Patterns In DNA

January 23rd, 2021

Derek Leung	*Timothy Tran*	*Cynthia Leung*	*Jasmine Guan*	*Dillen Padhiar*	*Marcos Arevalo*
Introduction	*Data*	*Background*	*Investigations*	*Investigations*	*Theory*

1. Introduction

Human cytomegalovirus (CMV) is a virus with health impacts severe enough to be lethal to people with a suppressed or deficient immune system. Given the risk people take with exposure to this virus and the medical advancements of the present day, it follows that scientists have decided to learn how to stop this virus from reproducing. The best way to do this is to figure out how the virus reproduces in the first place. While we know the information we need will be in the DNA, at the origin of replication, finding this segment of information can be a difficult task.

The DNA contains all the information necessary for the virus to grow, survive, and reproduce. It is made up of 4 bases: adenine (A), cytosine (C), guanine (G), and thymine (T), which appear in many different patterns to make up the sequences that construct DNA. The most important point here is that certain patterns may indicate important sites on DNA such as the origin of replication.

In order to figure out what kinds of patterns might flag the origin of replication for CMV, we refer to prior knowledge about a couple viruses in the same family which have already been researched. One of these similar viruses, herpes simplex has been found to have an origin of replication marked by a complementary palindrome of 144 bases. Another, Epstein-Barr, has several short palindromes and close repeats clustered at the origin of replication. These complementary palindromes are defined as a sequence that reads in reverse as the complement of the forward sequence where A complements T and G complements C. For example, AATT and GCACCGGTGC would be considered to be complementary palindromes.

In the case of CMV, the longest palindrome in the DNA is 18 base pairs long and the DNA, in total, contains 296 palindromes where each is between 10 and 18 base pairs long. Based on the knowledge of the similar viruses, we speculate that the clusters of palindromes in CMV may flag the origin of replication for this virus. The test to determine if the origin of replication is present in a sequence is to see if an individual segment is able to replicate. Without direction, these tests can be numerous and too costly in terms of both time and money as a result of the quantity. Therefore, our goal is to search the DNA for unusual clusters of complementary palindromes so that they might be isolated for testing.

2. Data

This dataset is composed of 296 palindrome locations along the DNA chain of the human cytomegalovirus, or CMV for short. There were other palindromes found, but only the locations of ones that were at least 10 letters were included in the dataset. The longest ones were found at locations 14719, 75812, 90763, and 173893 along the 229,354 long strand.

In this dataset, only 1 feature was recorded, and that was the location of the palindrome. This feature is considered a categorical variable despite having digits, as it does not make sense to perform calculations on the individual numbers of the locations. Doing so could give us a number not relevant to this feature, which is the locations of the palindromes with at least 10 letters long.

With these locations, we can find the parts of DNA that are the origin of replication, and develop strategies in combating this virus. By grouping these locations into any length interval, we can see that there appears a cluster of palindromes around the 93,000th and 195,000th pairs of DNA. We can also observe that the palindromes present higher spikes of palindromes per interval. However, these two cases do not appear when using a random number generator, which makes it logical to examine around these two locations to see if the replication code exists here.

3. Background

In this project, we will be focusing on determining the origin of replication of the human cytomegalovirus (CMV), a member of the herpes virus family. CMV infection varies geographically, but typically 10-15 percent of children are infected with CMV before the age of 5, and infection levels increase during young adulthood, where the CMV infection presents symptoms similar to mononucleosis. Like most viruses, CMV becomes harmful when it enters the productive cycle, and this poses major risk to immune-depressed patients. Two viruses, the Herpes simplex and Epstein-Barr virus, within the same family as CMV have complementary palindromes as the marker for the origin of replication. The Herpes simplex is marked by a long palindrome of 144 letters while the Epstein-Barr virus has several short palindromes and close repeats clustered at the origin of replication; therefore biologists believe that clusters of palindromes may be the marker in CMV.

In a 1999 paper, Leung et al. studied the application of scan statistics in DNA sequence analysis, specifically on the number of palindromes that form clusters within the herpesvirus family. The paper demonstrated that the location of palindromes are uniformly distributed along the unit interval. Additionally, the study examined three Poison-type calculations before computing the r-scan statistics. With the compound Poisson approximation, Leung et al. were able to identify important clusters within various DNA sequences of the herpesvirus family and were able to identify the origin of replication for three specific DNA virus sequences. Ultimately, the study

concluded that using Poisson approximations are useful in bioinformatics in determining clusters; however, the authors noted that a cluster does not always signify an important biological site, as biology and statistical significance are two separate fields.

A 2005 paper by Leung et al. studied the methods for determining nonrandom clusters within the herpesvirus genome. The paper provides a mathematical basis to demonstrate that the Poisson process can be used to approximate the location of these complementary palindromes in randomly generated DNA sequences. This was a critical finding because scan statistics was typically used to identify unusual clusters of palindromes within a DNA sequence, and scan statistics requires the assumption that points representing palindromes are independently and uniformly distributed on a unit interval. The analysis of the paper not only demonstrated the use of Poisson process and scan statistics, but also proved the importance of length of the palindrome as a consideration for determining unusual clusters. Ultimately, the outcome of the paper demonstrated that the method in the paper can be utilized to prioritize which segments of DNA can be tested first, which would expedite genome testing for origins of replication.

Another study in 2013 Tu et al. built upon the Leung et al. paper to determine a better method for estimating the occurrence rate of palindromes within a DNA sequence, and they tested this method on the herpes virus sequence bohv1. The paper is structured around the observation that in scan statistics, the commonly-used average rate estimation method can overestimate the null occurence rate by 50 percent, and this is especially seen when the null occurrence rate is very low. Tu et al. proposes a Markov based estimator and demonstrates that the Markov method is more robust against this hot-spot effect than the average rate method. Hot spots refer to locations of high palindrome occurrence in this instance. The analysis presents that the Markov rate estimate is more powerful in detecting nonrandom clusters from simply chance occurrences, especially when the Poisson process involves rare events with hot-spots.

4. Investigation

Scenario 1

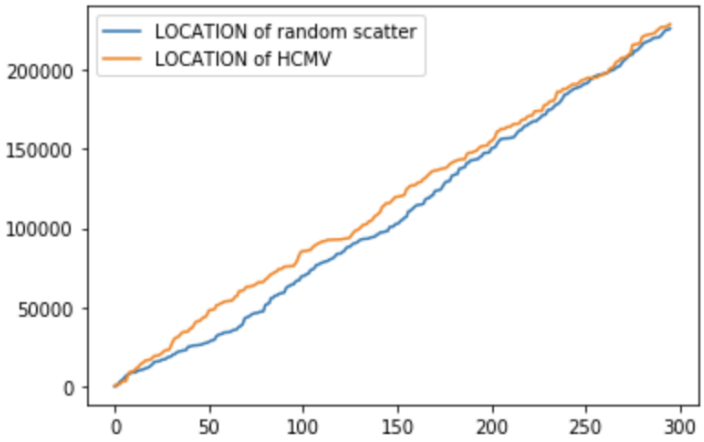
To investigate whether departures from a uniform scatter is indicative of potential sites for the “origin of replication”, we will first simulate 296 palindromes sites chosen at random along the DNA sequence of 229,354 according to a uniform distribution through a pseudo random number generator. Comparing the data to a random cluster can aid us in exploring whether longest palindromes or unusual clusters of palindromes could be valid candidates for CMV’s “origin of replication”.

Scenario 2

Next, we want to compare the palindrome location, spacing, consecutive pairs and consecutive triplets of the CMV data to the hypothesized distribution.

Location:

Graphic Exploration:



Caption: The palindrome distribution of the random scatter showcases a clear linear trend, which means the 296 palindromes are uniformly scattered around the potential sites of 229,354. When the palindrome distribution of CMV is overlaid with the random scatter, even though the distribution poses a general upward linear trend, there are visible departures from the random scatter around the 25th to 130th palindrome and small departures around 130th to 170th palindrome.

Testing Distribution:

In order to explore whether the location of palindromes of CMV follows a Uniform Distribution, we divided the DNA sequence of 229,354 into 20 intervals with equal length, calculated the expected number of palindromes in each interval, which is 296/20, found the observed number of palindromes in each interval, and performed Chi-Square Goodness of Fit test. Since there isn't any estimated parameters, the chi-square test statistic follows a χ_{19} distribution.

Goodness of Fit Test:

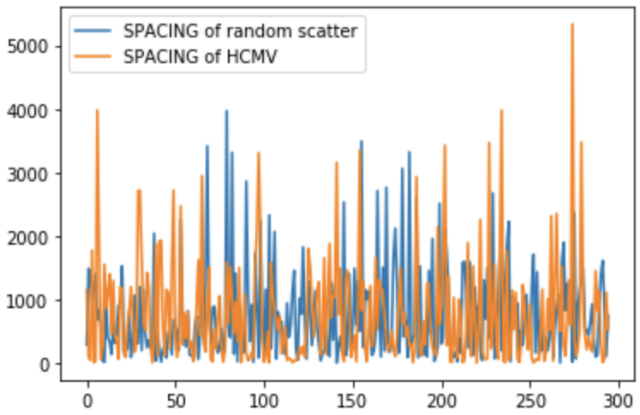
- Null Hypothesis: location of palindrome of CMV follows a Uniform Distribution
- Alternative Hypothesis: location of palindrome of CMV does not follow a Uniform Distribution
- As a result, we have:

Test Statistic = 17.92

P-Value = 0.528

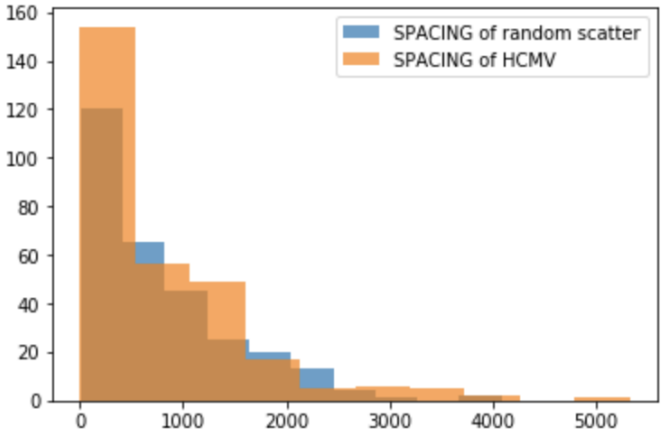
Conslution: fail to reject the null hypothesis

Spacing:



Caption: As indicated by the above graph, the spacing between consecutive palindromes across the 296 palindromes are extremely variant both in the case of random scatter and CMV. This indicates that there isn't a regular spacing between consecutive palindromes across the DNA sequence.

A noticeable difference is that there exists a significantly higher spike of MCV around the 270th palindrome. This indicates that there are two palindromes that are significantly farther away from each other than the average spacing.



Caption: To summarize the distribution of spacing, the above histogram showcases counts of spacing between significant intervals. The spacing of random scatter showcases a downward exponential trend. The overlaid CMV distribution also showcases exponential trends but has a much higher peak at the front of the graph.

The above graph indicates that, in CMV case, there are more palindromes with smaller spacing between each other than in random scatter , indicating clusters or potentially unusual clusters.

Testing Distribution:
To test whether the spacing between consecutive palindromes follow an exponential distribution, we divided the DNA sequence into equal-length intervals, used the Maximum Likelihood

Estimator, which is the inverse of average spacing, to estimate $\hat{\lambda}$ of the exponential distribution, utilized the Exponential CDF to calculate the expected counts in each interval, summarized the observed counts in each interval, and used these two counts array to perform Chi-Square Goodness of Fit Test.

The initial few bins between the observed CMV and the random scatter shows a visible deviance. And this is shown in the goodness of fit test when the number of bins are altered. The result of test is sensitive to the number of bins we divide.
Goodness of Fit Test:

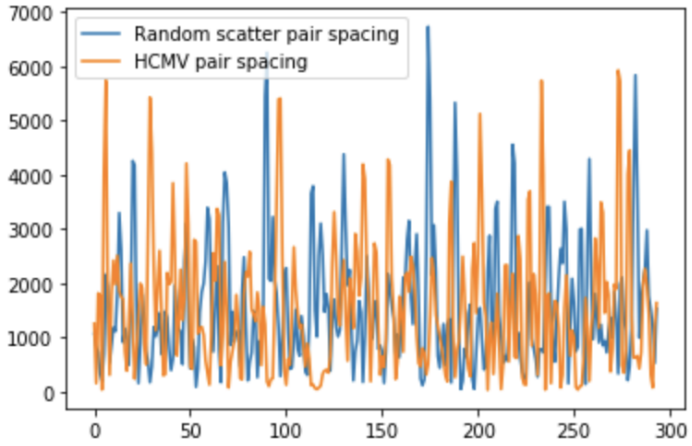
- Null Hypothesis: the spacing between consecutive palindromes of CMV has exponential distribution.
- Alternative Hypothesis: the spacing between consecutive palindromes of CMV does not have exponential distribution.
- Case 1: number of interval = 8

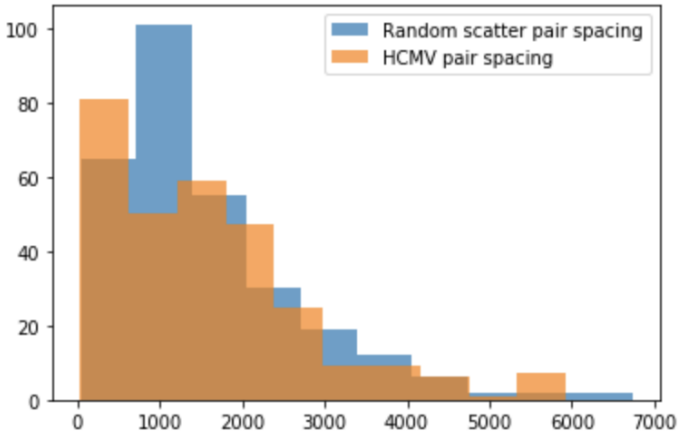
  Distribution ~ χ_6
  Test Statistic = 16.63
  P value = 0.07
  Conclusion: fail to reject null hypothesis
- Case 2: Number of interval = 9

  Distribution ~ χ_7
  Test Statistic = 7.03
  P value = 0.43
  Conclusion: reject null hypothesis
- Case 3: Number of interval = 10

  Distribution ~ χ_8
  Test Statistic = 17.28
  P value = 0.03
  Conclusion: reject null hypothesis

As seen by the 3 cases above, we can see that the test for whether spacing takes up exponential distribution is highly sensitive to the number of intervals we're separating the data into. As shown by the graphic, it is likely due to the deviation in counts at the initial spacing around 0 ~ 500. A possible explanation for this sensitivity to the number of intervals is that, as the number of intervals increases, the expected counts histogram will become closer to the probability density function of the Exponential Distribution and the deviations become more sensitive for the test.



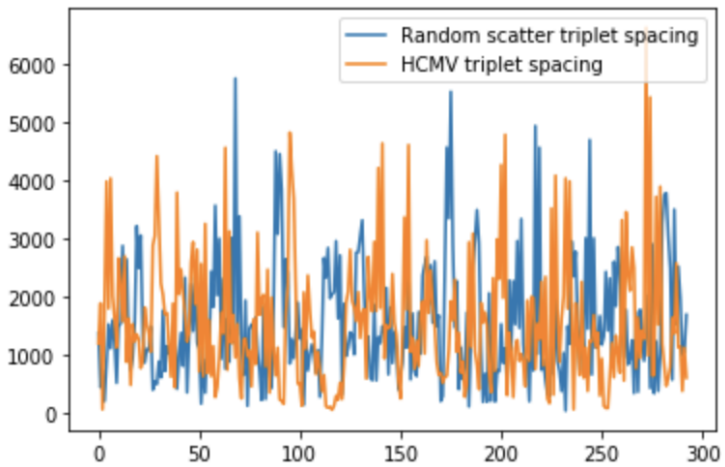


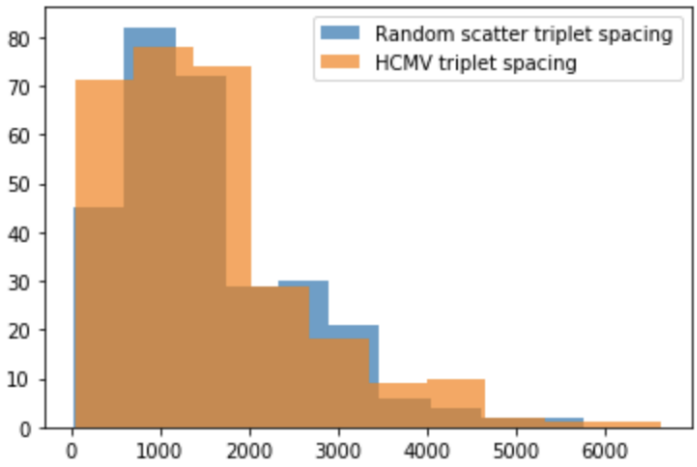
Consecutive Pairs:

Caption: The figure on the left indicates variable pair spacing of palindromes on for both random scatter and CMV, as expected from the distribution of spacing. The figure on the right showcases an interesting distribution of palindromes of random scatter and CMV.

The distribution of random scatter resembles a Gamma Distribution with drastic increase to a peak and slow decrease to 0. However, the distribution of CMV does not follow this similar trend but still poses a slight downward exponential trend. This is an early indication that the pair spacing between palindromes might not follow the uniform random scatter.

Consecutive Triplets:

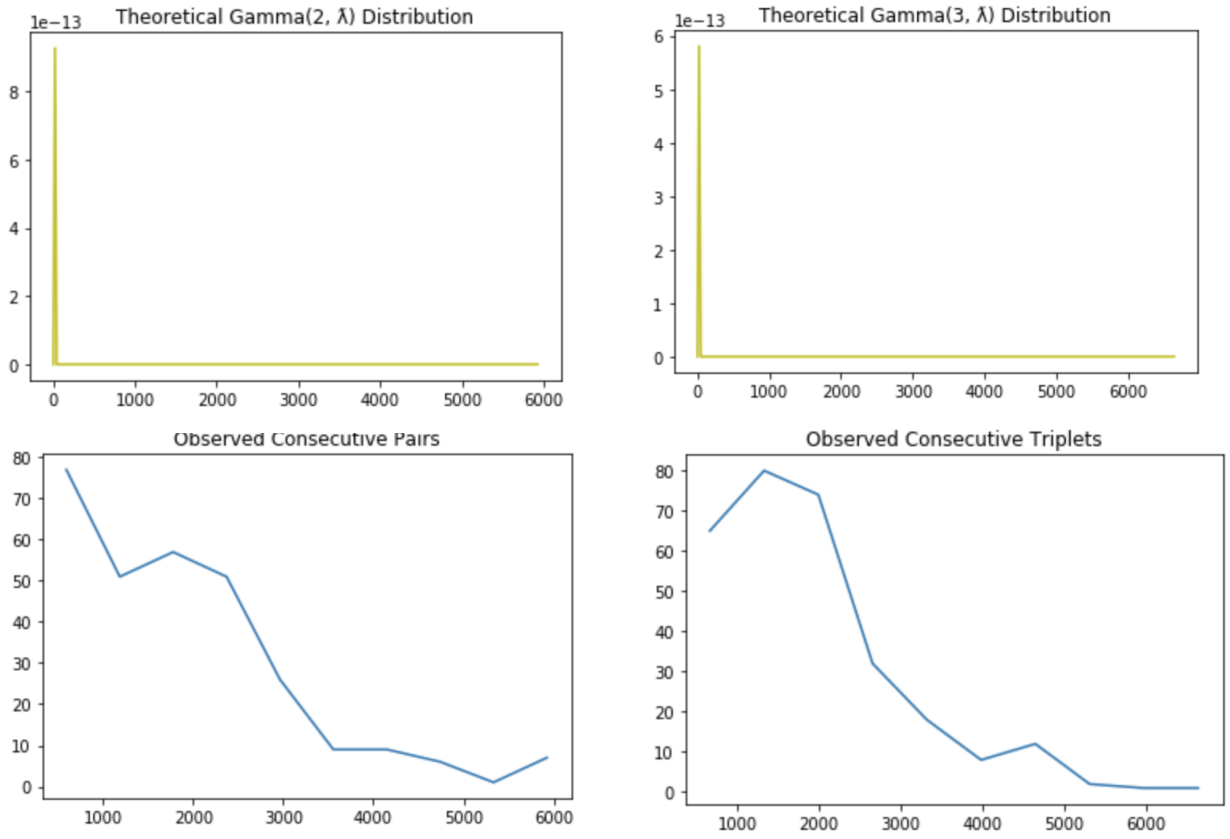




Caption: Consecutive triplets show the same pattern as consecutive pairs. The variation across the DNA sequence appears to be random.

The front of the distribution of consecutive triplets of CMV appears to be significantly different from the random scattering, which again, has signs of Gamma Distribution.

Testing Distribution: Consecutive Pairs & Triplets



Caption: The theoretical gamma distribution with the estimated parameter looks significantly different from the observed distribution of consecutive pairs and consecutive triplets.

We use MLE to estimate λ̂ of the Gamma Distribution. For consecutive pairs, our null hypothesis is that data follows a Gamma(2, λ̂) distribution; for consecutive triplets, our null hypothesis is that data follows a Gamma(3, λ̂) distribution.

$$\hat{\lambda} = r / \bar{x}$$

Gamma (2, λ̂): r = 2

Gamma(3, λ̂): r = 3

As shown above, the theoretical gamma distribution with the estimated parameter looks significantly different from the observed distribution of consecutive pairs and consecutive triplets. This also corresponds to the result of our Goodness of Fit Test.

Consecutive Pairs: P Value = 0

Consecutive Triplets: P Value = 0

Conslution: reject null hypothesis

As indicated by the test, we reject the null hypothesis that consecutive palindromes follow a Gamma Distribution. Since Gamma Distributions are the sum of exponential distributions, the fluctuation in the test for exponential distribution above might have been magnified in the sum of pairs and triplets scenario.

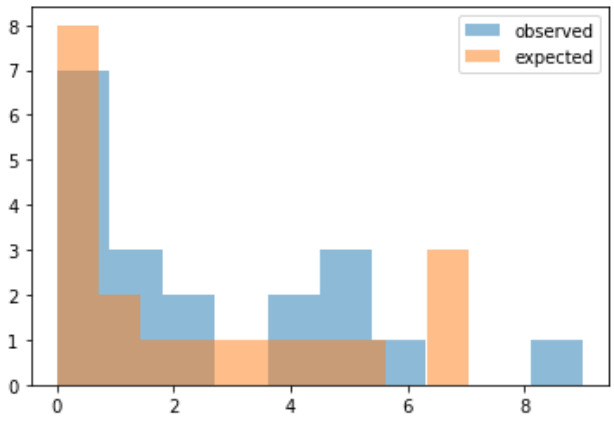
Scenario 3

After exploring location and spacing of the palindromes, we will shift our focus towards the distribution of counts of palindromes. We will once again divide the DNA sequence into equal length intervals, counted the number of palindromes in each interval, and used the counts as the values of another distribution.

This distribution has counts k as the independent variable ranging from 1 ~ maximum number of palindromes in each interval, and the number of intervals that has k palindromes contained in it as the dependent variable. For our null hypothesis, we speculates that this particular distribution follows the Poisson distribution. We estimated the Poisson(ƛ) parameter with $\lambda = \frac{296}{\#\ of\ bins}$. With this estimated parameter, we calculated the empirical distribution of counts across 1 ~ maximum number of palindromes in each interval, calculated the observed counts distributed across, and passed those values into the Chi-Square Goodness of Fit Test. Similar to the Exponential Distribution case, the choice of interval number has an influence over the results.

Goodness of Fit Test:

- Null Hypothesis: the counts of palindromes in each intervals has Poisson Distribution
- Alternative Hypothesis: the counts of palindromes in each intervals does not have Poisson Distribution



Caption: This histogram shows how many bins are expected to have k-palindromes versus how many we had in our observed data.

- Case 1:

Interval Size = 5000
Test Statistic = 206.14
P Value = 1.66e-34
Conclusion: reject the null hypothesis

- Case 2:

Interval Size = 4500
Test Statistic = 750.65
P Value = 1.47e-148
Conclusion: reject the null hypothesis

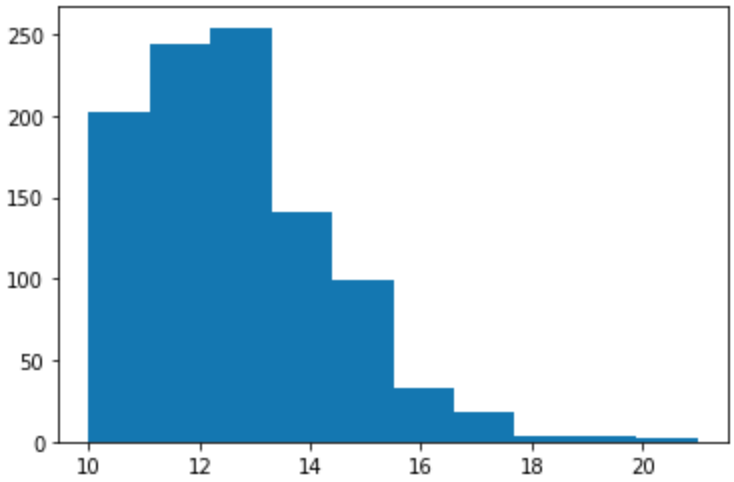
- Case 3:

Interval Size = 5500
Test Statistic: 39.92
P Value = 0.00092
Conclusion: reject the null hypothesis

Scenario 4

Here we will try to determine whether or not a cluster is chance occurrence or a potential replication site. We do this by examining the largest cluster in our data and running tests to see if the amount of palindromes contained in this cluster is unusual compared to our simulations. For our hypothesis tests for this section:

1. Null Hypothesis: The biggest cluster in the observed data is not under the simulated distribution.
2. Alternative Hypothesis: The biggest cluster in the observed data under the simulated distribution.



Caption: This histogram of the maxes of the simulations shows how uncommon our observed max of 18 is thus exemplifying it as an outlier.

- Case 1:

  Interval Size = 5000
  P Value = 0.002
  Conclusion: reject the null hypothesis
- Case 2:

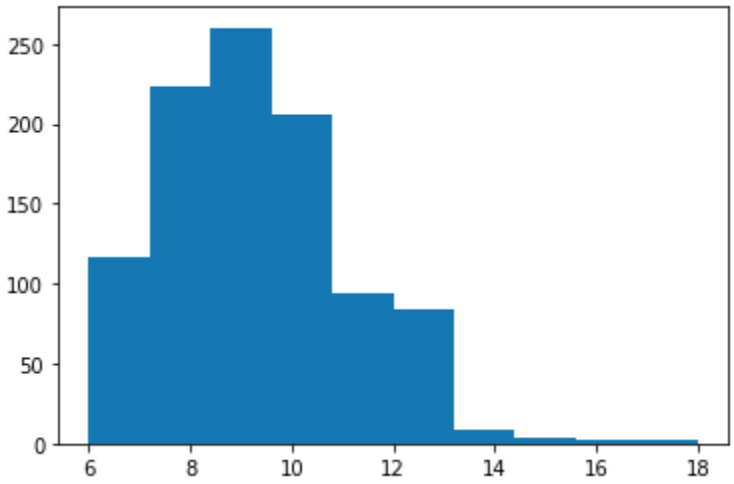
  Interval Size = 4500
  P Value = 0.025
  Conclusion: reject the null hypothesis
- Case 3:

  Interval Size = 5500
  P Value = 0.001
  Conclusion: reject the null hypothesis

To test these, we randomly generated counts of bins with certain amounts of palindromes (ie. 3 bins with 10 palindromes) using a poisson distribution. We then took the max of theses counts which corresponds to the biggest cluster and stored it into a matrix. We ran this simulation 1000 times to collect 1000 max counts and then compared them directly to the max of our observed data's biggest cluster. We tested this over 3 different bin amounts and found that our p-value did not change significantly. Thus, we reject the null hypothesis and conclude that the observed data is not under the simulated distribution. With this, we can consider the biggest cluster to potentially be a potential replication site however we cannot fully prove that it is one. If biologists were to look at this data, it would be wise for them to start looking at the clusters with the most palindromes to find the origin of replication.

Scenario 5

To approach this question in a different way, we chose to fully simulate new datasets instead of just generating exact counts of the cluster using the poisson distribution. In these tests, we generated a new dataset each simulation similar to how we did in Question 3 with our observed dataset. We then find the cluster with the most palindromes and save its count of palindromes to an array with all max counts. We then compare all values in this array to our observed max to see if it exceeds it. We find with 3 different bins counts it does not change our p-value and thus our conclusion from Question 4 holds.



Caption: As we can see in this figure, the observed max of 18 is not common compared to multiple other maxes.

Scenario 6

In addition to what we've done above, our additional investigation will see if looking at spacing distance rather than count of palindromes would give us indication of whether there are unusual clusters to reinforce our findings from above. We designed a hypothesis test, similar to the one tested in problem 4, but instead of counts of palindromes, we focused on spacing distance. For our hypothesis test:

1. Null Hypothesis: The smallest average spacing distance of a bin of the observed data is under the simulated distribution.

2. Alternative Hypothesis: The smallest average spacing distance of a bin is not under the simulated distribution.

3. Test Statistic: The minimum average spacing distance among the bins of equal interval lengths.

To do this, we randomly generated locations of each palindrome under the uniform distribution. Then, we calculated the spacing distance between each palindrome from the randomly generated data, created bins for each one, and found the average spacing distance for

each bin. After, we found the minimum of the average spacing distance among all the bins and stored this value. This was repeated 1000 times, and three different interval lengths (4500, 5000, 5500 per each bin) were used to test for sensitivity. The results are as follows:

Length of Interval	p-value
4500	0.003
5000	0.013
5500	0.010

Ultimately, for all interval lengths, we are able to reject the null hypothesis. The minimum average spacing distance that we see among the observed bins is unusual, and these results further prove that spacing between clusters of palindromes can also be another consideration when looking for unusual clusters. These results also make sense when considering the results of the previous simulation with max number of palindromes, as a smaller spacing distance would correlate to a higher number of palindromes within a bin.

Summary

In summary, we are able to identify unusual clusters of palindromes by comparing the observed to simulated data. Our results show that spacing between palindromes and number of palindromes within a given interval are important to identifying unusual clusters. For future research, biologists can examine the number of palindromes and spacing between palindromes to identify potential origin of replication sites and have a more efficient method of testing for these sites rather than simply performing random tests. Through scan statistics, these sites of abnormally high number of palindromes with minimal spacing can be the starting point for biologists to determine the origin of replication for viruses.

5. Theory

5.1.1 The Poisson Process

In order to understand a random model to describe the behavior of counts for a uniform random scatter, the usage of the Poisson Process is essential. This process allows us to approximate the number of occurrences in an independent and identically distributed distribution with intensity λ, where λ is the expected count. Defined as $\{ N(t), t > 0\}$, the Poisson Process consists of

non-negative, integer valued random variable N(t) for any t > 0, for which the probability mass function (p.m.f.) of N(t) is given by the Poisson p.m.f. (described further) with λt. Then,

$$P[\ N(t) = k] = \frac{(\lambda t)^k}{k!} * e^{-\lambda t}$$

Hence, for

$$t = 1 : P\ [\ N(1) = k] = \frac{(\lambda)^k}{k!} * e^{-\lambda}$$

$$t = 2 : P\ [N(2) = k\ ] = \frac{(\lambda 2)^k}{k!} * e^{-\lambda 2}$$

and henceforth. We then define S to be the arrival times of the instances of X where $S_1, S_2, \dots S_n$ such that $S_1 < S_2 < \dots S_n$. To more accurately account for inter-arrival times, we then define $Y_1, Y_2, \dots Y_n$ where $Y_1 = S_1, Y_2 = S_2 - S_1, Y_3 = S_3 - S_2, Y_4 = S_4 - S_3$, and so on.We can then define Z to be the time from t until the first arrival after t occurs, which will be approximated by the exponential distribution of λ. Then $P\ [\ Z > a] = e^{-\lambda a}$, which is conditioned on N(τ) , where $\tau \in (0, t)$.

Thus we have seen that the distribution of the instances of X has a poisson distribution and that the inter-arrival times have an exponential distribution. In reality however, a third component is needed to ensure the Poisson Process, which involves a uniform distribution of the arrival times (S). To show this, it is important to recall that N(t) represents the number of instances of X in the interval of time (0, t]. Then, we know $N(t_1) - N(t_2)$ is distributed by $N(t_1 - t_2) = N(t_1, t_2)$ where $t_1 > t_1$. It is therefore implied that $N(t_1)$, $N(t_1, t_2)$, $N(t_2, t_3)$, etc are statistically independent. Following this derivation, we use our previously defined variables to develop W_1 such that $Y_1 \mid S_2 = t_2$. W_1 is then uniformly distributed at $(0, t_2]$. It then follows that $(S_1, S_2, \dots S_n) \mid N(t) = n$ is uniform as required.

5.1.2 The Homogeneous Poisson (Counting) Process

More specifically, we can specify the Poisson Process to be homogenous if

- The underlying rate λ at which counts occur does not change with location, a trait commonly referred to as homogeneity
- The number of counts per interval is independent of the rest
- No two counts can occur at the exact same location

5.2 The Poisson Distribution

Letting X be a discrete random variable, X is approximately distributed by the Poisson distribution with rate λ > 0 such that $E\ [\ X\ ] = \lambda$ and the $Var\ [\ X\ ] = \lambda$. Additionally, the probability mass function follows such that

$$P[X=k] = \frac{\lambda^k}{k!} * e^{-\lambda} \text{ where } k \in \mathbb{Z}_{\geq 0}.$$

λ however is unknown, so instead we estimate it using $\hat{\lambda}$, which is defined as the observed average of the discrete random variable X within each unit interval, where average is defined to be

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

This process of estimating $\hat{\lambda}$ is called a method of movement, but we could also use maximum likelihood estimation to reach the same conclusion.

5.3 Hypothesis Test

In hypothesis testing, you test two assumptions, more formally hypotheses, using the data available to you. Typically, there are two hypotheses: the null hypothesis (H_0) and the alternative (H_1). Once the test is concluded, you either accept or fail to reject the H_0. Typically, the H_0 contains the answer to the scientific question (normally yes) and one tests to see if this claim is true. The test as a whole can have one of 4 types depending on the fundamental question being investigated. If you are testing towards large positive or negative values it is classified as a one-sided test. Similarly, if you test for both large positive and negative values it is classified as a two-sided test. However, if you are testing for distributions and simply state the null hypothesis is false, this is classified as a complex test. Alternatively, if you specify an alternative type of distribution it is then classified as a simple test.

When performing hypothesis tests however, there are key errors, called Decision Errors, one can make when arriving at a conclusion. If you fail to reject the null hypothesis even when it should have been rejected, a type II error occurs. On the other hand, if you reject the H_0 but should have failed to reject, a type I error occurs. Typically, society relies more on Type I error not occurring, so hypothesis tests are designed in a way that type 1 errors can be controlled using the level of a test, or how confident you want to be in your results, and is dependent on the H_0. Type II error is instead represented by the power of the test, where power is defined as 1 - type II error. This power is a curve that is dependent on the H_1, which is then dependent on far the H_1 is from the H_0.

5.4.1 Chi-Square Goodness-Of-Fit Test

The Chi-Square (χ^2) test, a type of parametric test, is the testing of a distribution such that it follows some distribution. Because of this, the H_0 is always that it follows some distribution which is specific to the test. The $\chi\chi\chi\chi\chi\chi^2$ test relies on the probabilistic result stemming from a mathematical theorem which originated in 1900: start with the summation

$$\sum_{j=1}^{r} \frac{(v_j - np_j)^2}{np_j}$$

where v_j is a random dependent variable with binomial distribution (n, p_j) and np_j is its expectation. From there, the summation will converge in distribution to a χ^2 random variable with r - 1 degrees of freedom.

To perform the actual test, you first start with the null hypothesis H_0, which states that the independent observations $W_1, W_2, \dots W_n$ have a given distribution, such as uniform, poisson, or exponential. Then, depending on the type of distribution you are testing against, you will calculate the unknown p_j by choosing the amount of intervals that will occur the distribution, defined as $B_1, B_2, \dots B_r$. To choose r, in practice it is done so that either $np_j \geq 5$ for all j, or $r = 2 * n^{2/5}$(which sometimes results in a large but acceptable r). However, r is an example of a tuning parameter and thus sensitivity analysis is crucial to ensure that the results do not change drastically by a small change in r.

5.4.2 Chi-Square Test for Uniform Distribution

If we perform the Chi-Square against the uniform distribution [0, L] where L is the length of the distribution, we can then compute p_j such that

$$p_j = P(W \in B_j) = \int_{Bj} Density$$

We then reject the H_0 if the observed test statistic is larger than the theoretical percentile we are testing in comparison to χ^2r-1 , usually 95%, with the belief that H_0 is true and that the test statistic follows the χ^2 distribution.

5.4.3 Chi-Square Test for Poisson Distribution

Under the assumption that H_0 is true, we estimate lambda using a maximum likelihood estimator for the observations as defined in 5.2. From there, each p_j can be calculated by $P(Poisson(\hat{\lambda}) \in B_j)$. The test statistic is approximately models by the χ^2r-1 distribution, and is modeled by the previously defined theorem in 5.4. For example, if the first probability phat is composed of k = {0,1,2}, we can separate

$$\begin{aligned} & P((Poisson(\lambda) = 0) \cup (Poisson(\lambda) = 0)) \cup Poisson(\lambda) = 0)) \\ = \; & P(Poisson(\lambda) = 0) + P(Poisson(\lambda) = 1) + P(Poisson(\lambda) = 2) \\ = \; & \frac{(\lambda)^0}{0!} * e^{-\lambda 0} + \frac{(\lambda)^1}{1!} * e^{-\lambda} + \frac{(\lambda)^2}{2!} * e^{-\lambda} = e^{-\lambda}[1 + \lambda + \lambda^2] = H_0 \end{aligned}$$

Then, if the test statistic is larger than the tested percentile of the χ^2 distribution, we fail to reject the H_0. Alternatively, you could instead use p-values to determine the outcome of your hypothesis test. In this case, the p-value is defined as the P [Theoretical distribution of T under the H_0 > observed value of the test statistic. If the p-value is above the percentage threshold, you fail to reject the H_0.

5.5 Testing for Uniform Distribution

To test a distribution for uniform distribution, you first establish your H_0 as the location of X having a uniform distribution, where X is a discrete random variable, and you H_1 as the location of X not having a uniform distribution. More specifically, the H_0 states that each interval I within the distribution contains 1/I of all instances of X within I. Letting L_1, L_2, … L_n be the locations of X, L is assumed to be an independent random variable that is uniformly scattered under the null hypothesis. Then you make use of the previously defined χ^2 r-1 to compute the test statistic, then then compare it to the desired percentage threshold and reject or fail to reject accordingly.

5.6 Testing for Exponential Distribution

When testing for an Exponential Distribution, we are in fact trying to see if the distance between to consecutive locations of the previously defined X has an exponential distribution λ. In other words,

P[Distance between the first and second instance of any two sequential Xs > t] = $e^{-\lambda t}$

If the goal is to calculate consecutive pair distances, the H_0 would follow that the distance between to consecutive instances of X follows a Gamma Distribution of (2, λ). Similarly, if the goal is consecutive triplets, then the H_0 follows a Gamma Distribution of (3, λ) and so on so forth for consecutive α with Gamma Distribution (α ,λ). The Gamma Distribution has the probability distribution function

$$f(x;\alpha,\beta) = \{\frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \; if\; x \geq 0,\; 0\; otherwise \;\} \text{ where } \Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$

You then calculate your test statistic, and follow the same procedure of rejecting or failing to reject the null hypothesis.

5.7 Scan Statistic

In order to test for an unusual cluster of instances of X, a Scan Statistic is needed to check for the grouping of X. More formally, you first compute T_m to be the max number of instances of X over m intervals. Additionally, we assume under the null that the number of instances of X

follows the Poisson distribution (λ) and that T_m has a distribution of a maximum of independent Poisson (λ) random variables. To do this, we compute

$$P[T_m < a] = P[\text{maximum of independent Poisson } (\lambda) < a]$$
$$= P[\text{ all of independent Poisson } (\lambda) < a \text{ }]$$
$$= P[\text{first interval has Poisson } (\lambda)$$
$$= P[\text{Poisson } (\lambda) = 0] + P[\text{Poisson}(\lambda) = 1] + \ldots + P[\text{Poisson}(\lambda) = a]$$
$$= e^{-\lambda m} * (1 + \lambda + \frac{\lambda^2}{2!} + \ldots + \frac{\lambda^a}{a!})$$

One then computes the p-value by using the new observed test statistic, labeled *t*, to find

$$P[T_m > t] = 1 - P[T_m < t] = 1 - \{e^{-\hat{\lambda}} * (1 + \hat{\lambda} + \frac{\hat{\lambda}^2}{2!} + \ldots + \frac{\hat{\lambda}^t}{t!})$$

Where $\hat{\lambda}$ = The maximum likelihood estimator of the parameter λ of the Poisson Distribution. Once the p-value is obtained, you reject or fail to reject the H_0 as standard for hypothesis testing.

Works Cited

1. Leung, Ming-Ying et al. “Nonrandom clusters of palindromes in herpesvirus genomes.” Journal of computational biology : a journal of computational molecular cell biology vol. 12,3 (2005): 331-54. doi:10.1089/cmb.2005.12.331
2. Leung MY., Yamashita T.E. (1999) Applications of the Scan Statistic in DNA Sequence Analysis. In: Glaz J., Balakrishnan N. (eds) Scan Statistics and Applications. Statistics for Industry and Technology. Birkhäuser, Boston, MA. https://doi.org/10.1007/978-1-4612-1578-3_12

3. I-Ping Tu, Shao-Hsuan Wang, Yuan-Fu Huang "Estimating the occurrence rate of DNA palindromes," The Annals of Applied Statistics, Ann. Appl. Stat. 7(2), 1095-1110, (June 2013)