CASE STUDY 3: SEARCH FOR THE UNUSUAL CLUSTER IN THE PALINDROMES

Luying Jiang A13513110
Siyao Mi A13806003
Nuo Chen A13709502
Jining Jin A13688539
Zhijian Hu A92070159

Introduction

The human cytomegalovirus (CMV) is a potentially life-threatening disease for people with suppressed or deficient immune system. For example, CMV infection is typically unnoticed in healthy people, but can be life-threatening for the immunocompromised, such as HIV-infected persons, organ transplant recipients, or newborn infants. To develop strategies for combating the virus, scientists study the way in which the virus replicates: 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, coded message made from a four-letter alphabet: A, C, G, T. DNA sequences contain many patterns, as the alphabet is small. Some of these patterns may flag important sites on the DNA, such as the origin of replication. Complementary palindrome is one type of pattern. In DNA, the letter A is complementary to T, and G is complementary to C, and complementary palindrome is a sequence of letters that reads in reverse as the complement of the forward sequence: GGGCATGCCC

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. This process can be very time consuming and expensive without leads on where to begin the search. 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.

In this lab we will search for unusual clusters of complementary palindromes. Particularly, we are investigating "How do we find clusters of palindromes? How do we determine whether a cluster is just a chance occurrence or a potential replication site?".

Data

In 1990, the DNA sequence of Cytomegalovirus (CMV), a virus, was published and Leung et al employed search algorithms to screen the sequence for many types of patterns. Under this circumstance, by ignoring the palindromes which are shorter than 10 letters, there are 296 palindromes were found, which are at least 10 letters long. And he found that 18 letters were the longest ones and occurred at the locations of 14719, 75812, 90763 and 173893. The CMV DNA is 229,354 letters long.

By segmenting the DNA chain into intervals of base pairs and count the number of palindromes found in each interval, we can group the data of the 296 palindromes. Through drawing the histograms, we can observe the clusters of palindromes appear in at least two locations, which around the 93,000th and 195,000th pairs of DNA. And then we can formulate the hypothesis about the clusters at these two locations. By comparing the histograms of the actual palindromes to the histograms of the random generated numbers, we can observe the random sets. Besides, by drawing the intervals with palindromes, we can observe that the palindrome present higher spikes of number of palindromes per intervals. And there exist few outliers of intervals no matter the length of the interval. But the intervals of random hits do not display outliers. In this way, the outliers on the DNA are atypical and it is worth examining for the replication code.

Data Limitation: ignoring the palindromes that are less than 10 letters.

Background

The basis of DNA come in four types: adenine, cytosine, guanine and thymine, or A,C,G,T for short. As the two strands of nucleotides are connected at the bases, forming complementary pairs, the bases on one strand are paired to the other strand. The CMV DNA molecule contains 229,354 complementary pairs of base pairs. Once infected, CMV lays dormant and only becomes harmful when the virus enters a productive cycle in which it quickly replicated, posing a great risk for people with vulnerable immune system. By locating the origin of replication of CMV, we may help with the design of an effective vaccine against it.

Investigation

Scenario 1: Random Scatter

We will investigate the structure in the data indicated by departures from a random scatter of palindromes across the DNA. An important thing we should notice is that a random uniform scatter does not mean that the palindromes will be equally spaced. There will be some gaps on the DNA where no palindromes occur, and there will be some clumping together of palindromes.

Multiple random scatters were generated in which 296 palindromes, chosen by a pseudo random number generator, were randomly scattered along a DNA sequence of 229,354 base pairs. Then the locations of the palindromes, the spacing between palindromes, and the counts of palindromes in non-overlapping regions of the DNA are compared with the real data. The structure of the real data is compared with the structure of a Monte Carlo simulated random scatters from the uniform distribution.

Location:

Figure 1 depict a simple dot plot of the palindrome locations. We will compare the simulated strip plot from simulated data, with the strip plot that generated from real data. From the graph, it is difficult to recognize any patterns in either the real data or the random scatters. However,

we can observe that the palindrome locations are not exactly equally spaced. There is no distinguished cluster which means it is possibly uniformly distributed.

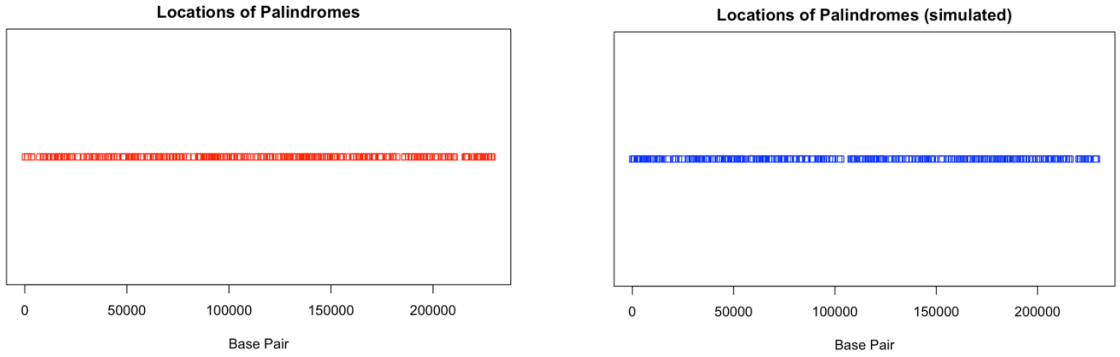


Figure 1. Simple Dot Plot of Palindrome Locations

Figure 2 are the histograms of the palindrome locations. From the histograms, we can see that the original data appear to be clustered around the 90000th base pair. The intervals with higher counts possibly contain those palindromes. However, in the random scatters, clusters are not as apparent, which indicates the data's departure from a random scatter.

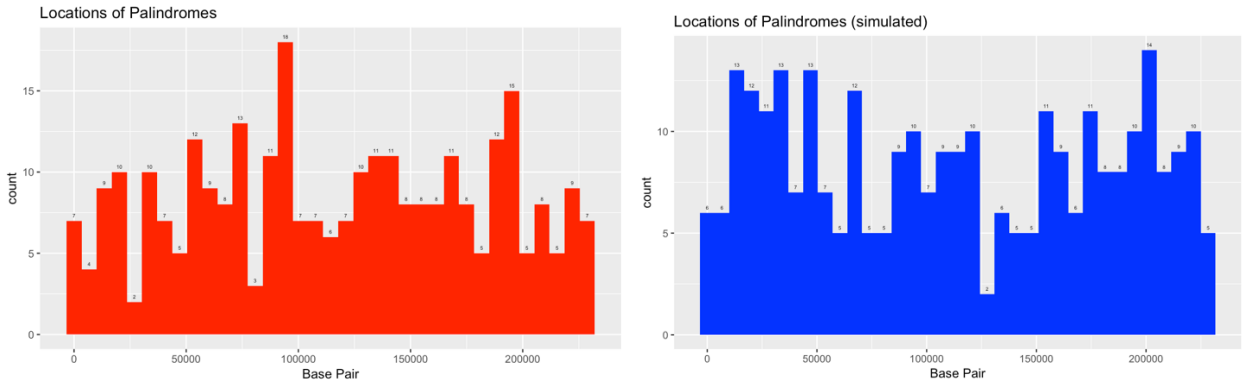


Figure 2. Histogram of Palindrome Locations

Spacing:

Figure 3 depict scatterplots of spacing between consecutive palindromes. From the graph, again it is difficult to recognize any patterns in either the real data or the random scatters. We will further investigate the spacing in the Scenario 2.

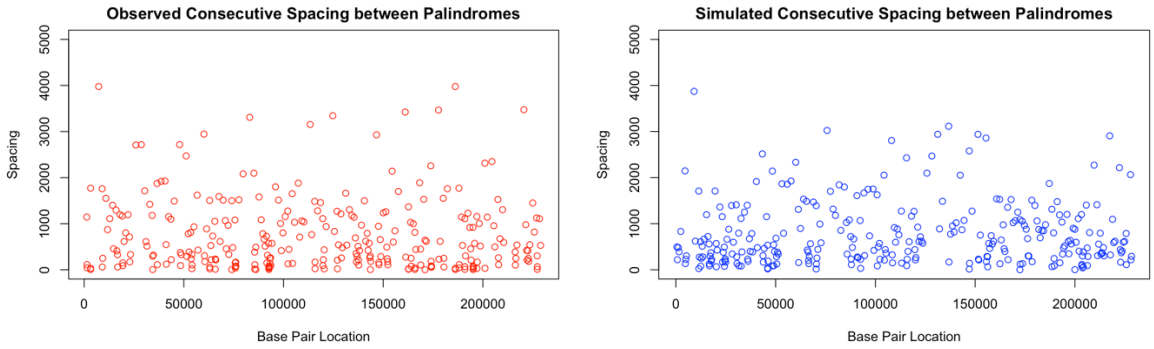


Figure 3. Scatterplot of Spacing Between Consecutive Palindromes

Counts:

Figure 4 depict histograms of the number of palindromes in non-overlapping regions of length 2,500 base pairs. From the graph of the real data, there are outliers in the distribution's right tail. Outliers are not as apparent in the random scatters which indicates the data's departure from a random scatter.

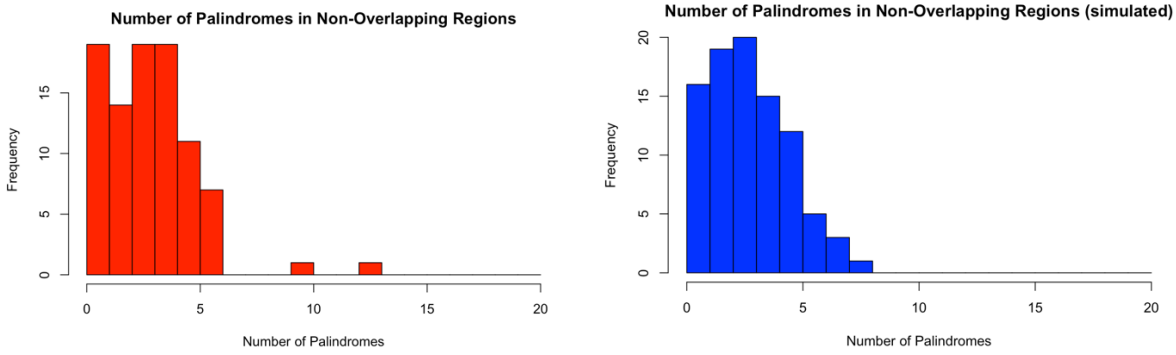


Figure 4. Histogram of Number of Palindromes in Non-Overlapping Regions

In conclusion, after comparing the locations of palindromes, the spacing between palindromes, and the counts of palindromes in non-overlapping regions of the DNA, we can see that the data appears to departure from a random scatter indicating the existence of clusters of palindromes, a potential replication site. In the following scenarios, the probability will be discussed in detail by implementing graphical and statistical methods.

Scenario 2: Locations and Spacings

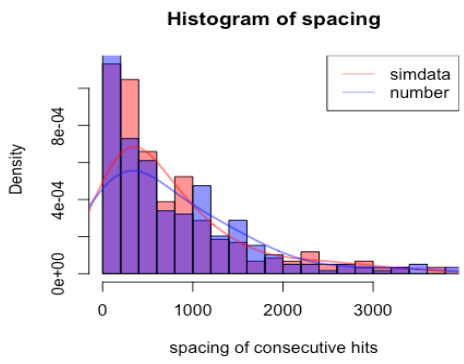
Since the given data "hcmv.txt" provides the locations of the occurrence of palindromes, we compute the spacing between two consecutive palindromes by setting the occurrence of first palindrome as 0. Then we get 295 results with minimum of 1, maximum of 5333 and average of 775.5.

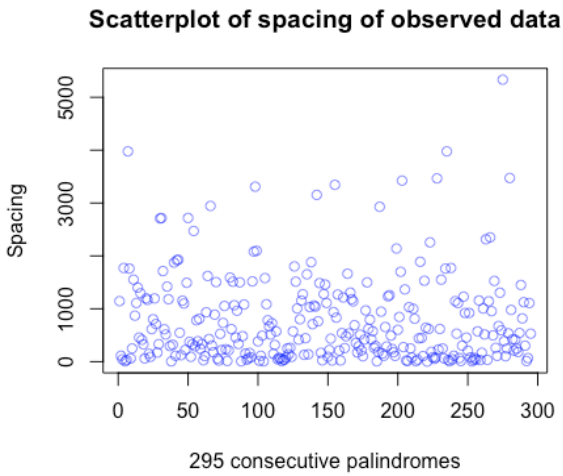
In order to get the clearest data distribution, we choose k=20 and divide results into 20 intervals. As we can see from the table, the data distributed in each interval is enough for observation and also not too concentrated. Thus, we believe k=20 is an appropriate choice.

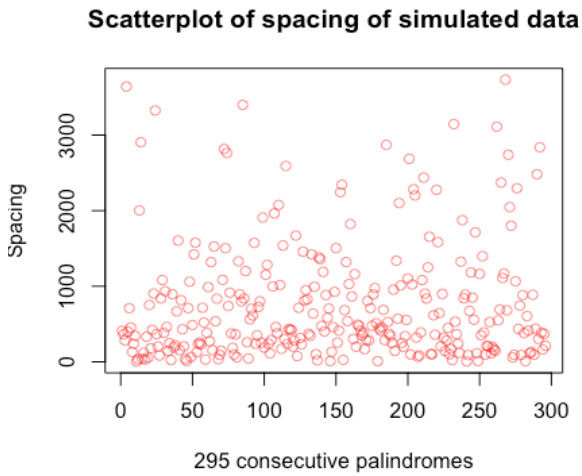
[0,267]	(267,533]	(533,800]	(800,1.07e+03]
106	47	30	26
(1.07e+03,1.33e+03]	(1.33e+03,1.6e+03]	(1.6e+03,1.87e+03]	(1.87e+03,2.13e+03]
31	19	10	7
(2.13e+03,2.4e+03]	(2.4e+03,2.67e+03]	(2.67e+03,2.93e+03]	(2.93e+03,3.2e+03]
4	1	4	2
(3.2e+03,3.47e+03]	(3.47e+03,3.73e+03]	(3.73e+03,4e+03]	(4e+03,4.27e+03]
3	2	2	0
(4.27e+03,4.53e+03]	(4.53e+03,4.8e+03]	(4.8e+03,5.07e+03]	(5.07e+03,5.33e+03]
0	0	0	1

We assume that the spacing between consecutive palindrome follows the exponential distribution. The we use the max likelihood estimation to calculate our λ based on exponential distribution. The likelihood function gives the data as $L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$, and the log-likelihood function $l(\lambda) = nlog(\lambda) - \lambda \sum_i x_i$. By solving the last equation for λ, we obtain: $\hat{\lambda} = \frac{1}{\bar{x}} = 0.00129$.

According to upper calculation, we are given the estimated λ and use it to simulate a group of 295 exponential distributed random variables for the comparison between spacings of observed data and of random scatter. We firstly draw the histograms of the spacing of the sample using blue color. As we can observe from the shape of the histogram and its density line, it is very likely that the data of spacing is exponentially distributed. To further explore our guess, the second histogram and density line is drawn with 295 simulated data following exponential distribution λ=0.00129. The graph is in pink to contrast with the sample graph in blue. It can be observed that when spacing between palindromes is relatively large, the sample data fits exponential distribution better compared to small spacings. However, we can see that no matter the length of the spacings, there are always difference in density between our sample and the distribution supposed to be. In order to further the comparison, we then draw scatterplots for each group of data. It is obvious that the sample data follows loosely the dot pattern of exponential distribution. Therefore, it would seem logical to deduce that the unusual clusters exist in observed data, and further tests are needed for the examination of fitness of exponential distribution though the differences are not huge.



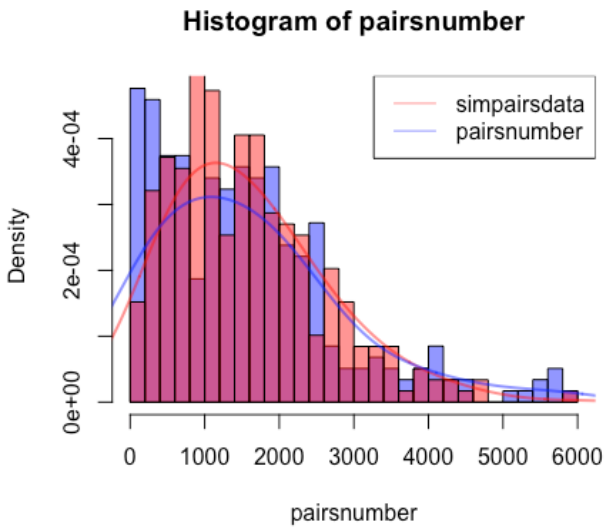


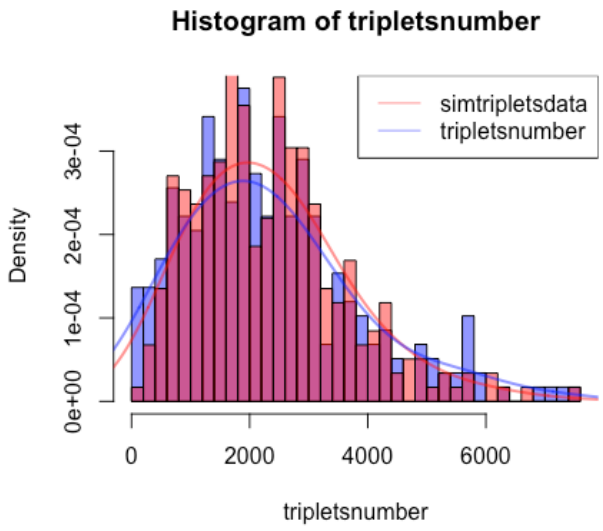


What also requires examination is spacings between consecutive pairs and triplets palindromes. According to the slides discussed in class and previous results, we assume that the distances between consecutive hits follow an exponential distribution. And thus it is able to conclude that spacings between consecutive pairs A, and triplets palindromes B, both follow a Gamma distribution respectively with parameters (2, λ), (3, λ).

We firstly draw the histogram of A, which is spacings between consecutive pairs. As the graph is left-skewed and assumption that it follows gamma distribution, we generate a random gamma distribution with the same parameter λ=0.00129 and shape equals 2, for the sake of comparison. It is clearly indicated by the graph (Histogram of pairs number) below that there are apparent diversions between our sample number distribution and the random generated distribution, especially in the interval from 0 to 2000 in spacings. With further numerical statistics test, we know that the density of spacings between two consecutive hits in our data is larger than the random scattered situation in the interval (0,400). Thus, the existence of unusual clusters is implied in this interval.

Similarly, deviations exist between histogram of spacing of triplets' palindromes and random gamma distribution simulated. Since the graph representing distribution of spacings between triplet palindromes is skewed slightly to the left and assumed to be gamma distribution, we generated another gamma distribution with random variables using again the λ=0.00129 and shape equals 3. As we can see from the graph, the sample fits mostly well with the gamma distribution though differences still exist. In the interval (0,200) of spacings, the density of triplets spacings is apparently larger than that of gamma distribution, which again implies the possibility of clusters when spacings are relatively small.





We know that Gamma distribution is the sum of exponential distribution. We conduct a hypothesis test to test whether our sample follows exponential distribution or not.

$H_0 = the\ spacing\ between\ consecutive\ palindrome\ follows\ the\ exponential\ distribution$
$H_1 = the\ spacing\ between\ consecutive\ palindrome\ does\ not\ follow\ the\ exponential\ distribution$

We already have the estimated λ by max likelihood estimation, and therefore we could calculate the expect value assuming H_0 is true (the spacing follows the exponential distribution). Our sample size is 295, and we calculate the expected value of 295 numbers in different interval which following the exponential distribution randomly. Following the exponential distribution, we have the density function is $\int \lambda e^{-\lambda x} dx$ with $\lambda = 0.00129$. We run the whole process 500 times and calculate the average to get a more accurate expected value shown in the table below.

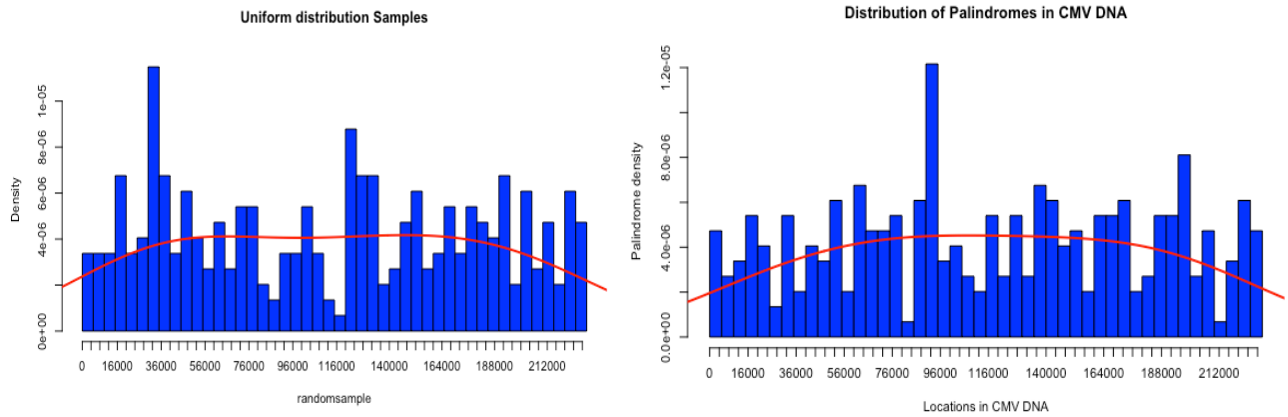
Spacing Interval	Observation	Expected Value
[0,267]	106	86.316
(267,533]	47	61.026
(533,800]	30	43.630
(800,1070]	26	30.920
(1070,1330]	31	21.022
(1330,1600]	19	15.784
(1600,1870]	10	10.984
(1870,2130]	7	7.372
(2130,2670]	5	9.582
(2670,3200]	6	4.658
(3200,5333]	8	4.404

After comparing the observation and expected value, it is obvious that there is significant difference between observation and expected value, which seem like corresponds to our assumption that the spacing between consecutive palindrome does not follow exponential distribution.

Then, we use chi-square test to get the accurate result. The definition of chi-square test is $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$. Based on the data shown in the table above, we calculate χ^2= 23.766 by using R. Assuming that our significant level $\alpha = 0.05$, we know that our degree of freedom is 9 (numbers of interval - 1), and then we calculate the p-value = 16.919 by using R. Thus, since 23.766 > 16.919, we reject H_0 and in favor of H_1. We conclude that the spacing between consecutive palindrome does not follow the exponential distribution.

Scenario 3: Counts

In order to determine the counts of palindromes in various regions of the DNA, we want to test whether our data follows the uniform distribution and the Poisson distribution. In order to visually understand the sample, we did two histograms of the distribution. There are 296 palindromes and we divide CMV DNA into 57 non-overlapping regions of equal average length 4000.



We first generated a uniform distribution based on a random sample in the same interval. Second, we generated a distribution of palindromes in CMV DNA and get the above right graph. By comparing these two figures, the density curve is roughly following the uniform distribution. Then in order to further test our hypothesis, we did a Chi-Squared Test statistic to verify whether the distribution of the palindromes does really follow the uniform distribution in different intervals.

H_0= the data follows the uniform distribution
H_1= the data does not follow the uniform distribution

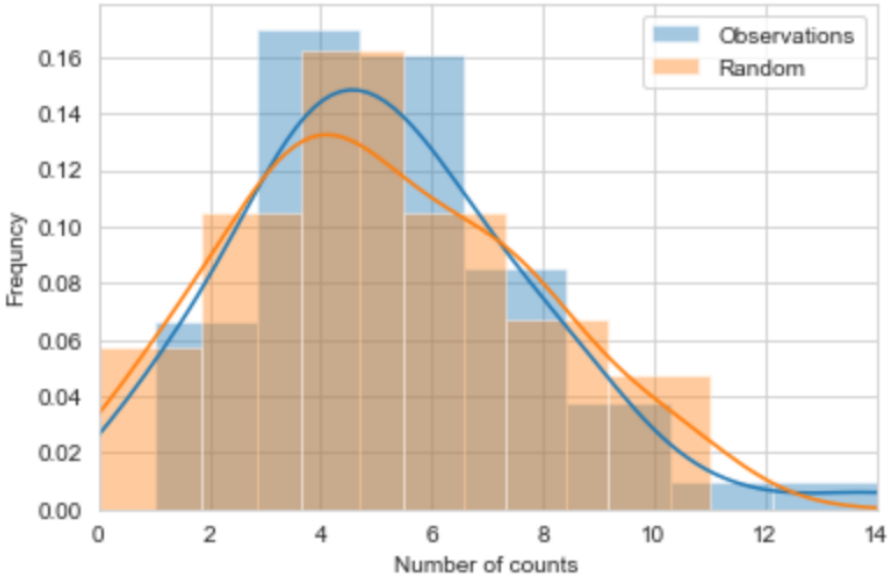
According to the formula, $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = n \sum_{i=1}^{k} \frac{(\frac{O_i}{n} - p_i)^2}{p_i}$ and by choosing $\alpha = 0.05$, we get our sample's chi-squared=72.189 with p-value=0.07 at the 57 intervals, which indicates that we fail to reject the null hypothesis. Thus, the distribution of CMV DNA does follow the uniform distribution. Besides, in further support our hypothesis of uniform distribution, we regroup the data into 38 subintervals in order to satisfy the requirement of Chi-square test that the expected value of the sample observations in each interval should be at least 5. At this new subinterval, we get our sample's X-squared=21.865 with the p-value=0.978, which strongly indicate that our data fail to reject the null hypothesis. Thus, our distribution of CMV DNA does follow the uniform distribution.

Next, we want to further investigate if the number of counts of palindrome in 57 intervals fit into Poisson distribution. We will take two different division strategies to regroup the data set.

Method:

A: Divide the data set to intervals with length equal to 4000

The project first generates random Poisson data sets and compares the graph with given observations:



Second, we use statistics to further examine whether the observed data is like Poisson distribution:

Because we want to perform chi-square test, we need to ensure that every interval expected value is larger than 5(np>5). Therefore, p needs to be calculated by Poisson distribution density function: $P(x=k)=\frac{\lambda^k e^{-k}}{k!}$, where $\lambda = \frac{296}{57} = 5.19$. In this case, $P(x=0,1,2)=e^{-5.19}(1+5.19+\frac{5.19^2}{2!})=6.23$, $P(x=3)=\frac{e^{-5.19}5.19^3}{3!}=7.39$

The result is presented as follows:

	Palindrome count	Number of Observed	Interval expected
0	0-2	7	6.23
1	3	8	7.39
2	4	10	9.60
3	5	9	9.97
4	6	8	8.63
5	7	5	6.40
6	8	4	4.15
7	9+	6	4.64

But the expected intervals of sixth group and seventh group are less than 5, so we need to regroup the observations:

	Palindrome count	Number of Observed	Interval expected
0	0-2	7	6.23
1	3	8	7.39
2	4	10	9.60
3	5	9	9.97
4	6	8	8.63
5	7	5	6.40
6	8+	10	8.79

The third step is to perform Chi-square goodness-of-fit test:
H_0: The sample distribution is Poisson Distribution
H_1: The sample distribution is not Poisson Distribution

$$\chi^2 = \frac{(7-6.23)^2}{6.23} + \frac{(8-7.39)^2}{7.39} + \frac{(10-9.60)^2}{9.60} + \frac{(9-9.97)^2}{9.97} + \frac{(8-8.63)^2}{8.63} + \frac{(5-6.40)^2}{6.40} + \frac{(10-8.79)^2}{8.79}$$

The statistics is 0.775, and p-value is 0.992, which is larger than 0.05 significance level. So we fail to reject H0. We can conclude that the counts of interval fit Poisson distribution.

B. Divide the sample evenly to intervals with approximate equal length

Another approach is to divide the data evenly----with every interval with length 4016. The result of this approach is represented as following:

	Palindrome Count	Number of observed	Interval expected
0	0-2	10	6.23
1	3	4	7.39
2	4	7	9.60
3	5	10	9.97
4	6	7	8.63
5	7	12	6.40
6	8+	7	8.79

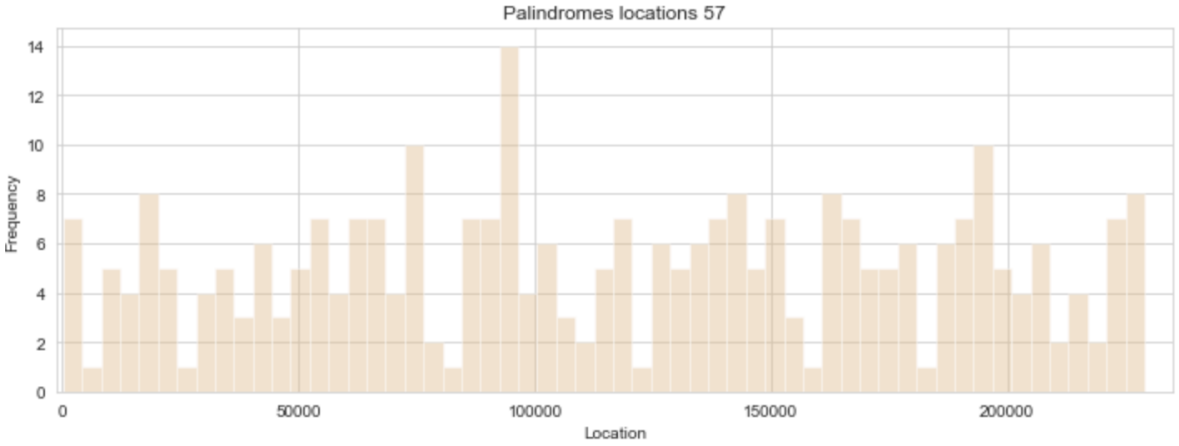
H_0: The sample distribution is Poisson distribution
H_1: The sample distribution is not Poisson distribution

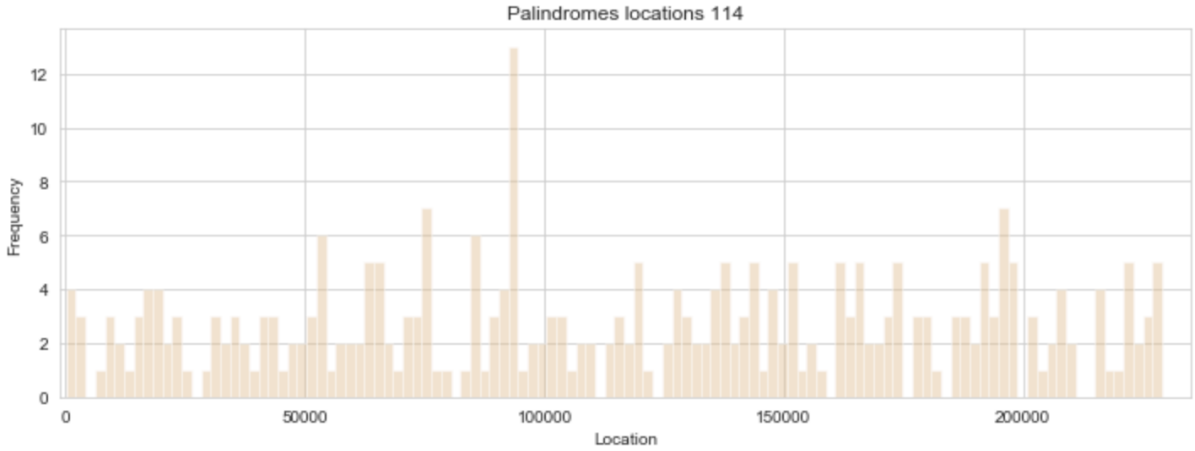
The chi-square statistic is 10.113 and p-value is 0.112, which is less than 0.05, we reject H0.So the project generates two different statistics and results according to different approach. The difference is caused by distinct interval strategy and special clusters of palindrome because we already know that palindromes show higher frequency around location 96000 and 196000 a. When we divide the data with length 4000, the clusters are probably separated evenly to different intervals. However, palindromes aggregate in several intervals when we use the second approach, causing the chi-square value very high.

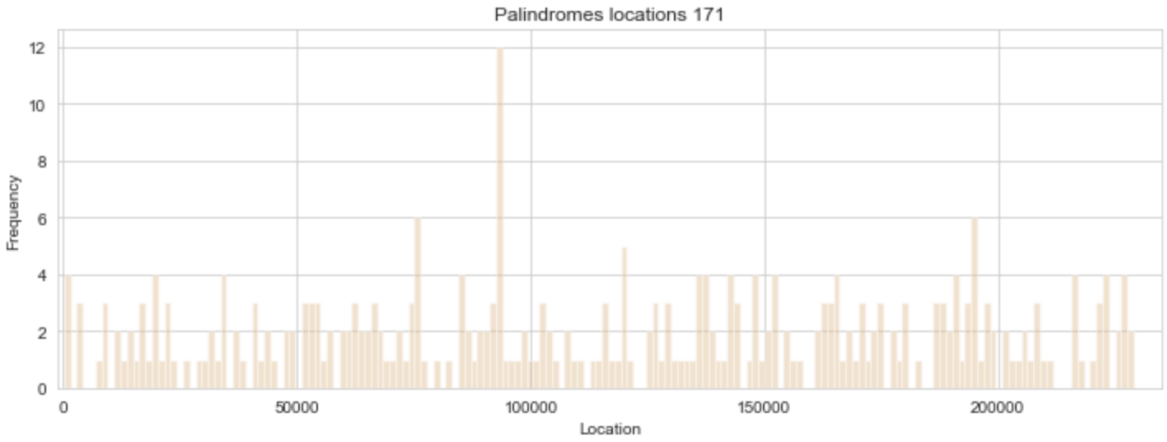
Scenario 4: Biggest Count and Clusters:

According to theories of DNA and RNA sequence, “Palindromic sequences (inverted repeats) flanking the origin of DNA replication with the potential of forming single-stranded stem-loop cruciform structures have been reported to be essential for replication of the circular genomes of many prokaryotic and eukaryotic systems.” If we want to search for origins of replication, it is best to find regions with largest number of palindromes or the biggest clusters of palindromes.

The following graphs are palindromes distribution among virus alphabetic sequence according to 57, 114 and 171 intervals.



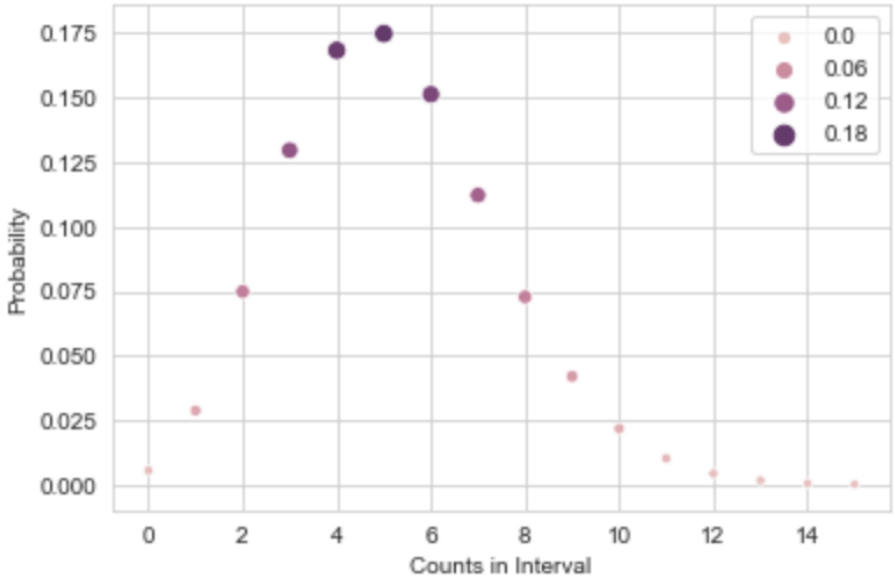




From the graphs, one can roughly conclude palindromes seem to aggregate in three locations: 75000, 95000 and 19000. To further investigate the palindrome clusters, the projects summarize palindrome counts in 57 intervals.

7, 1, 5, 4, 8, 5, 1, 4, 5, 3, 6, 3, 5, 7, 4, 7, 7,
4, 10, 2, 1, 7, 7, 14, 4, 6, 3, 2, 5, 7, 1, 6, 5, 6,
7, 8, 5, 7, 3, 1, 8, 7, 5, 5, 6, 1, 6, 7, 10, 5, 4,
6, 2, 4, 2, 7, 8

The first interval with high counts is [72422, 76436] with 10 counts, the second is [92490,96503] with 14 counts and the third is [192830, 196844] with 10 counts. Because we already prove the distribution is Poisson Distribution, we can calculate the occurrence based on Poisson probability density function:



We can observe that when occurrence approaches 14, the probability is close to 0. So the interval with 14 counts of palindrome can be considered as a special case.

In this part, we investigate the distribution of counts of palindromes in intervals. We establish hypothesis that it fits Poisson distribution. Then we use chi-square test to prove this hypothesis.

Based on Poisson Distribution, the project compares the observed data with theoretical values and concludes that palindromes aggregate in interval [92490,96503]. Biologically, we highly recommend researchers focus on this interval because the aggregations of palindromes mean the origins of replication.

Scenario 5: Additional Hypothesis

Another approach to Clustering:

In order to get more precise locations of cluster of palindromes, we can try another algorithm “Jenks natural breaks optimization” to compute the intervals of clusters. This algorithm is to minimize each class’s average deviation from the class mean, while maximizing each class’s deviation from the means of the other groups. The detailed algorithm can be referred in reference.

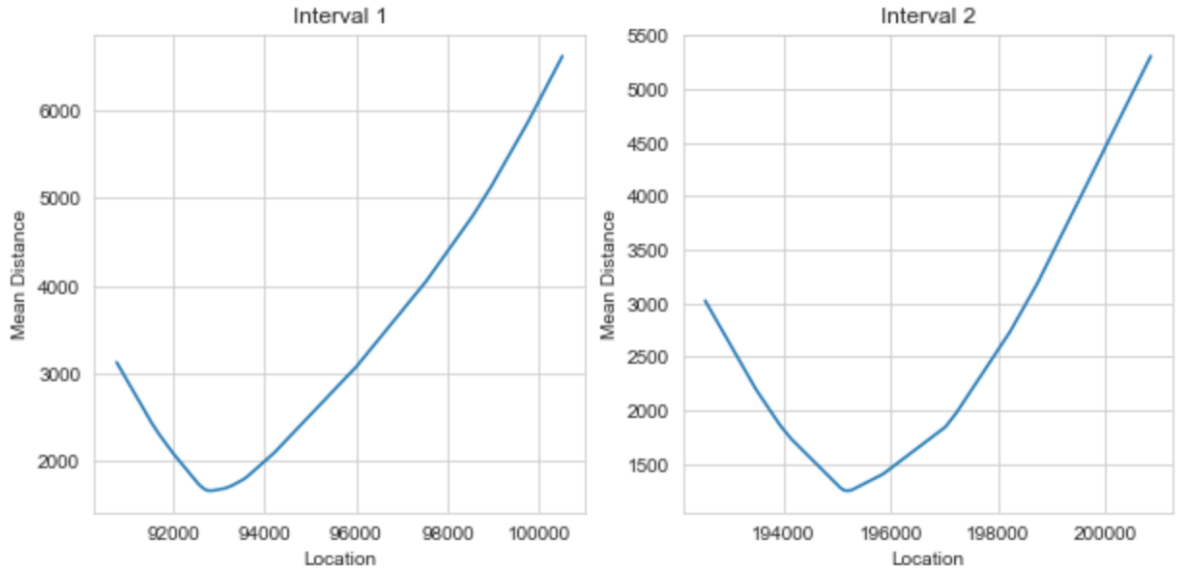
After implementing Jenks natural breaks, we get distinct demarcations and counts of palindromes in each interval:

6, 4, 3, 6, 6, 4, 2, 3, 5, 3, 4, 3, 4, 4, 6, 2, 4,

6, 6, 5, 9, 2, 7, 5, 16, 5, 4, 3, 4, 6, 7, 6, 3, 4,
9, 3, 7, 6, 6, 4, 8, 5, 3, 4, 5, 3, 4, 6, 7, 10, 5,
3, 3, 6, 5, 8, 9
(Also see Appendix)

The highest counts increase to 16, but the interval [90763, 100517]is approximately same as previous analysis. Besides, another interval [192527, 200857] with 10 counts appear to be another potential candidate for origins.

The next process is to focus on two intervals. If we want to narrow the region of origin of replications, the best way is to apply cluster analysis. Thus, we define a center of cluster as **the point which minimizes the average distance to other palindromes in the same interval.**



Both intervals represent local minimum: 92783 in the first interval and 195151 in the second interval.

Conclusion:

In this part, we investigate the distribution of counts of palindromes in intervals. We establish hypothesis that it fits Poisson distribution. Then we use chi-square test to prove this hypothesis.

Based on Poisson Distribution and “Jenks natural breaks optimization” algorithm, the project compares the observed data with theoretical values and concludes that [90763, 100517] and [192527, 200857] appear to be most possible regions of origins of replication. More precisely, 92783 and 195151 are potential “centroids” for clusters. Biologically, we highly recommend researchers focus on two “centroid” because the aggregations of unusual palindromes mean the origins of replication.

Theory:

Goal

Our goal is to understand a random model that describes the behavior of ”counts” of the number of palindromes and for a ”uniform” aka random scatter of palindromes. We want to determine the estimation procedure in such a model. Also, we want to understand how to find statistical discrepancies between a model with clusters and model without clusters. (i.e. Is a model a good model? Can we formulate spacings as well as counts in the model? What is a hypothesis tests? How is uniform distribution related to the problem?)

The Homogeneous Poisson Process

The Homogeneous Poisson Process is 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. The process arises naturally from the notion of points haphazardly distributed on a line with no obvious regularity. It is widely used to model random points in time or space. Let us assume that we are observing number of occurrences of certain event over a specified period of time. We can consider them as happening under a Poisson Process provided they satisfy below conditions.

1. The underlying rate λ at which points, called hits, occur and is such that it doesn't change with location (homogeneity).
2. The number of points falling in separate regions are independent.
3. No two points can land in exactly the same place.

If we denote number of occurrences during a time interval of length t as X(t) then

$$\mathrm{P}(\mathrm{X}(\mathrm{t}) = \mathrm{n}) = \frac{e^{-\lambda t}(-\lambda t)^n}{n!}$$

Checking the Homogeneous Poisson Process

The poison process is a good reference model for making comparisons because it is a natural model for uniform random scatter. The strand of the DNA can be thought of as a line, and the location of a palindrome can be thought of as a point on the line. According to uniform random scatter model, palindromes are scattered randomly and uniformly across the DNA. All the three properties are met: The chance that one tiny piece of DNA ha palindrome in it is the same for all tiny pieces of the DNA (Property 1). The number of palindromes in any small piece of DNA is independent of the number of palindromes in another, non-overlapping piece (Property 2&3).

Maximum Likelihood

Maximum likelihood estimates the unknown parameter by the λ-value that maximized the likelihood function. The likelihood function gives the data as $L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$, and the log-likelihood function $l(\lambda) = nlog(\lambda) - \lambda \sum_i x_i$. By solving the last equation for λ, we obtain: $\hat{\lambda} = \frac{1}{\bar{x}}$.

Exponential and Gamma distribution

Distances between successive hits follow an Exponential distribution with the density function of $\int \lambda e^{-\lambda x} dx$. Distances between the hits that are two apparatus, follows a Gamma distribution with parameters 2, λ.

Hypothesis test

Hypothesis test is used for the $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ goodness-of-fit test and the test for the maximum number of palindromes in an interval. The null hypothesis H_0 and the alternative H_1 are formulated. Then P-value is calculated through test statistics and compared to significance value α. If $P \geq \alpha$, the alternative hypothesis is ruled out, and the null hypothesis is valid.

Counts and The Poisson Distribution

Counts in different regions do follow the Poisson distribution with the rate λ. And the formula for the Poisson distribution is P (k points in a unit interval) $= \frac{\lambda^k}{k!} e^{-\lambda}$. In this formula, λ represents the rate of hits per unit. However, λ is unknown. In this case, there are two methods of estimation. The first one is method of moments, which is estimating the empirical average number of hits per unit interval. Another method is maximum likelihood method.

Maximum Number of Hits

Under the Poisson process, the number of hits in a set of non-overlapping intervals is independent observations, which implies that the greatest number of hits can be represented as the maximum of independent Poisson random variables. In light of this, we can get P(maximum count over m intervals $\geq$k) = 1-P(maximum count over m intervals <k) =1-P(all interval counts <k)^m =1-[$\lambda^0 e^{-\lambda} + \cdots + \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda}]^m$. Then, if the chance is small, it can indicate the cluster is larger than the expected from the Poisson process. In this case, we can use the maximum palindrome counts as a test statistic.

Method of Moments

For this part, our group find E(X) where X has Poisson distribution with λ, then express the rate λ in terms of E(X), and replace E(X) with $\bar{x}$ to produce an estimate $\hat{\lambda}$.

Chi-Squared Test Statistics

Chi-Squared Test is the method to test the relationship between categorical variables. The formula is $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = n \sum_{i=1}^{k} \frac{(\frac{O_i}{n} - p_i)^2}{p_i}$. O_i and E_i is the observed and theoretical frequencies; k is the number of intervals; n is the number of samples; $p_i = \frac{E_i}{n}$ is the expected probability in the i-th interval. Since the size of the test statistic is the measure of the fit of the distribution, then if the test statistic is large, that will indicate the distribution is the lack of fit.

Besides, when we construct the hypothesis test for a discrete distribution, we compare the counts with the expected counts under the null hypothesis, where $\mu_j = np_j \; and \; p_j = P$. Since $\sum p_j = 1$, then $\sum_j \; u_j = n$.

P-value

With m-k-1 degrees of freedom, where m is the number of categories and k is the number of parameters, we can compute p-value using Chi-squared test. If the P-value is small, the fit of distribution is doubtful. Then by calculating $\frac{sample\ count - expected\ count}{\sqrt{expected\ count}} = \frac{N_j - u_j}{\sqrt{u_j}}$, we can plot the standardized residuals.

Besides, if the total number of hits in an interval is known, the positions of hits are uniformly scattered across the interval. Applying that to our case for CMV DNA, our group compares the locations to the expected locations from the uniform distribution. For instance, if we divide the sample into 10 subintervals, we can expect each interval contain $\frac{1}{10}$ of the palindromes.

Jenks natural breaks optimization (for the additional hypothesis)

Jenks natural breaks optimization is a method of data classification for analyzing the cluster, which divides a range of numbers in order to minimize the squared deviation and maximize the variance within classes.

Conclusion:

All of the scenarios have shown that our data does not follow the uniform distribution, Poisson distribution and Gamma Distribution. Therefore, we highly suspect the existence of clusters and palindromes and there is very high probability to find the origin of replication in the particular interval with abnormal cluster.

Hence, in order to find the origin of replication, we suggest the biologists cut DNA into segments and test for the unusual big cluster. The process of detecting viruses can be very time consuming and expensive without leads on where to begin the search. 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. First, we need to generate a random uniform sample, and compared it with our real sequence sample. After that, we can locate abnormal clusters.

Work Cited

Bradic, Jelena. “Chapter 4: Patterns in Data.” MATH 189 Lecture, UC San Diego, 1 May 2018. Lecture.

Ryan KJ, Ray CG, eds. (2004). Sherris Medical Microbiology (4th ed.). McGraw Hill. ISBN 978-0-8385-8529-0. pp. 556, 566–9.

Appendix:

Intervals by “Jenks natural breaks optimization” algorithm
{(177.0, 4190.6140350877195, 6),
(3286.0, 8204.228070175439, 4),
(9333.0, 12217.842105263158, 3),
(12863.0, 16231.456140350878, 6),
(16812.0, 20245.070175438595, 6),
(20832.0, 24258.684210526317, 4),
(23241.0, 28272.298245614038, 2),
(28665.0, 32285.912280701756, 3),
(31503.0, 36299.52631578947, 5),
(34723.0, 40313.14035087719, 3),
(38626.0, 44326.754385964916, 4),
(42376.0, 48340.36842105263, 3),
(45188.0, 52353.98245614035, 4),
(48699.0, 56367.596491228076, 4),
(52629.0, 60381.210526315794, 6),
(55075.0, 64394.82456140351, 2),
(57123.0, 68408.43859649124, 4),
(61441.0, 72422.05263157895, 6),
(64502.0, 76435.66666666667, 6),
(68221.0, 80449.28070175438, 5),
(72553.0, 84462.8947368421, 9),
(76124.0, 88476.50877192983, 2),
(79724.0, 92490.12280701754, 7),
(86137.0, 96503.73684210527, 5),
(90763.0, 100517.35087719299, 16),
(94174.0, 104530.9649122807, 5),
(99709.0, 108544.57894736843, 4),
(102711.0, 112558.19298245615, 3),
(105534.0, 116571.80701754386, 4),
(110224.0, 120585.42105263159, 6),
(117097.0, 124599.0350877193, 7),
(121370.0, 128612.64912280702, 6),
(127587.0, 132626.26315789475, 3),
(129537.0, 136639.87719298247, 4),
(134221.0, 140653.49122807017, 9),
(138111.0, 144667.1052631579, 3),
(141201.0, 148680.71929824562, 7),
(143738.0, 152694.33333333334, 6),
(148821.0, 156707.94736842107, 6),
(152331.0, 160721.56140350876, 4),
(157617.0, 164735.1754385965, 8),
(164072.0, 168748.7894736842, 5),
(166372.0, 172762.40350877194, 3),
(168815.0, 176776.01754385966, 4),
(171607.0, 180789.6315789474, 5),

(174260.0, 184803.24561403508, 3),
(178574.0, 188816.8596491228, 4),
(182195.0, 192830.47368421053, 6),
(188137.0, 196844.08771929826, 7),
(192527.0, 200857.70175438598, 10),
(195835.0, 204871.31578947368, 5),
(198709.0, 208884.9298245614, 3),
(202198.0, 212898.54385964913, 3),
(206000.0, 216912.15789473685, 6),
(210469.0, 220925.77192982458, 5),
(217076.0, 224939.3859649123, 8),
(223544.0, 228953.0, 9)}