LOCATING PALINDROME CLUSTERS IN DNA

Navin Souda
University of California, San Diego
nsouda@ucsd.edu

Evan Velasco
University of California, San Diego
emvelasc@ucsd.edu

Shu-Ho Yang
University of California, San Diego
shy096@ucsd.edu

Yu-Ming Lin
University of California, San Diego
yul857@ucsd.edu

Andrew Chavez
University of California, San Diego
adchavez@ucsd.edu

1 Introduction

The human cytomegalovirus (CMV), commonly known as human herpesvirus-5, can be a life threatening disease when affecting individuals with suppressed or compromised immune systems. A common goal for many researchers in the medical field is to develop methods for combating the virus and to help at- risk individuals. Some research has lead scientists to believe the way to achieve this is through studying the way in which the virus replicates. More specifically, scientists are in search of a unique location on the virus' DNA sequence that may hold the code for its self replication, referred to as the origin of replication. Similar to humans, the virus' DNA contains the instructions required for it to form, grow, survive, and replicate itself. The human DNA string is comprised of four nucleotides: Adenine (A), Cytosine (C), Thymine (T), and Guanine (G). From these four base pairs, patterns can be made to expand the amount of information that can be stored with the pairing of A-T and C-G. For example, complementary palindrome is one type of pattern in which in on one strand of the DNA reads five-prime to three-prime forward on one and then matches the sequence to create a double helix. It is believed that some patterns in particular sites may be important in the virus' replication, so they are flagged and examined more closely.

The herpes virus can be categorized into different families. Within the family of CMV, two viruses that are similar to one another are marked by complementary palindromes. The first, Herpes simplex virus, is marked by a long palindrome that is comprised of 144 letters. The second, Epstein-Barr virus, is marked by several short palindromes that closely repeat clustered at the origin of replication. Meanwhile for CMV, the longest palindrome is 18 base pairs, with a total of 296 palindromes between 10-18 base pairs long. It is hypothesized by biologists that clusters of palindromes within CMV may play the same role as the single long palindrome found in Herpes simplex or as in the short repeat cluster in Epstein-Bar virus.

When analyzing the origin of replication encoded in CMV, its DNA is spliced into segments and tested to see if it can replicate itself. If the spliced set of pair cannot replicate itself, then the scientists conclude that the origin of replication must not reside in that sequence. Since there are many different ways to divide up the DNA strand, the process of analyzing each possible sequence is very time consuming. This means a significant loss in money if the work the scientists put in leads them to another dead end. To narrow down this search and potentially reduce the amount of tests needed to find the origin, statistical investigations are done to identify unusually dense clusters of palindromes.

In this investigation, we will conduct a statistical analysis on the DNA sequence in the CMV to identify any unusual clusters of complementary palindromes. To do so, we must ask how we can devise a method to find these unusual clusters and how can we determine if a given cluster is a possible site of the origin of replication. This investigation will then provide an informed suggestion to those scientists investigating CMV and potentially narrow down their search for the origin of replication.

2 Data

In 1980, the DNA sequence for CMV was published by Chee et al. and made publicly accessible for further investigation. Its DNA was found to be 229,354. Over a decade later in 1991, Leung et al decided to implement a search algorithm in the investigation with the goal of screening the sequence for the different types of patterns. With this implementation, 296 palindromes were found that had at least 10 letters in them, any found with less than 10 letters were ignored. The longest of which were 18 letters long and occurred in locations 14719, 75812, 90763, and 173893.

When studying the group of data obtained from the 296 palindromes, the DNA is spliced into different segments with varying size and combinations of nucleotides. Data suggests that independent of the length of interval; There appear to be clusters of palindromes in at least two locations. The clusters are thought to be around the 93,000th and 195,000th pairs of DNA. From this, scientists infer that the clusters at these two flagged locations are exceptions within the typical DNA structure and thus not due to chance. Histograms of the actual palindromes are then compared to randomly generated numbers with no extrapolated patterns in clusters at any given point.

The data used for this investigation includes a list of the locations that the 296 palindromes in the human cytomegalovirus DNA.

3 Background

CMV affects anywhere from 30%-80% of the population depending on the location. This virus is also able to enter an individual and lie dormant for a long period of time while not presenting symptoms. Although the virus can remain harmless for most people, for those with compromised immune systems once it enters productive cycles it can reactivate illnesses. It is also estimated that 10%- 15% of children come into contact with the virus before the age of five. It is thus crucial for scientists to find the origin of replication for the virus, and learn how to stop its reproduction to help save these individuals' lives [2].

For most individuals, coming in contact with the human cytomegalovirus is not life threatening, or particularly harmful to them. In extreme cases CMV can put an individual's life in serious jeopardy if they have a compromised immune system. For instance, when someone receives an organ transplant the patient is put on an immunosuppressant to reduce the risk of the body rejecting the organ. During this period of time, a patient is vulnerable to having potentially fatal complications if they come in contact with CMV. Interestingly, studies in the human cytomegalovirus have begun to indicate that these infections have some sort of association with malignant glioma, the most common primary brain tumor. Malignant gliomas are generally a rapid growing tumor that is most often fatal, despite current treatments. It is also believed CMV can be reactivated under certain conditions of inflammation and immunosuppression. CMV gene products can interfere with and dysregulate multiple cellular pathways, which can be deadly for an individual that has a malignant glioma. Research at the University of Alabama stated that they found a high percentage of malignant gliomas are infected with CMV, and the harmful gene products are expressed in the tumors. These findings suggest that CMV may potentially play an active role in glioma pathogenesis [1].

The method used in the process of researching CMV is not unique to this virus, in fact many retrovirus research have adopted this method. For example, when studying the human immunodeficiency virus, or HIV, scientists have examined a large number of integration target site sequences and found that a statistical palindromic consensus centered on the virus-specific duplicating target site sequence [3]. They are able to do this with the same method as CMV to cut the DNA into segments and analyze them. They too, use statistical methods to help narrow down their workload and further exemplify the usefulness of statistical analysis.

4 Investigation

The goal of our investigation is to use a Poisson process to locate any "large" and "unusual" clusters in the DNA. These clusters would be the most likely locations for the place of origin of human CMV.

4.1 Creating a Random Scatter

We will begin our investigation by creating a uniform random scatter. The purpose of this plot is to create a standard in which to compare our observed DNA data to. In order to achieve this random scatter, we take a sample of size 296 from a discrete uniform distribution from the interval [0, 229,354]. The sample is done without replacement because we cannot have multiple palindromes at the same location. The distribution of the sample is shown below.

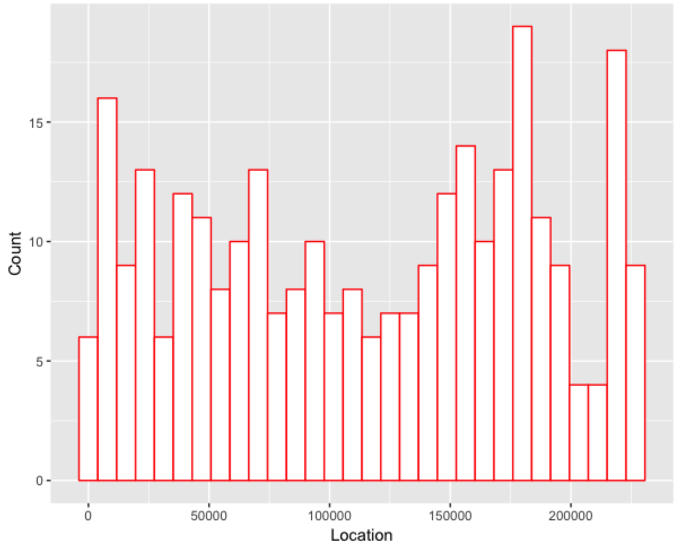

Figure 1: Distribution of the location of our simulated palindromes.

4.2 Locations and Spacing

Now, we will compare the locations found in our observed data with those from our simulated data. Looking at the histogram below, we see that both have seemingly random peaks and troughs. Overall, they don't look too different.

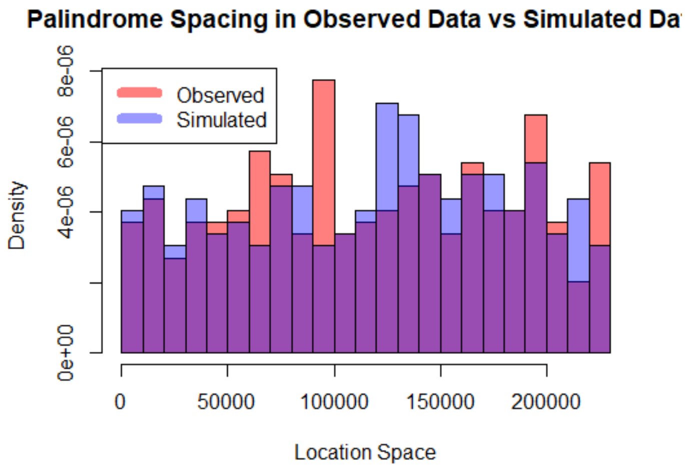

Figure 2: Histograms comparing distribution of locations in observed and simulated data

We can also consider quantile-quantile plots to determine whether or not these locations have similar distribution. We can see that for the most part the data follows the straight line, except for some curvature in the middle. Thus, it seems the locations of palindromes are for the most part similar in our observed data to that of a truly random scatter.

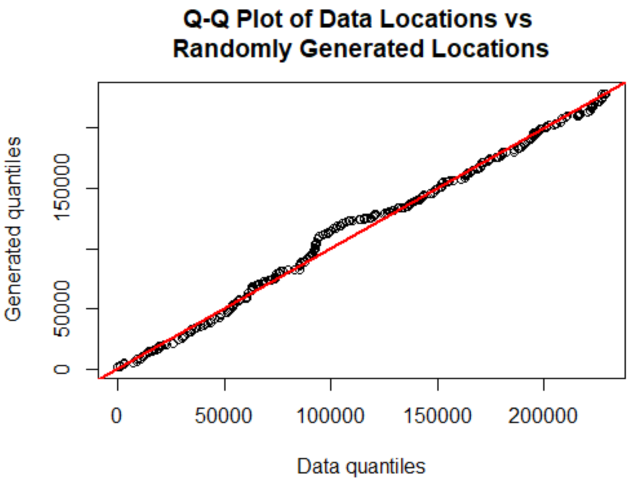

Figure 3: Plot comparing quantiles of the locations of the true data with the randomly generated data.

Next, we can compare the spacing between our observed data and simulated data by overlaying the two histograms and analyzing their trends. We can see that the observed data seems to have slightly heavier tails. There are more points concentrated at the lower end, as well as points with much higher values than those in the simulated data.

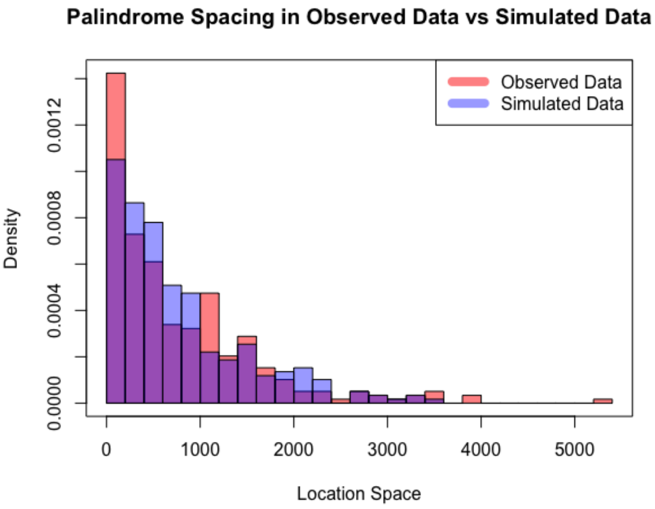

Figure 4: The graph above compares the spacing between palindromes found in our observed data with the spacing of palindromes in our simulated data.

Like before, we can also consider the similarity (or lack of it) through the use of quantile-quantile plots, shown below. We see that the observed data has a much heavier upper tail than the simulated data, indicating there are more large gaps between palindromes than one might expect in a truly random scatter.

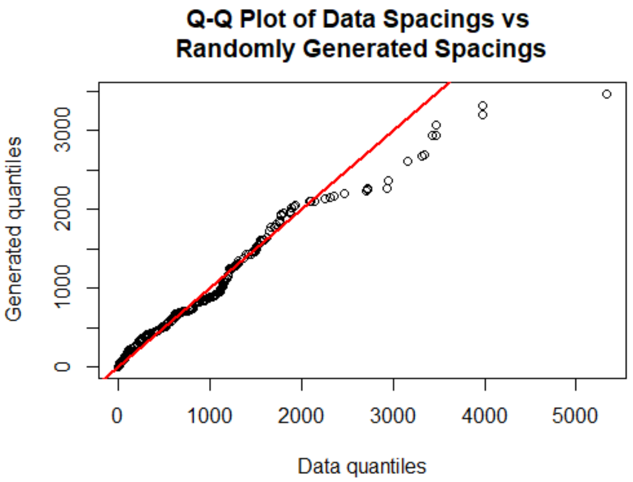

Figure 5: Plot comparing quantiles of the spacing of the true data with the randomly generated data.

A quick visual analysis leads us to believe that the palindromes in our observed data and simulated data are similarly distributed in location, but not necessarily in spacing. This may be something to keep in mind when partitioning the sequence into intervals, as spacing can affect how clusters are seen in the data. Finally, we can look at the distribution of spacing between consecutive pairs and triplets below.

4.3 Counts

We will now consider the number of palindromes when separating the DNA into 60 non-overlapping regions of equal length. If the palindromes are truly randomly scattered, we would expect the counts of the regions to have a Poisson distribution with rate parameter λ. So, we want to compare the counts found in our observed data with the counts we would expect from a Poisson distribution. A natural way to do this is to perform a chi-square goodness of fit test. Since we don't know λ, we can estimate it as $\bar{x} = \frac{296}{60} = 4.9\bar{3}$. Then we can make a table of the number of intervals we observed with a certain count (or range of counts), as well as the number of intervals we would expect with these counts given our estimate $\hat{\lambda}$. This table is shown below.

Num. Palindromes	Observed	Expected
0-2	10	7.82
3	9	8.65
4	9	10.67
5	10	10.52
6	5	8.65
7	10	6.10
8+	7	7.59

Figure 6: Observed number of intervals containing certain amounts of palindromes and expected number of intervals containing those many palindromes

Using these values we can then calculate a chi-square statistic, which we find to be 4.99. This test statistic has approximately the χ^2 distribution with 7 - 1 - 1 = 5 degrees of freedom. This gives us a p-value of .417, meaning the probability we would see something at least as large as this under our null (i.e. random scatter) is .417. Thus, we fail to reject the null, indicating our observed counts of palindromes may follow a Poisson distribution.

4.4 The Biggest Cluster

Finally, we would like to decide whether or not the region which contains the largest cluster of palindromes could contain a potential origin of replication. That is, we want to see whether or not our maximum value for a cluster (let us call it c_{max}) is out of the ordinary when compared to a truly random scatter. As stated before, a random scatter would indicate that the number of palindromes in a region has a Poisson distribution. Using this knowledge, we can use the maximum observed palindrome count as a test statistic and get a p-value explaining the probability of getting this count as out maximum under the hypothesis of random scatter. The derivation of this p-value is explained in the theory section. It is calculated as $1 - (P(\text{one interval count} < c_{max}))^{60}$. In our case, $c_{max} = 14$, thus using our previous estimate for λ, we get a p-value of 0.0362. Thus, we have strong evidence indicating that this cluster is not a result of random scatter, and for that reason may be a good place to look for the origin of replication.

4.5 Conclusion and Recommendation

I would recommend that a researcher looking for the origin of replication look at the sequence between the 92,000th base pair and the 96,000th base pair. It is in this region that we found an abnormally large cluster of palindromes, 14 to be exact. We have checked that our data seems to follow for the most part a random scatter in terms of palindromes, and thus we found that the probability of getting a cluster of 14 palindromes in a interval of that size is quite unlikely. Thus, this may indicate that the origin of replication is in this area.

4.6 Additional Hypothesis

In addition to the work we have done, we want to have some quantification of the difference that region length makes in the tests we performed, and how it affects our findings. It should be clear that too small of an interval length will lead to us breaking apart clusters of palindromes, thus not being able to locate them. On the other hand, intervals that are too large will lose information as well by not taking into account gaps between clusters. We will test both of these to verify our ideas. First we will repeat our previous procedures of chi-square testing and testing for the maximum cluster size with 20 regions and then with 100 regions.

Num. Palindromes	Observed	Expected
0-12	7	5.69
13-16	7	7.97
16+	6	6.33

Figure 7: Observed number of intervals containing certain amounts of palindromes and expected number of intervals containing those many palindromes when there are 20 regions.

We tabulate the palindrome counts for 20 regions and create the chi-square table above. Already there is an issue. In order to maintain expected counts of greater than 5, we have to condense our table to only 3 bins. This is a problem in more ways than one. First of all, it makes our data seem much more homogeneous than it may actually be, in the same way that a histogram with only 2 or 3 bins likely won't accurately represent the density. Secondly, this causes issues because our chi-square test will have only 1 degree of freedom. This means the accuracy of our chi-square approximation is worse, and our test is less stable. Computing the test statistic, we get a value of .438, which corresponds to a p-value of .508, which would mean we fail to reject the null that the counts have Poisson distribution. Our maximum palindrome count was 23. Using the same test as before to calculate the probability of getting this as out maximum, we get a p-value of .290, indicating that this is not such a departure from the norm. Indeed, at a 0.05 level, we would say that this is not an abnormal maximum at all. This shows us the second problem with small bin size - it is very likely to get large "clusters" of palindromes because even palindromes with a lot of space between them will be considered part of the same cluster.

Num. Palindromes	Observed	Expected
0	46	45.53
1	78	67.38
2	33	49.86
3	30	24.60
4+	13	12.63

Figure 8: Observed number of intervals containing certain amounts of palindromes and expected number of intervals containing those many palindromes when there are 200 regions.

Now we will do the same when dividing the sequence into 200 regions. We see the chi-square table above. Our issue here is almost the opposite of above. In this case, we see only a few different values of counts. Though it is not seen in this table, we have 9 intervals with 4 palindromes, 2 intervals with 5, 1 with 7 palindromes and 1 with 9 palindromes. That is, there is not very much variation in the cluster sizes and furthermore, due to the nature of the Poisson distribution, even slight departures from the average value will seem significant. Indeed, when using our maximal value test, we get a probability of 0.005 that we see a value as extreme as 9 in the circumstances. This means we are more and more likely to reject the null as m increases, simply because lambda will go down, and if the maximum does not decrease at a similar rate, then the probability that we see such a maximum will be very low. This may make our data seem as though it is not Poisson distributed, and looking at the chi-square test, this is verified by a p-value 0.035. Again, though, since our bin sizes are so large and encompass so much space, this may not be the greatest indicator of fit.

5 Theory

5.1 Creating a Random Scatter

We model the random scatter of palindromes through something called a homogeneous Poisson process. This process has 3 distinct characteristics, from which all of its other properties can be derived. The first is that palindromes occur at some rate, λ, and this rate is invariant at every point in the DNA sequence. The second characteristic is that the number of palindromes that occur in two separate regions are independent of each other, and the final characteristic is that we cannot have 2 palindromes occur at the same point in the sequence, which makes logical sense as well.

Using this knowledge, we can now derive some properties of the Poisson process. The first is that the number of palindromes in a region of unit length have Poisson(λ) distribution. Secondly, the spacing between hits has the exponential distribution with parameter λ. This is derived fairly easily from the following logic:

$$\begin{aligned}&P(\text{distance between first and second hit} < t)\\ =&P(\text{the interval of size t has no hits})\\ =&e^{-\lambda t}\end{aligned}$$

The last step follows from the fact that the number of hits a region of size t (that is, t times the unit length) is distributed as Poisson(λt). One final property we have is that the location of the palindromes (over the whole length of the sequence) are uniformly distributed. It is this property that we utilize when generating our own random scatter model of palindromes. We simply sample without replacement from a vector going from 1 to 229,354 (the length of our DNA sequence) 296 times to match the number of palindromes we found in our observed data. We use these numbers from the sample as the locations of our palindromes, and we can then compare this with our observed data to look for abnormalities.

5.2 Locations and Spacing

As stated in the previous section, if our data was truly randomly scattered, we would expect the locations of our data to be uniformly distributed over the length of the sequence, and we would expect the spaces between palindromes to be exponentially distributed. After generating our randomly scattered data, we can compare the distributions of spacing and location with that of our observed through our usual graphical methods - histograms and quantile-quantile plots. In histograms, we look for a similarity in density, that is, that there are no large differences in the histograms of the two, be that large peaks, long tails, or apparent differences in center. In quantile-quantile plots, if the two data sets follow the same distribution, we expect to see the points follow the line $x = y$. Thus, if we notice significant deviations from that, it indicates a difference in distribution. Possible differences include curvature, a difference in slope, or skewing of the tails.

5.3 Counts

By the properties of the Poisson process, we expect the counts in each region of the sequence to be i.i.d Poisson(λ). Unfortunately, we can never know the true λ, thus our best choice is to estimate λ by some $\hat{\lambda}$. There are a couple of common techniques for parameter estimation. We will first discuss maximum likelihood estimation (MLE). Maximum likelihood estimation, as the name indicates, determines an estimate for a parameter by maximizing the likelihood function (often times what is maximized is the log of the likelihood function, but the result is the same). The likelihood function for some random sample $x_1, ..., x_n$ from random variable X with density or mass function $f_X(x)$ is

$$\prod_{i=1}^{n} f_X(x_i)$$

Thus, maximization amounts to taking the derivative and finding the root of the function (when possible). The derivation of the MLE for λ for the Poisson distribution is shown below.

$$\frac{\partial}{\partial\lambda} l(\lambda) = \frac{\partial}{\partial\lambda}[\sum_i x_i log(\lambda) - n\lambda - \sum_i log(x_i!)]$$

$$= \sum_i \frac{x_i}{\lambda} - n = 0$$

Solving the last equation for λ, we get $\hat{\lambda} = \bar{x}$. Thus, the MLE for λ in a Poisson distribution is simply the sample mean. There are a few properties of MLEs that often make them desirable estimators. While some MLEs may be biased, all MLEs are consistent - that is, as sample size approaches infinity, the value of the MLE is guaranteed to converge to the true value of the parameter. Furthermore, the MLE is similarly asymptotically efficient, in that it achieves the Cramer-Rao lower bound on its variance as the sample size grows to infinity. Finally, the MLE is asymptotically normally distributed. In our case, our $\hat{\lambda}$ has an asymptotic normal distribution with mean λ and variance $\frac{1}{nI(\lambda)}$, where $I(\lambda)$ is the Fisher information matrix. Fisher's information matrix is defined as $I(\lambda) = E[\frac{\partial}{\partial\lambda} log f_\lambda(x)]^2 = -E[\frac{\partial^2}{\partial^2\lambda} log f_\lambda(x)]$. We can use this to make an asymptotic confidence interval for λ, that being, at a 95% confidence level, $\hat{\lambda} \pm 1.96\sqrt{nI(\lambda)}$.

The second form of estimation we will discuss is method of moments estimation. This method starts by finding, for some random variable X, $E[X]$ in terms of the parameters of the distribution. Then, solve one of the parameters in terms of $E[X]$. Next, we find the estimator of the parameter by replacing $E[X]$ with $\hat{x}$. Using the Poisson distribution as an example, we have that $E[X] = \lambda$. Then we just replace λ by $\hat{\lambda}$ and $E[X]$ by $\bar{x}$, and find that $\hat{\lambda} = \bar{x}$, the same as the MLE.

Now that we have an estimator for the rate of our Poisson process, we want to check if our counts actually have the Poisson distribution with this rate parameter. To do this, we use the chi-square goodness-of-fit test. To utilize this test we first split the DNA sequence into regions of equal sizes, and count the number of palindromes in each region. Then, we tabulate how many sequences have a certain count, that is, how many intervals have no palindromes, how many have 1, how many have 2, and so on. We then calculate the expected number of intervals for each count under the null hypothesis that the counts are Poisson distributed. We do this by calculating the probability that we would get a certain count in a Poisson($\hat{\lambda}$ distribution, and multiplying it by the total number of regions. For example, as we split the sequence into 60 equal regions, giving us a of our expected count for 0 would be $P(X = 0) * 60 = (60)\frac{4.9\bar{3}^0}{0!}e^{-4.9\bar{3}} = .4321$. So the expected number of intervals with count 0 is .4321. Then, the final step in the preparation is to combine counts such that we have all the expected counts $\geq$ 5, as this is required for the stability of the chi-square test. Thus, we combined counts 0, 1 and 2 into one bin, and counts of 8 and above into another. Now, we are ready to compare the chi-square test statistic, which is simply

$$\sum_{k=1}^{m} \frac{(E_k - O_k)^2}{E_k}$$

where we have bins $1, 2, ..., m$, with bin k having expected count E_k and observed count O_k. Under the null hypothesis, that is, the palindromes are randomly scattered (meaning the palindrome counts have Poisson distribution), the test statistic has approximately χ^2 distribution with 7 - 1 - 1 = 5 degrees of freedom. The 7 comes from the number of bins, one is always removed automatically, and then one more is removed because we had to use the data to estimate the rate parameter. The value of our statistic came out to 4.99, and so we can finally calculate a p-value by looking at P($\chi^2_5 \geq 4.99$) = .417. This indicates to us that we cannot reject the null, and that it's possible the palindrome counts per region have Poisson distribution. Checking the residuals (shown in the graph below), we see they are all less than 1, indicating that none of our bins show a distinct lack of fit.

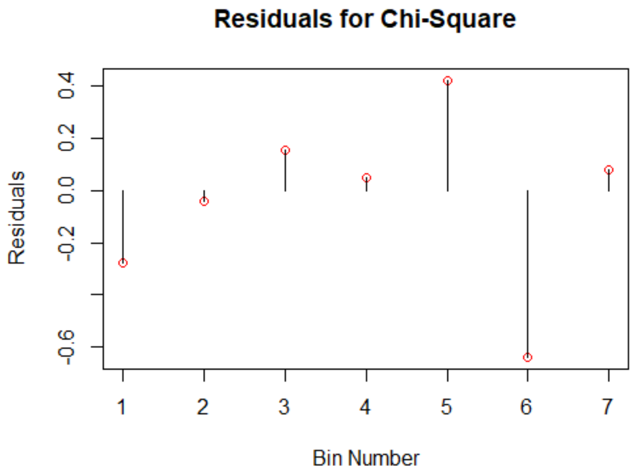

Figure 9: The graph above shows the residuals for each bin of our data. 1 corresponding to the bin 0-2, and 7 corresponding to 8+, with individual numbers in between them.

We will use a hypothesis test to classify whether a cluster of palindromes is significant. The null and alternative hypothesis are shown below.

$$H_0 : \text{There is not an unusually large cluster}$$

$$H_A : \text{There is an unusually large cluster}$$

Now, we must define what it means to be considered unusually large. In order for a cluster to be potentially significant, it must be larger than the max of an i.i.d. Poisson distribution. We will define the max count over 60 intervals as our test statistic. The process for obtaining our p-value for our defined test statistic is shown below.

$$\begin{aligned}
&P(\text{maximum count over 60 intervals} \geq k)\\
=&P(max(Y_1, ..., Y_{60}) \geq k)\\
=&1 - P(max(Y_1, ..., Y_{60}) < k)\\
=&1 - P(\text{all } Y_i \text{ are} < k)\\
&ie : (Y_1 < k) \cap (Y_2 < k) \cap ... \cap (Y_{60} < k)
\end{aligned}$$

Since our events are independent from one another, we can apply the multiplication rule of probability.

$$\begin{aligned}
=&1 - P(Y_1 < k) * (Y_2 < k) * (Y_3 < k) * ... * (Y_{60} < k)\\
=&1 - [P(Y_1 < k)]^{60}\\
=&1 - [\lambda^0 e^{-\lambda} + ... + \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda}]^{60}
\end{aligned}$$

In general, this equation can be written as $1 - [P(Y_1)]^m$ or $1 - [\lambda^0 e^{-\lambda} + ... + \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda}]^m$, where m is a positive number representing the number of intervals the data is split into.

References

[1] C. S. Cobbs, L. Harkins, M. Samanta, G. Y. Gillespie, S. Bharara, P. H. King, L. B. Nabors, C. G. Cobbs, and W. J. Britt. Human cytomegalovirus infection and expression in human malignant glioma. *Cancer Research*, 62(12):3347–3350, 2002. ISSN 0008-5472. URL http://cancerres.aacrjournals.org/content/62/12/3347.

[2] C. for Disease Control. Cytomegalovirus (cmv) and congenital cmv infection. URL https://www.cdc.gov/cmv/index.html.

[3] X. Wu, Y. Li, B. Crise, S. M. Burgess, and D. J. Munroe. Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. *J Virol*, 79(8):5211–5214, Apr 2005. ISSN 0022-538X. doi: 10.1128/JVI.79.8.5211-5214.2005. URL https://www.ncbi.nlm.nih.gov/pubmed/15795304. 15795304[pmid].