# Lab02 - CMV DNA Analysis

*Jonathan Stuart*

*3/21/2018*

## Introduction

Human cytomagalovirus (HCMV) is a disease that effects somewhere between 30% and 80% of the population, varying geographically. Though the virus lays dormant once it infects a host, it can become harmful when it enters a productive cycle, replicating many thousands of copies of itself [1]. In this productive phase, it is most harmful to individuals with chronically weakened immune systems, and, additionally, is the leading cause of mental retardation and deafness at birth [2]. Being part of the Herpes family, biologists have hypothesized that patterns in the distribution of the virus' DNA base pairs might give insight into the location of the site of replication. In an effort to mitigate the effects of this particular disease, we are faced with the problem of studying the distribution of base pair palindromes in order to locate the virus' replication site. If we are able to develop methods to efficiently identify the locations of palindromes, we will have given researchers a reliable approach toward controlling the disease by impeding its spread and treating those already infected.

Along a strand of HCMV DNA, we have observed the existence of 229, 354 DNA base pairs. It has been further observed that, within these base pairs, 296 palindromes can be found, each between 10 and 18 base pairs long. Given that two other viruses in the Herpes family have replication sites marked by abnormal characteristics of base pair palindromes, research is underway to determine if HCMV does as well. To assist you in your research, we will determine if clusters of palindromes are non-random, and thereby indicative of DNA replication sites. In the process of this investigation, we will answer the following questions:

1. How do we find clusters of palindromes?

2. How do we determine whether a cluster is a chance occurrence or an indication of a replication site?

## Methodology

In order to answer these questions, we used inferential statistics to find a distribution that fit the data well. Once we had a distribution that the data could likely have come from, we examined whether or not the area with the maximum number palindromes could have been a chance occurrence or not. From this, we made our determination of where researchers should look for clusters of palindromes and, by implication, DNA replication sites.

We used intervals of length 500, 2500, 4000, and 7000 base pairs which divided the data set into 458, 91, 57, and 32 intervals, respectively. This process was important due to the fact that the interval size had an impact on the goodness of fit of the Poisson distribution to the data; it was important to choose an interval size that was neither to large, nor too small. Once the data was partitioned into intervals of these sizes, we counted the number of palindromes in each interval. From these counts, we were then able to conduct a series of hypothesis tests to determine how well the Poisson distribution could describe the data we had been given.

Starting with the chi-squared goodness of fit test, we treated each interval size and corresponding palindrome count to determine how closely the Poisson distribution fit the given data. We then calculated residual values and again sought to answer the question of how closely the Poisson distribution fit the given data. Next we constructed 95% confidence intervals for the parameters $\lambda$ for each of the 4 intervals. From there, we examined the inter-arrival distances of hits of palindromes and asked how closely they adhered to a exponential and gamma distributions. Finally, we examined the p-value for the maximum number of palindromes counts, which allowed us to answer the question of whether or not a cluster of palindromes was significant for identifying a DNA replication site.

## Results

First we needed to get counts of palindromes for each of the interval sizes mentioned above, 500, 2500, 4000 and 7000. We can see a table of those counts here.

Then, assuming applicability of the Poisson distribution, we can take $\hat{\lambda}$ to be $\bar{x}$, an unbiased estimate for $\lambda$. We then created vectors of observed and expected values, which can be seen below for each of the intervals.

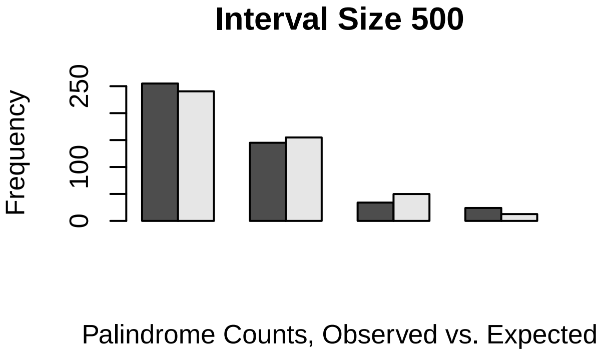
### Tables of Expected and Observed Counts

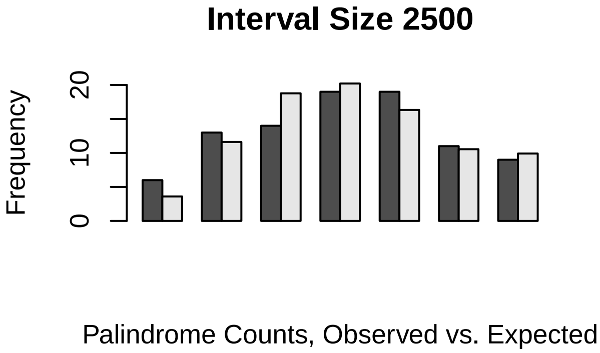
| Interval Size | Expected Count | Observed Count |
|---|---|---|
| 500 | 241 | 255 |
| | 155 | 145 |
| | 50 | 34 |
| | 13 | 24 |

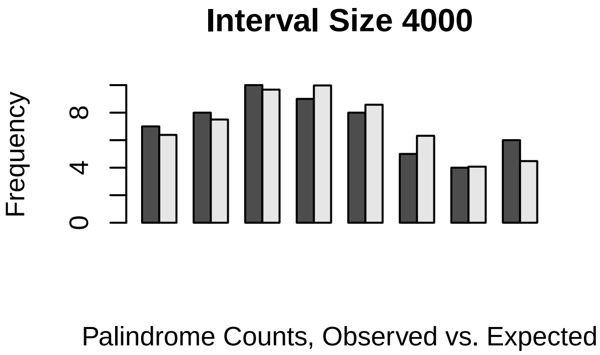
| Interval Size | Expected Count | Observed Count |
|---|---|---|
| 2500 | 4 | 6 |
| | 12 | 13 |
| | 18 | 14 |
| | 20 | 19 |
| | 16 | 19 |
| | 11 | 11 |
| | 10 | 9 |

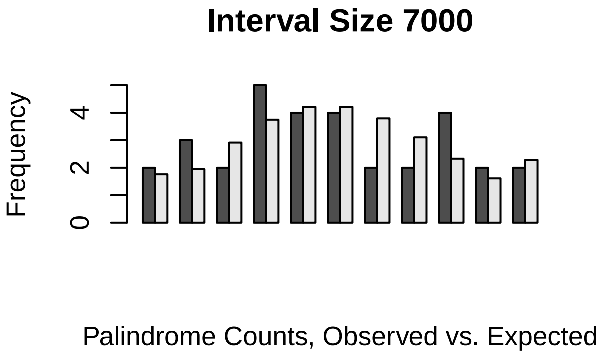
| Interval Size | Expected Count | Observed Count |
|---|---|---|
| 4000 | 6 | 7 |
| | 8 | 8 |
| | 10 | 10 |
| | 10 | 9 |
| | 9 | 8 |
| | 6 | 5 |
| | 4 | 4 |
| | 4 | 6 |

| Interval Size | Expected Count | Observed Count |
|---|---|---|
| 7000 | 2 | 2 |
| | 2 | 3 |
| | 3 | 2 |
| | 4 | 5 |
| | 4 | 4 |
| | 4 | 4 |
| | 4 | 2 |
| | 3 | 2 |
| | 2 | 4 |
| | 2 | 2 |
| | 2 | 2 |


Interval Size 500


Interval Size 2500


Interval Size 4000


Interval Size 7000

With these counts, we then conducted the chi-square goodness of fit test for the Poisson distribution with parameter $\hat{\lambda}$.

**Table of Chi-Square Values**

| Interval Size | Chi-Square Value |
|---|---|
| 500 | 0.99988 |
| 2500 | 0.84980 |
| 4000 | 0.96025 |
| 7000 | 0.93652 |

Here, the closer the values are to 1, the better the indication of fitness of the Poisson distribution to the observed data. From these data it appears that 4000 would be the optimal interval size to use.

Then, looking at the standardized residuals to see more closely where the fit occurs for each of the observed count totals, we found that the interval size of 4000 showed the best results, as residual values less than three in absolute value indicate goodness of fit. Here, we have a table to average residual values.

**Table of Average Residual Values**

| Interval Size | Average Residual |
|---|---|
| 500 | 1.79 |
| 2500 | 0.6 |
| 4000 | 0.3 |
| 7000 | 0.5 |

We then generated the following 95% confidence intervals for the parameters $\lambda$ for each sized interval of base pairs. From these data, the tightest confidence interval appears to come from the 500 base pair interval.

**Table of Confidence Intervals**

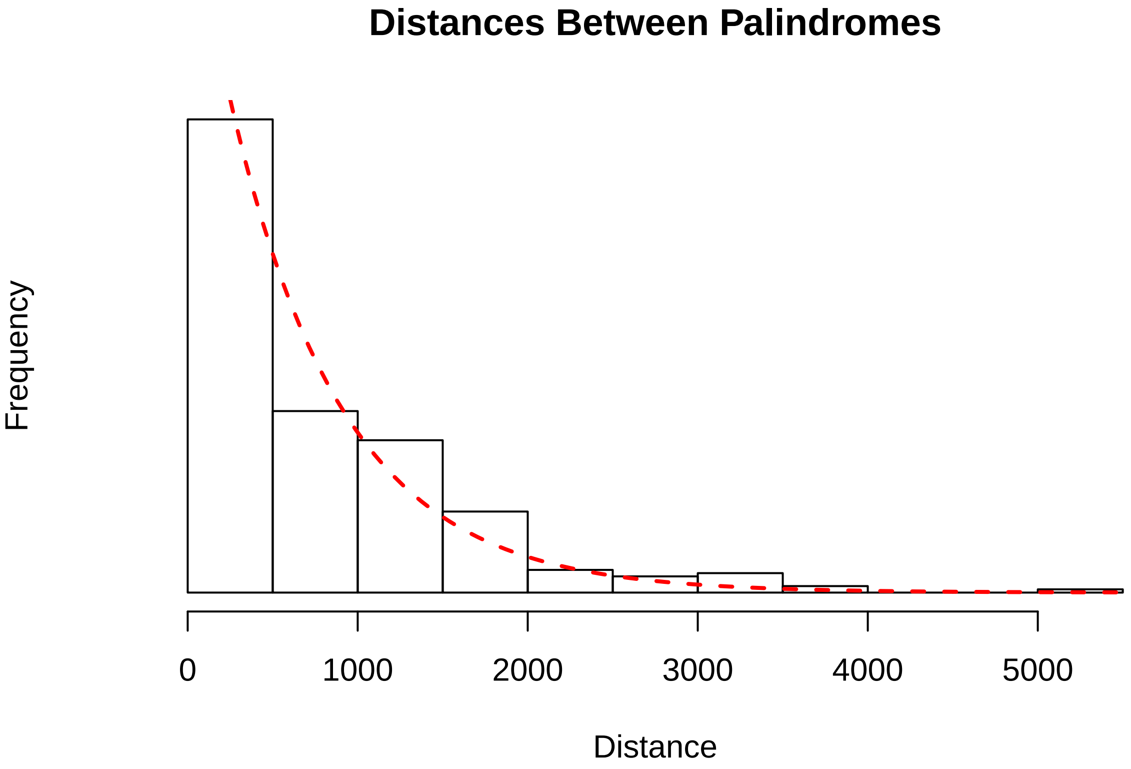
| Interval Size | Confidence Interval |
|---|---|
| 500 | 1.112293 |
| 2500 | 2.491117 |
| 4000 | 3.147584 |
| 7000 | 4.157788 |

Looking then toward the likelihood that the maximum palindrome count for each interval was at least as large as the highest count, we came up with the following p-values.

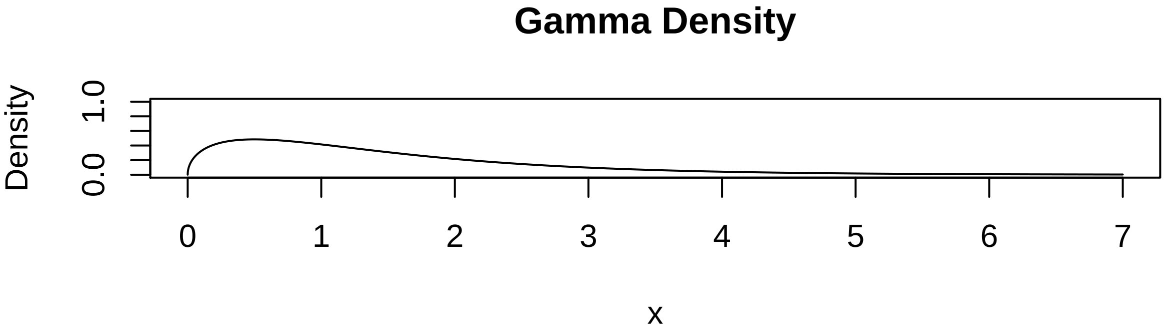
**Table of Maximum Count P-values**

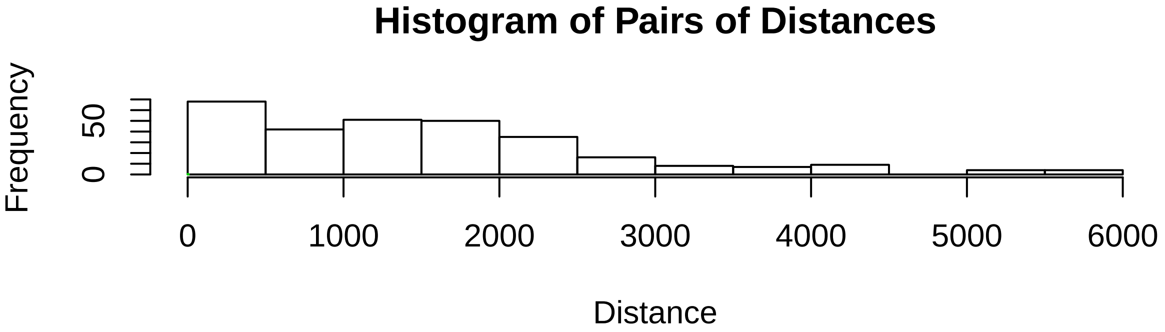
| Interval Size | Confidence Interval |
|---|---|
| 500 | 0.0001904997 |
| 2500 | 0.003140679 |
| 4000 | 0.05190829 |
| 7000 | 0.1603417 |

Next we examined the distance between hits, as the inter-arrival time should approximate the exponential distribution if the Poisson distribution governs the data.



We also examined pairs, triplets, etc. of distances to inquire after the fitness of the gamma distribution. Here we have plots for pairs of distances.





Aggregating these test results into a single metric, we determined that the Poisson distribution does effectively describe the data when considering intervals of 4000 base pairs.

## Discussion

Taking the results of all of these hypothesis tests into account, we decided upon an interval of 4000 base pairs, and made the conclusion that the Poisson distribution did in fact fit the given data. Though the observed and expected counts were quite close for all interval sizes, it was through the chi-squared goodness of fit tests and the associated residual values that the 4000 base pair interval showed its viability. And with regard to the maximum count p-value, this test statistic is the most straightforward about the usefulness of this approach in solving the questions laid out in the introduction. From the certainty of the locations of maximum counts that the maximum count test statistic reveals, we have our clearest guidance to researchers seeking DNA replication sites by identifying areas of palindrome clustering.

One limitation, however, was the fact that the exponential distribution fit the inter-arrival time of palindromes, while the gamma distribution did not fit pairs of inter-arrival distances. This could be due to the clustering of palindromes interfering with and obfuscating arrival times, and could also be due to the fact that the Poisson distribution is not an **exact** fit for the data. Assuming the Poisson and then estimating $\lambda$ could be problematic in that we are conducting our analysis with an incorrect parameter $\hat{\lambda}$.

## Conclusion

In summary, we have taken the two questions posed at the outset and carried out a statistical analysis of the given data. From that analysis, we have determined that the Poisson distribution does accurately account for the observed data, and that an interval size of 4000 base pairs is a useful division of a strand of HCMV DNA. Our analysis tells us that the likelihood of identifying clusters of palindromes at least as large as the expected and observed maximum counts is very high, thus providing reliable guidance to researchers as they attempt to better understand HCMV replication.