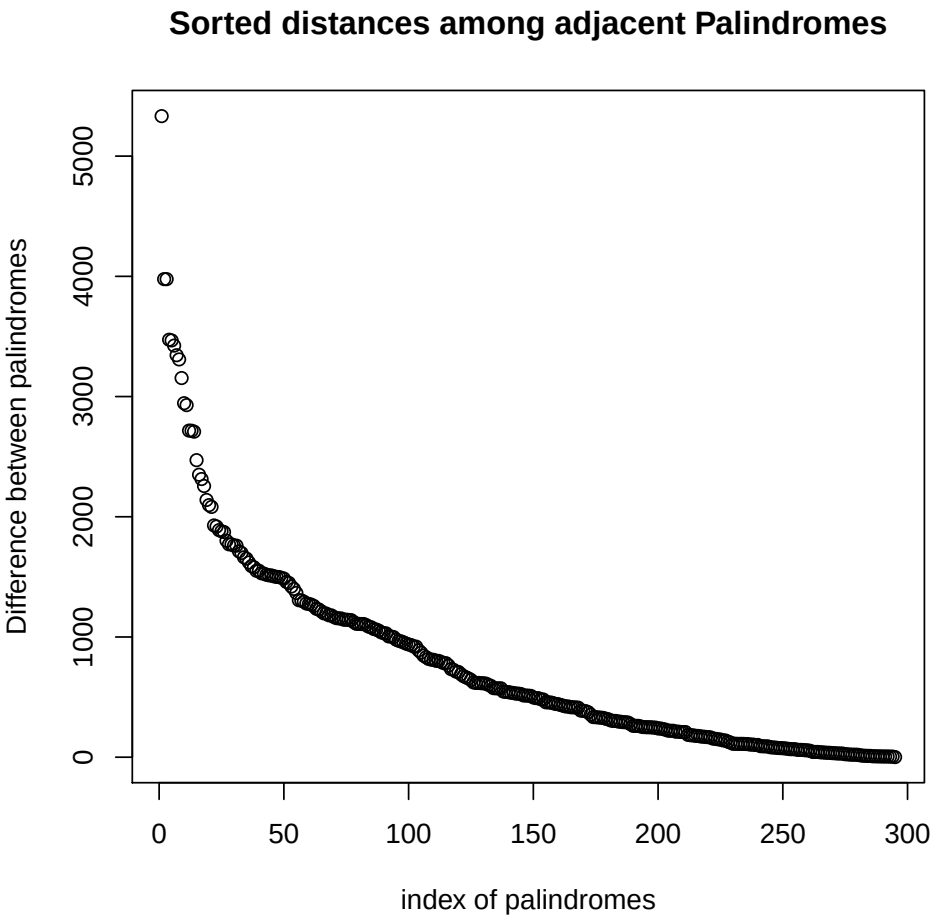


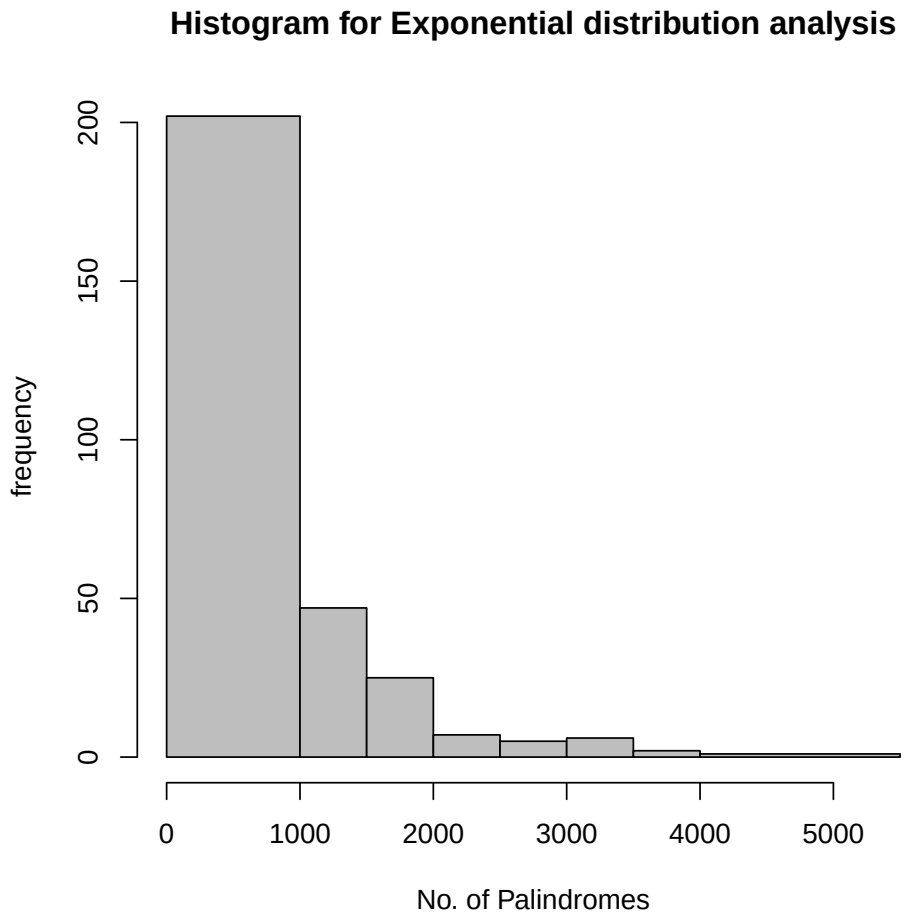
The Poisson Goodness-of-Fit test on simulated random data: The value of the chi-squared statistic is 118.862 and the p-value is 0.

Analysis: Here we observe the p-value to be much lower than what we got for the palindrome data. This implies that there is very less chance the palindrome data occurred randomly and it further strengthens our argument that palindrome data follows Poisson

5.3.5 Test B4

To check that distances between successive hits follows an Exponential distribution. In Part A, we plotted the scatter-plots for successive hits, but the graph was inconclusive, hence we see how good the exponential curve fits the spacing for $x[i+1] - x[i]$


Sorted distances among adjacent Palindromes
Difference between palindromes
0
1000
2000
3000
4000
5000
0
50
100
150
200
250
300
index of palindromes

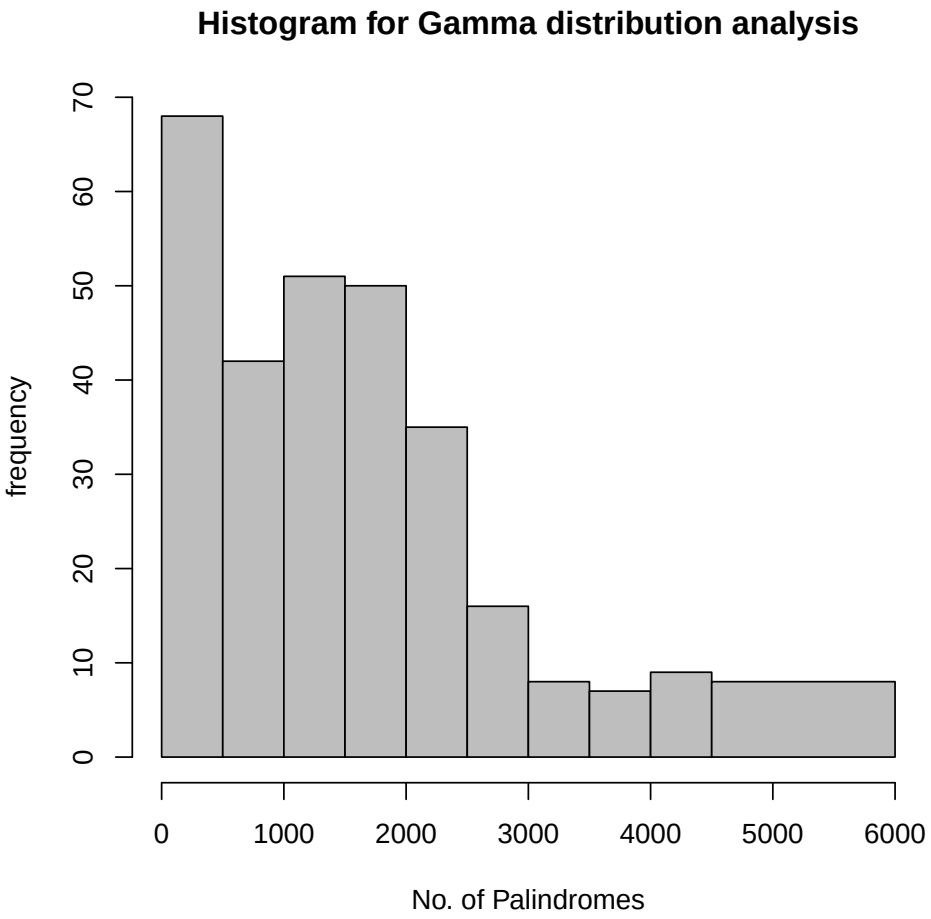


Exponential distribution Goodness-of-Fit test: The value of the chi-squared statistic is 8.153, the p-value is 0.319.

Analysis: We observe the the first two bins are the ones contributing to most error. But we notice that qualitatively this graph does follow the exponential curve to a decent approximation. The reason this makes sense is that Poisson and Exponential variables are related. Suppose that events (palindrome) occur in space (across DNA) according to a poisson process with parameter λ. So $X \sim Poisson(\lambda)$, and the probability that the next successive palindrome that is, until the first arrival is exponential with parameter λ. Hence we expected the graph of differences to follow a exponential curve.

5.3.6 Test B5

To check that Distances between the hits that are two apart, follows a Gamma distribution. The scatterplots in part A were inconclusive, hence we see how good the Gamma curve fits the spacing between points that are two distance apart (i.e. $x[i+2] - x[i]$)



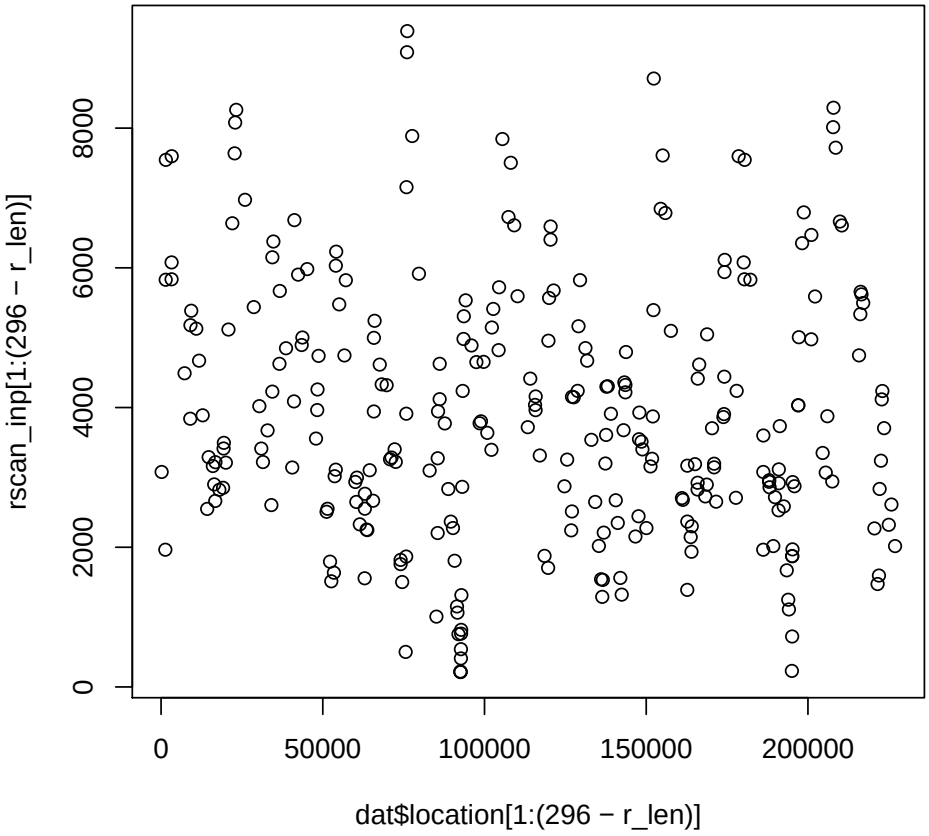
Gamma distribution Goodness-of-Fit test: The value of the chi-squared statistic is Inf, the p-value is 0.

Analysis: The reason alternate makes sense is that Poisson and Exponential variables are related. Suppose that events (palindrome) occur in space (across DNA) according to a Poisson process with parameter λ. So $X \sim Poisson(\lambda)$, then probability that the next k^{th} palindrome that is, until the k^{th} arrival after current palindrome is Gamma with parameter (k,λ). Hence we expected the graph of differences to follow a Gamma curve with (2,λ). The Gamma Distribution is a Good fit for this data.

6 Part C: Scan Statistics

Let $X[1], ..., X[n]$ be the order statistics of n i.i.d. uniformly distributed points. Let $Si = X[i+1] - X[i]$ be the consecutive spacing. Let $Ar(i) = Si + ... + Si + r - 1$ be the sum of the r adjoining spacings starting at X(i).

This Ar gives us the R-scan statistic. If Ar is too small, an unusual cluster is present. Ar will help determine which clusters are unlikely to occur by chance. We implemented the R-scan statistic to verify whether unusual clusters are indeed at 93000 and 195000 locations. We can clearly see from scatterplot below that the value of Ar for 93000 and 195000 locations is unusually low. Hence the unusual clusters must be located at these positions.



7 Conclusion and Recommendations

Hence we conclude that the best case to look for palindromes around these regions and these palindromes are related to the replication in the DNA

8 Citations

1. http://www.math.utep.edu/Faculty/mleung/presentation/nari0500/poster.pdf
2. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1197138/
3. http://www.cimt.plymouth.ac.uk/projects/mepres/alevel/fstats_ ch5.pdf
4. https://www.statisticssolutions.com/chi-square-goodness-of-fit-test/
5. http://www.pitt.edu/∼ super1/ResearchMethods/Ricci-distributions-en.pdf
6. https://www.inet.tu-berlin.de/fileadmin/fg234_ teaching/SS12/IM_ SS12/im12_ 05_ stats_ traceroute.pdf