Case Study 3

0. Author Contributions

Both authors (Felipe Lorenzi and Anna Preger) contributed equally to the writing of this report. They worked on it simultaneously and both typed and coded similar amounts in all sections of the report.

1. Introduction

In this report we are trying to study the origin of DNA replication by looking for unusual clusters of palindromes within a DNA sequence. Our dataset comprises the DNA locations of the 296 palindromes (at least 10 letters long) found in the Cytomegalovirus DNA molecule, which contains 229,354 complementary base pairs.

The main goal of this analysis is to investigate the distribution of palindromes across the CMV DNA sequence. In this analysis, we use numerical summaries and graphical methods to describe the distribution of palindromes across our dataset, as well as across 10 simulations of uniform distributions of 296 palindrome locations (out of 229,354).

We first compare their distributions and counts graphically, finding discrepancies which led us to further investigate. These findings suggest that a goodness of fit test could be pertinent to this data.

Our null hypothesis is that the palindromes are uniformly distributed across the intervals of base pairs in the virus' DNA, following a Poisson model. In addition to numerical summaries and graphical methods, we also perform a Pearson's Chi-squared test to further investigate if the Poisson Model is a good fit for the data.

We also calculated the distances between palindromes and groups (pairs and triples) of palindromes, in order to look for unusual clusters that might be found in the given DNA sequence.

Furthermore, we looked at the interval containing the largest number of palindromes in our dataset and recorded that value. We did this over eight different interval lengths. We did the same for our 10 simulations and compared those values with the real ones, finding that every interval containing the largest number of palindromes in our dataset had higher values than the corresponding interval in the simulations.

2. Basic Analysis

2.1 How Randomly are the 296 Palindromes Allocated?

Methods

In order to determine if our sample has unusually dense palindrome clusters or not, we decided to generate 296 random locations (uniformly distributed) along a DNA sequence of the same size as CMV's DNA sequence.

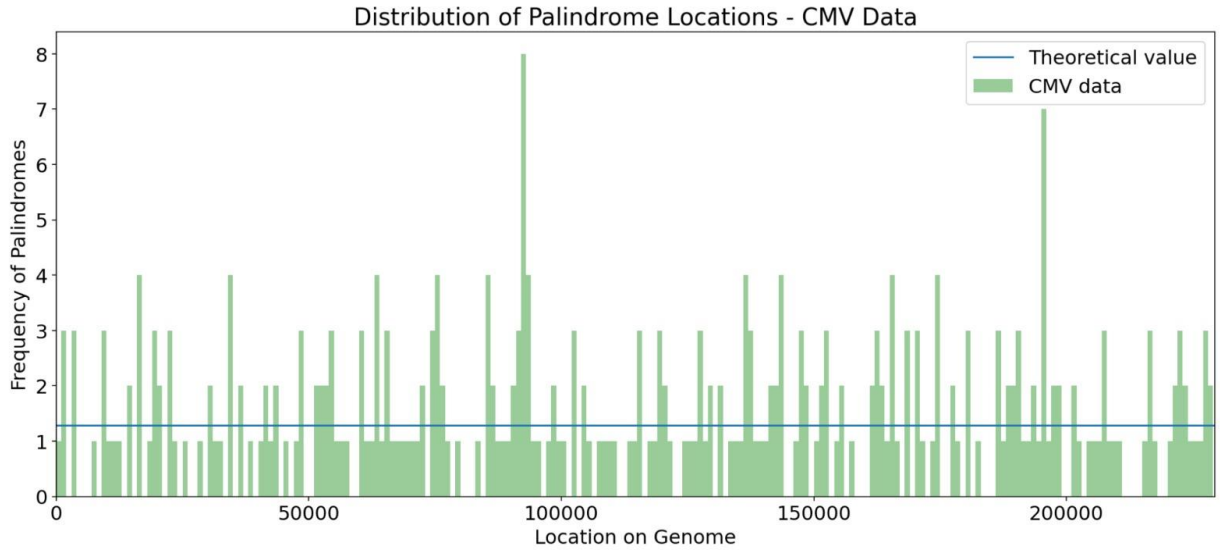
We used Python to create, first, a scatter plot of one random simulation in order to compare it with the distribution of the real data. We divided the DNA sequence into intervals of 1000 bases and of 4000 bases and counted the number of palindromes in each interval.

In order to account for random variability, we then generated 10 more random simulations, which we gathered into one boxplot. For each 4000 bases interval, there is a box which indicates the quartiles of the set of values (number of palindromes) from the ten simulations, together. We then plotted our data into a different histogram in order to compare it to the boxplot values.

On the histogram showing the distribution of the real data, we also plotted a line of the expected (theoretical) number of palindromes in each interval if these palindromes were to be equally distributed across the DNA sequence.

Analysis

Figure 0 – Interval = 1000 base pairs



We can see here that our dataset has much denser clusters at the following locations:
Between base pairs [92000, 93000] and between base pairs [195000, 196000].

Figure 1

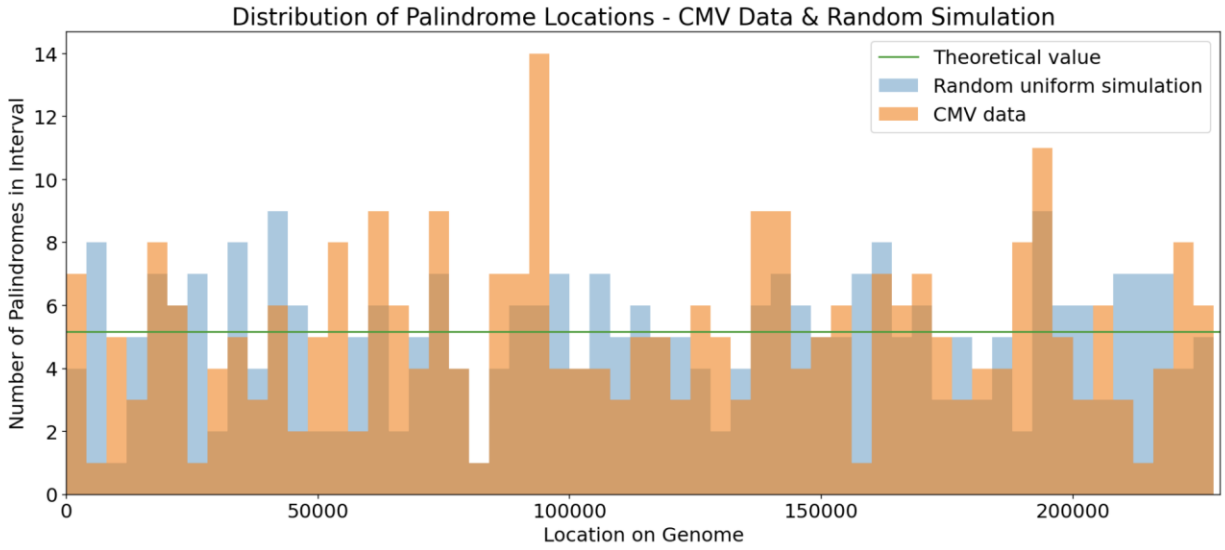


Figure 1- Comparison between sample and one random simulation

Figure 1 highlights some of the outstanding properties of our CMV sample. In the random simulation, counts never exceeded 9 palindromes per 4000-base interval. Meanwhile, in our dataset they did exceed that number twice, reaching 13 and 14 palindromes in two separate intervals.

Figure 2

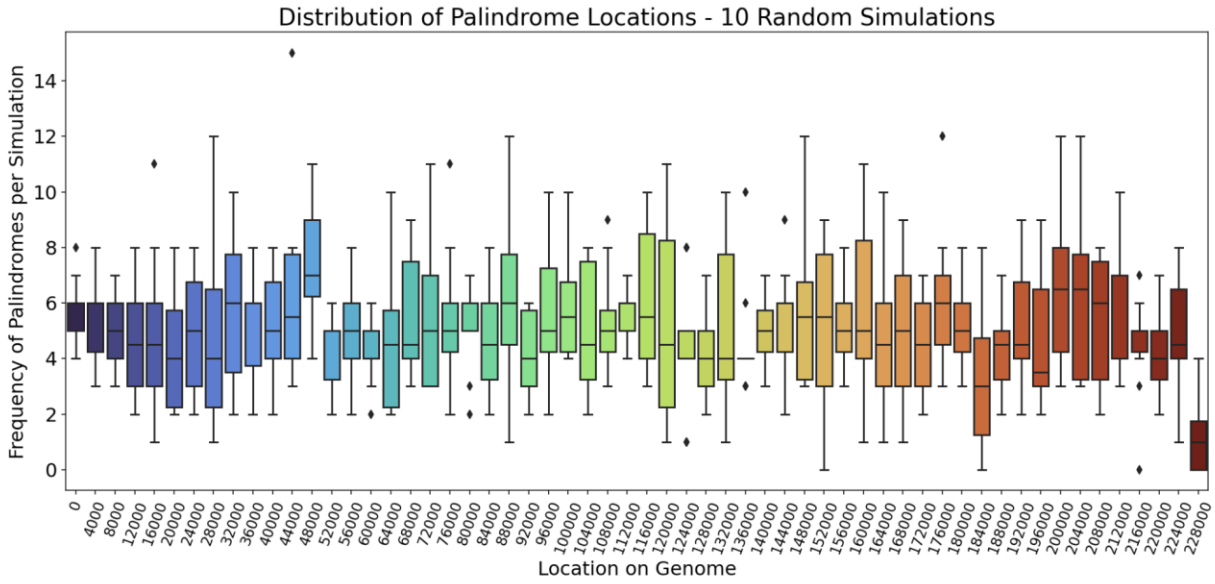


Figure 2 - Distribution of 10 random simulations. Each box shows the minimum, 25th percentile, median, 75th percentile, and maximum of the values observed in that interval across the 10 simulations.

Figure 3

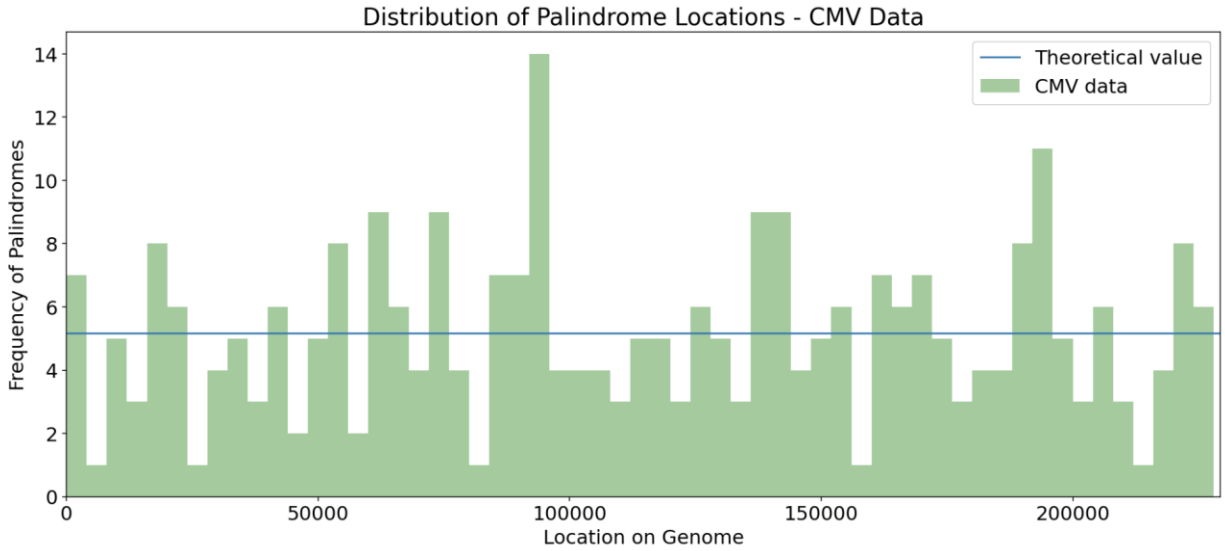


Figure 3 - Distribution of palindromes along DNA sequence in our sample of CMV.

Figures 2 and 3 can be compared to see how some of the values in our dataset are quite unusual. In 10 simulations, only one count of 14 palindromes in an interval was observed (between base numbers 44000 and 48000), and it was denoted an outlier by the Python system since the next largest value observed across the simulations in that interval was 8. Across all the simulations and intervals, the number of palindromes per interval usually peaked between 8 and 12 palindromes (never exceeding 12), with a considerable number of intervals whose largest number of palindromes was lower than 8.

Furthermore, the median number of palindromes per interval across the 10 simulations never exceeds 7. In contrast to this, our dataset has multiple intervals with more than 8 palindromes.

Conclusions

Our dataset appears to have two unusually dense clusters of palindromes that are rarely observed in random simulations. This does not necessarily indicate that our dataset does not follow a uniform distribution - however, it does prompt for further investigation.

2.2 Spacing Between Consecutive Palindromes

2.2 Spacing Between Consecutive Palindromes

Methods

In this section, we measured the number of bases in between pairs and triplets of palindromes in our dataset and in a random simulation. We plotted the results in order to compare the distribution of our dataset with a random one.

The plots created were histograms and empirical cumulative density (ECDF) functions of both our dataset and the simulated dataset. We also plotted a line on each graph showing the expected (theoretical) value that the ECDF should take if the palindromes were uniformly distributed along the DNA sequence.

Analysis

Figure 4

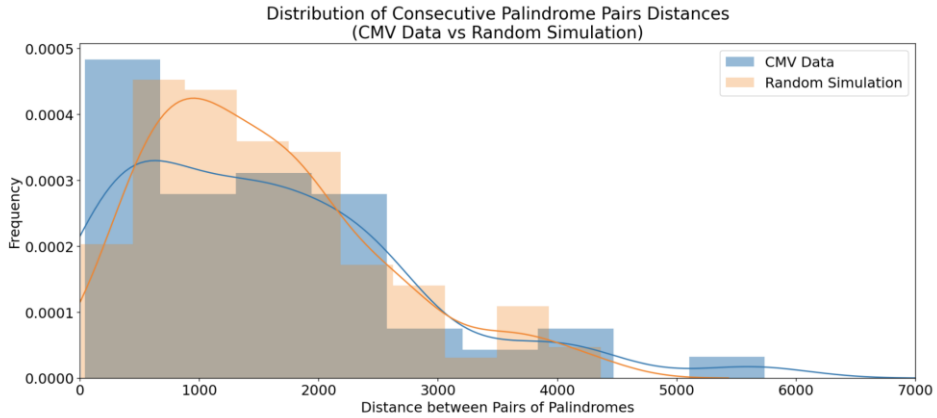


Figure 5

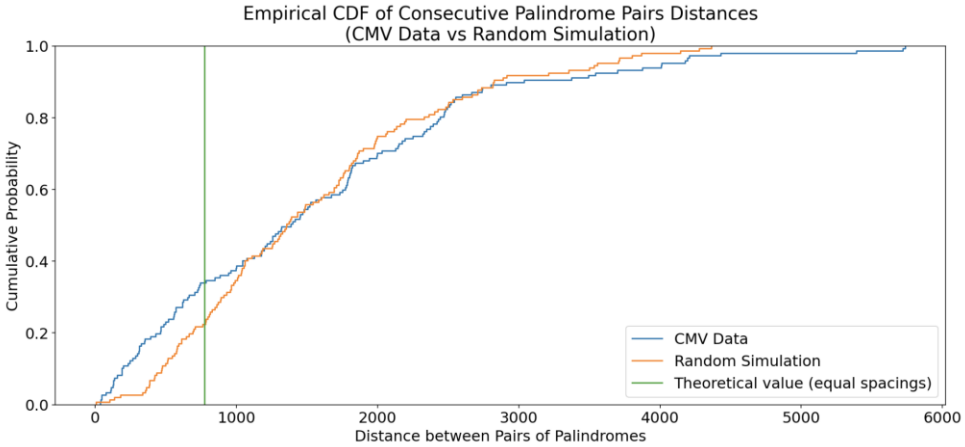


Figure 6

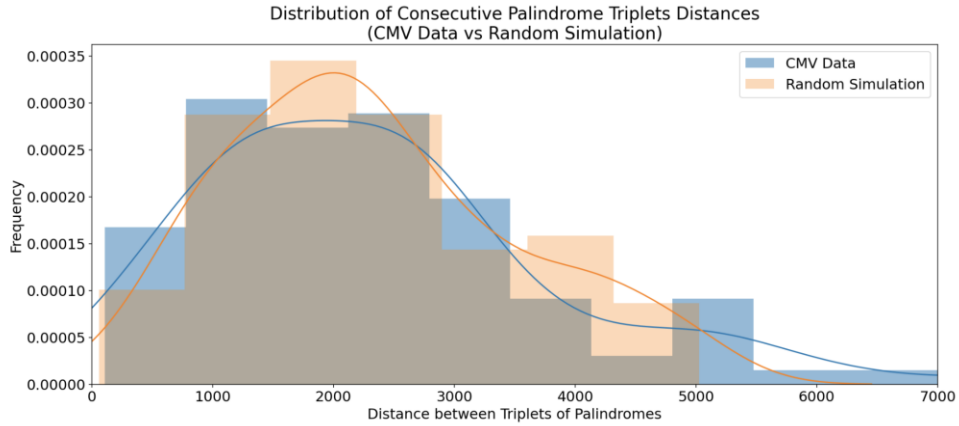
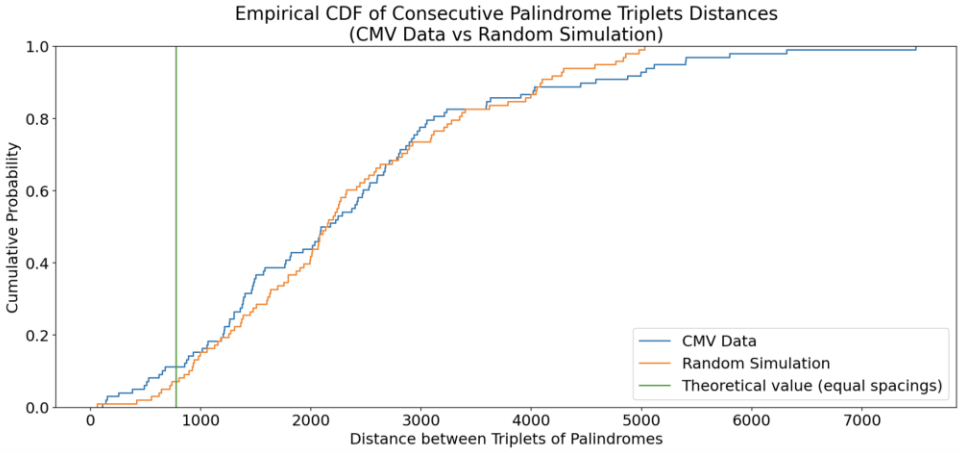


Figure 7



Our results show that pairs of palindromes that are close together (0 - 500 bases apart) quite a lot more frequent in our dataset than it was in a random simulation.

Conclusions

Our results indicate that the palindromes in our sample of CMV are clustered more densely than expected from a uniform distribution.

2.3 Expected Number of Intervals with X Palindromes

Methods

Here we again shift our perspective in order to observe the incidence of intervals with many (more than expected) palindromes in them. To do this, we grouped the 4000-base intervals in our dataset by the number of palindromes in them and counted the number of intervals in each group. We then performed the same procedure on 10 random simulations and plotted the results onto a boxplot.

In addition to this, we compared our results to the expected (theoretical) values if the palindromes were uniformly distributed along our DNA sequence. In order to do this, we plotted expected values from a Poisson distribution with rate of 5.162 (=296

palindromes / number of intervals) and performed a chi-square test between it and our dataset's values.

Analysis

Figure 7

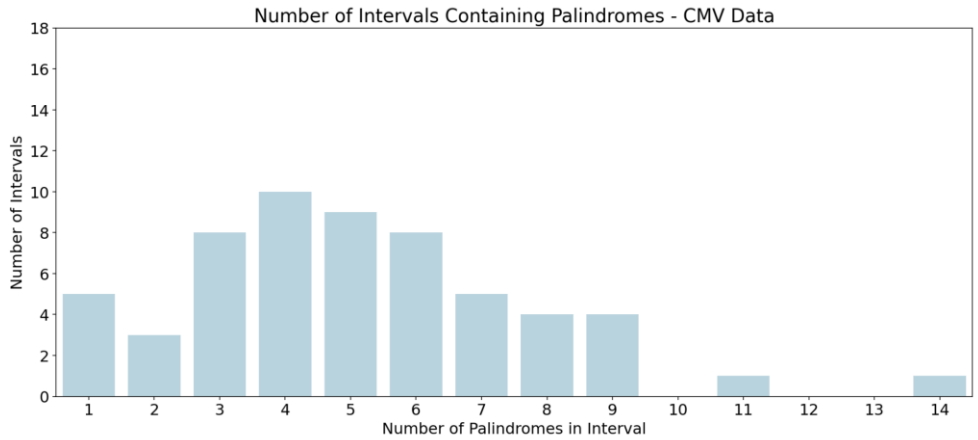
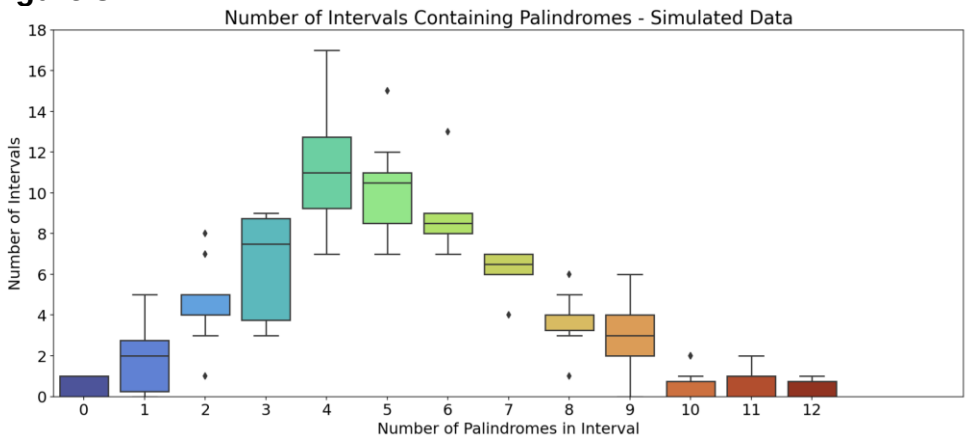
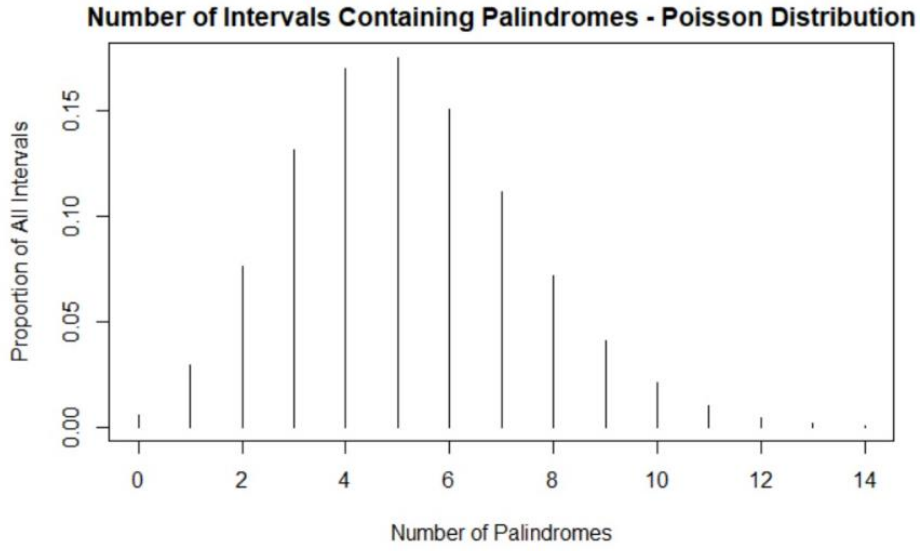


Figure 8



Figures 7 and 8 can be compared to see how our dataset had a large number of intervals with only 1 palindrome in it, and few intervals with relatively large numbers of palindromes in them (6 - 9 palindromes). However, our dataset did have one interval with 14 palindromes in it, which was never observed across 10 random simulations.

Figure 9



Chi-Square p-value: 0.00104 (significant)

R gave us a warning: "Chi-Squared approximation may be incorrect" because the expected values are small and therefore the approximations of p may not be right.

Conclusions

The difference between the distributions of the 10 simulations and our dataset, seen in Figures 7 and 8, might indicate that the palindromes in our dataset are more sparsely distributed at many places and then more densely distributed in a few places.

In addition, our chi-square test clearly indicates that the palindromes in our dataset are very likely not to be following a uniform distribution along the DNA sequence. However, this should be taken with a grain of salt since our dataset does not necessarily meet all of the assumptions necessary to perform a chi-square test.

2.4 Changing Interval Sizes and Indentifying Largest Cluster

Methods

Here, we looked at the interval containing the largest number of palindromes observed across all intervals in our dataset and compared it to that of 10 random simulations. We did over 8 different interval lengths - 500, 1000, 2000, 3000, 4000, 5000, 6000, and 8000 bases.

Analysis

Table 1 - (Simulated Data) Number of Palindromes in Largest Cluster per Simulation per Interval Size

	500	1000	2000	3000	4000	5000	6000	8000
0	4	6	6	9	11	14	13	16
1	5	5	8	11	15	12	16	20
2	4	5	9	10	10	12	15	16
3	4	5	8	10	11	14	15	17
4	5	5	7	9	11	12	13	16
5	4	5	7	9	13	11	15	15
6	3	7	8	9	12	13	16	16
7	4	5	8	9	11	10	17	17
8	4	5	7	10	11	15	14	19
9	3	5	8	11	10	11	15	18

Table 2 - (Simulated Data) Average Number of Palindromes in Largest Cluster per Interval Size

500	1000	2000	3000	4000	5000	6000	8000
4	5.3	7.6	9.7	11.5	12.4	14.9	17

Table 3 - (Real Dataset) Number of Palindromes in Largest Cluster per Interval Size

500	1000	2000	3000	4000	5000	6000	8000
8	8	12	13	14	18	19	21

We can see that our dataset has clusters that are much larger than the ones seen in our 10 simulations, especially when the interval lengths are small. Across all 10 simulations, the largest cluster of palindromes seen at intervals of length 500 bases had 5 palindromes in it, while our dataset had an interval of the same length with 8 palindromes in it. Similar differences between our dataset and expected values can be seen by comparing tables 2 and 3.

Conclusions

This section gives a strong indication that our sample has quite abnormally large clusters of palindromes. This was highlighted when we reduced the length of the intervals we were dividing our DNA sequence into.

3. Advanced Analysis

3. Advanced Analysis

Methods

In order to take a different perspective on how the palindromes in our dataset are clustered, we measured the spaces between each palindrome and its next palindrome, as well as the spaces between each palindrome and its second next palindrome and plotted the results. We also did the same for a simulated random dataset.

From these data, we created histograms and empirical cumulative density (ECDF) functions of both our dataset and the simulated dataset and compared the results. We also plotted a line on each graph showing the expected (theoretical) value that the ECDF should take if the palindromes were uniformally distributed along the DNA sequence.

Analysis

Figure 10

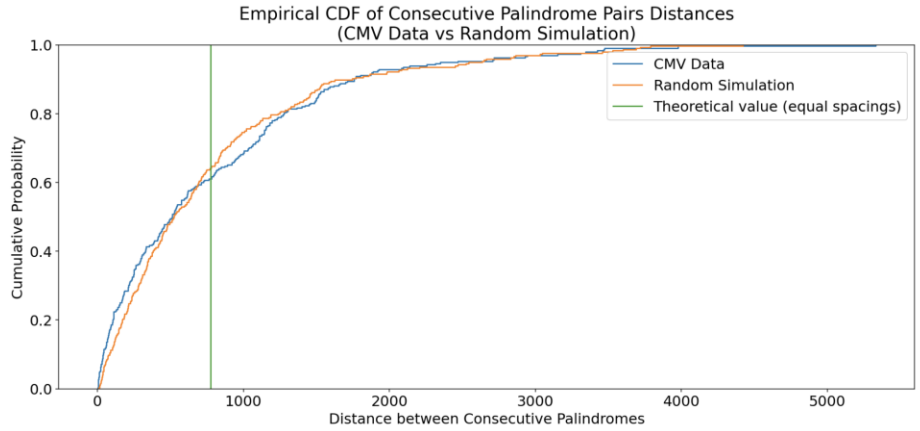


Figure 11

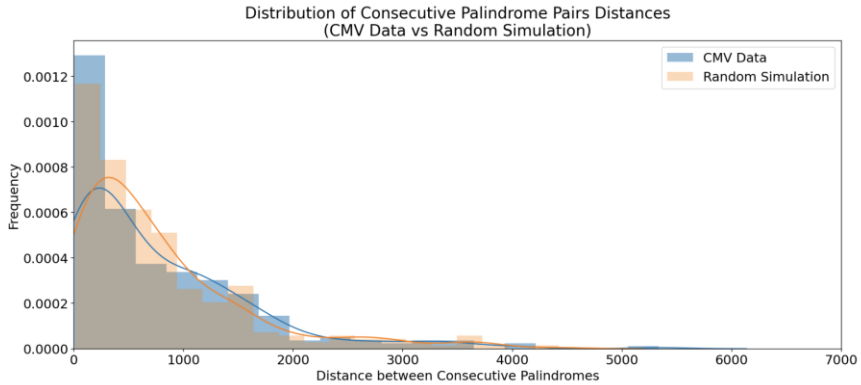


Figure 12

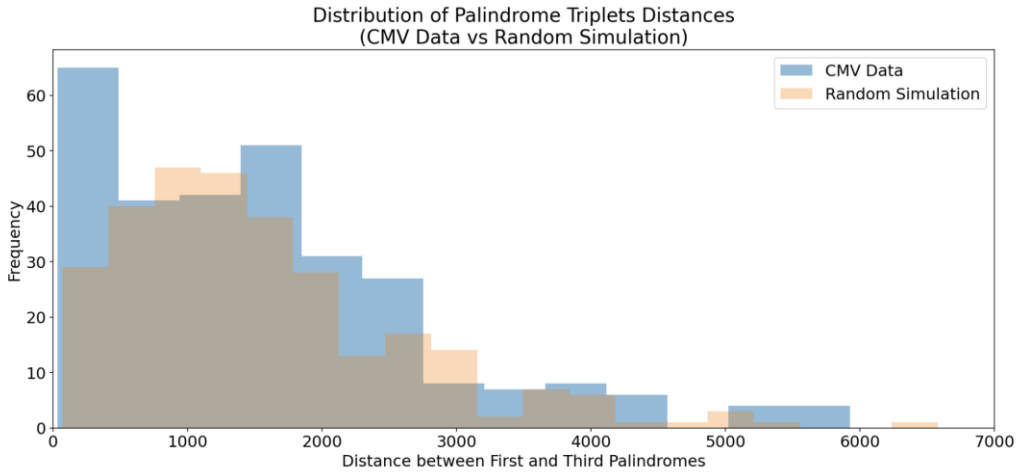
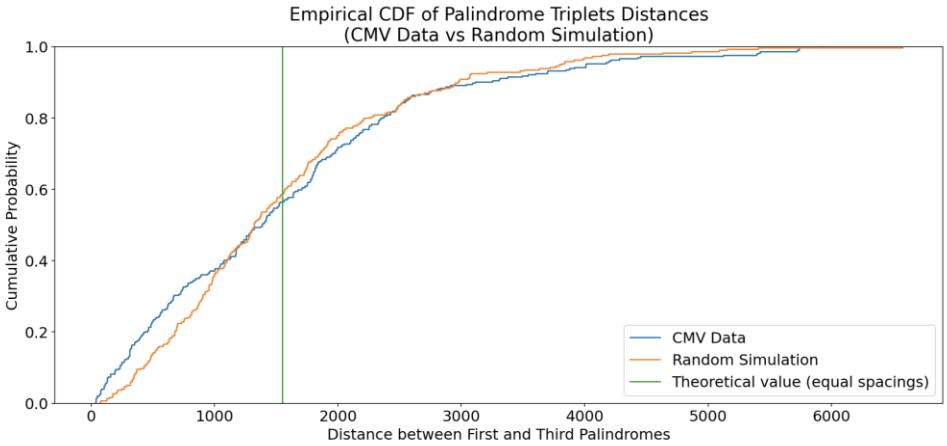


Figure 13



Figures 4 and 5 show that there is was a slightly larger chance of observing two palindromes close together (0 - 500 bases apart) in our sample than it was in a random simulation. Figure 5 also demonstrates that there was a smaller chance of observing palindromes that were far apart (> 800 bases apart) than in a random simulation.

Figures 6 and 7 show a much larger difference between our dataset and a random simulation. It can be seen that clusters of 3 palindromes close to each other (0 - 500 bases apart) were much more common in our dataset than they were in the random simulation.

Conclusions

Our findings in this section further indicate that our dataset has unusually dense clusters of palindromes, when compared to randomly scatter palindromes along a DNA sequence.

4. Discussion and Conclusion

Although we cannot establish that the palindromes in the Cytomegalovirus DNA are randomly distributed, unusual clusters were observed in our dataset. This report found graphical and numerical suggestions which point towards rejecting their distribution across intervals of base pairs as Uniform.

From calculating the distance between pairs and triplets of palindromes in our datasets and in our simulations, we found that pairs of palindromes that were close together were much more frequent in the real data than they were in simulations. This suggest that the dataset we studied had an unusual number of clustered values if compared with Uniform distributions.

A Pearson's Chi squared test was conducted in order to investigate this, and we rejected the Poisson Model as fit for our data. This suggests that the palindromes are not randomly distributed – portions of the DNA sequence with a higher number of palindromes might be relevant clusters indicative of replication sites. However, these results are only theoretical, since our data did not fit all the requirements for the Pearson's Chi squared test ($np>5$), which means that these results cannot be trusted in a real-world scenario.

To the biologist reading this report: in order to find more likely sites of replication within this virus' DNA sequence, look for clusters of palindromes within the data. We divided our data into intervals and compared them to intervals of uniformly distributed simulations of palindrome locations. We did this in order to pick out sites that were very unlikely to be densely packed with palindromes due to chance. The sites found were: [92000, 93000] and [195000, 196000]. These are the intervals of base pair locations where these largest clusters were found.

I would suggest that you first look at these two intervals when looking for the replication sites. In case you do not find the origin of the replication within these intervals, look at the next largest valued intervals, and so on. This should guide you towards the correct palindrome.