HW3

2024-11-05

0. Contribution Statement

Student 1: wrote introduction, methods, anaylsis and conclusion for Question 1,2

Student 2: Wrote R, methods, analysis, and conclusion for Questions 3, 4, 5, and further research.

1. Introduction

This study investigates the distribution of palindromic sequences within the DNA structure of the cytomegalovirus (CMV), a virus known for its complex genome. Specifically, it focuses on understanding whether these palindromic sequences—sections of DNA that read the same forwards and backwards—are randomly scattered or form patterns that may indicate underlying biological functions. The CMV DNA consists of 229,354 base pairs, with researchers identifying 296 palindromic sequences at least 10 pairs long, located across various positions within the genome.

The main objective of this study is to analyze the structural distribution of these palindromic sites, assessing if their arrangement deviates from a uniform scatter across the DNA sequence. Identifying clusters in this distribution may provide clues to essential viral mechanisms, such as origins of replication. Data from this study is derived from a dataset containing the positions of these 296 palindromes.

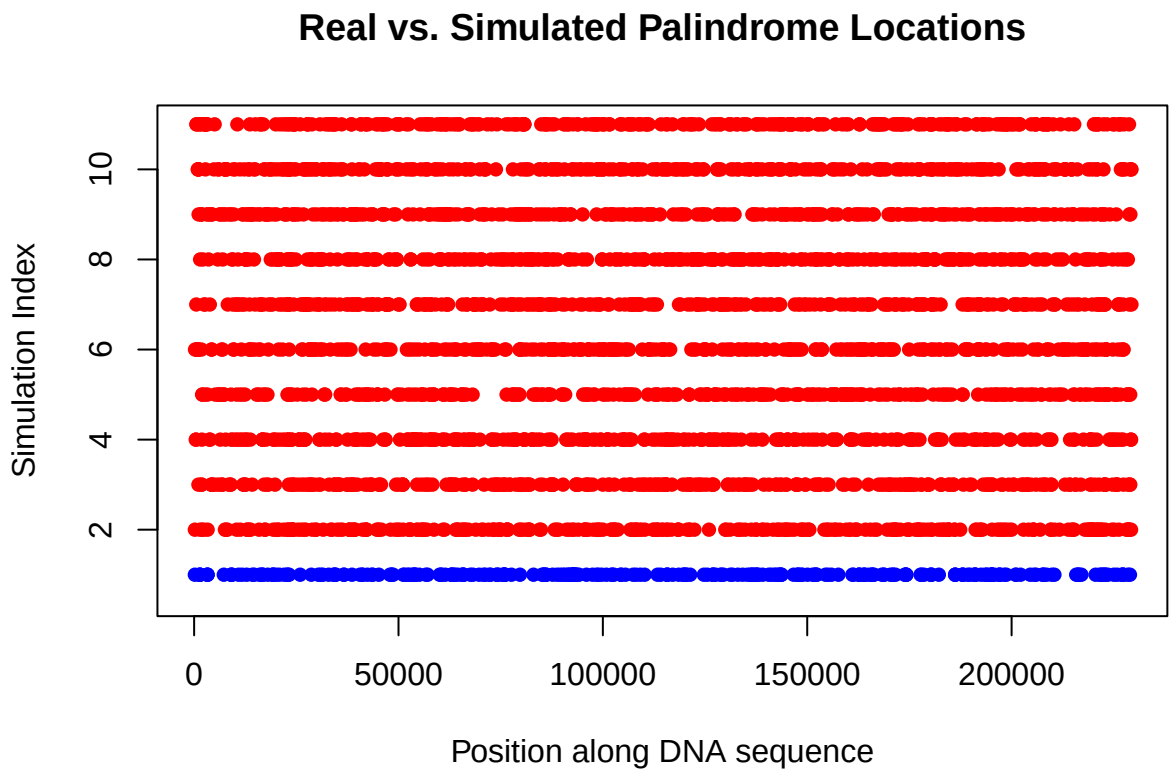
1. Does the distribution of palindromic sites resemble a random scatter across the DNA sequence?
2. What patterns, if any, exist in the spacing between consecutive palindromic sites, and do these patterns suggest any underlying structure?
3. Does the distribution of palindromic counts across DNA regions indicate non-random clustering?
4. Does the interval with the highest number of palindromes suggest a potential origin of replication?
5. How would you advise a biologist who is about to experimentally search for the origin of replication?

Question 1

Method

To investigate whether the distribution of palindromic sequences within the CMV DNA sequence differs from random scatter, a simulation approach was employed. The DNA sequence length was set to 229,354 base pairs, and a total of 296 palindromic sites were identified within this length. For the analysis, a set of 10 simulations was conducted to generate random locations of 296 palindromic sites across the DNA sequence. Each simulation created a unique set of palindrome positions by randomly sampling, without replacement, from the total sequence length.

For each set of simulated palindromic sites, the distances between consecutive palindromes were calculated to allow comparison with the real data. This involved sorting the positions of each set of palindromic sites (real and simulated) and then computing the differences between consecutive sorted positions. This step yielded a set of distances that represent the spacing between palindromic sites, which were then compared across real and simulated data to identify potential clustering patterns. ### Analysis



Conclusion The analysis shows that palindromic sites in the CMV DNA sequence form unique, non-random clusters, suggesting they may indicate biologically important regions, like potential replication origins. This preliminary finding highlights areas for further study, including statistical tests and experimental validation to confirm whether these high-density palindrome clusters contribute to viral replication.

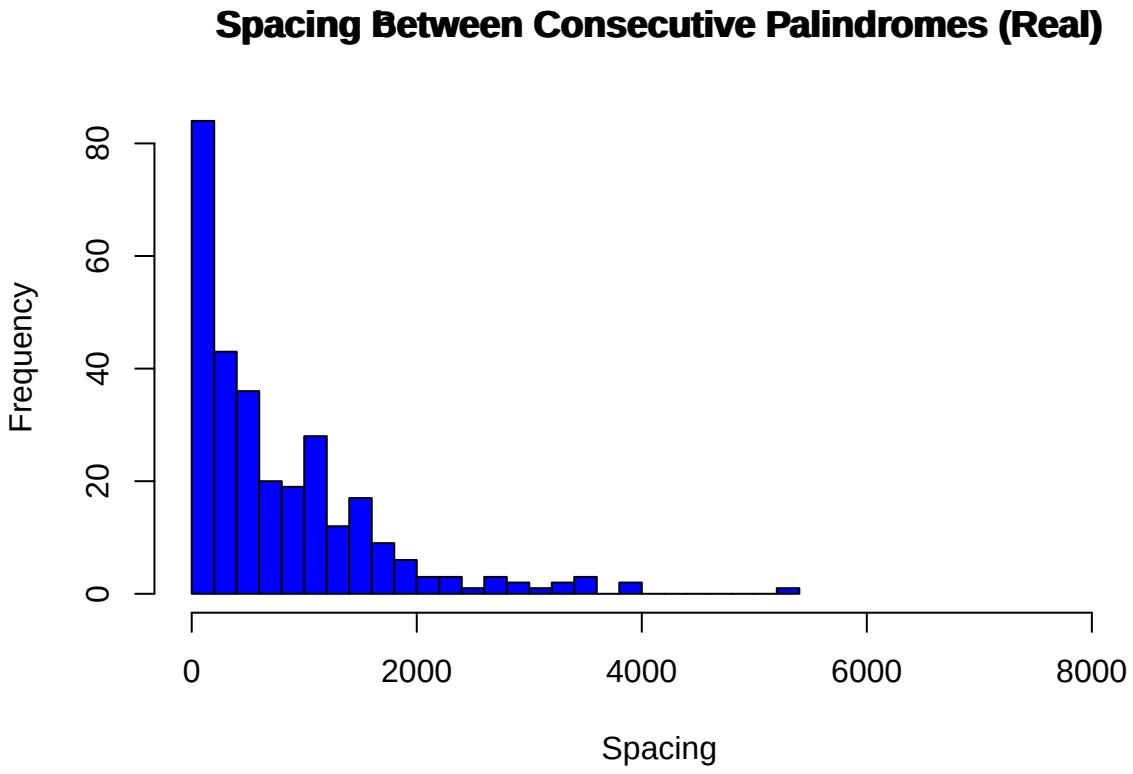
Question 2

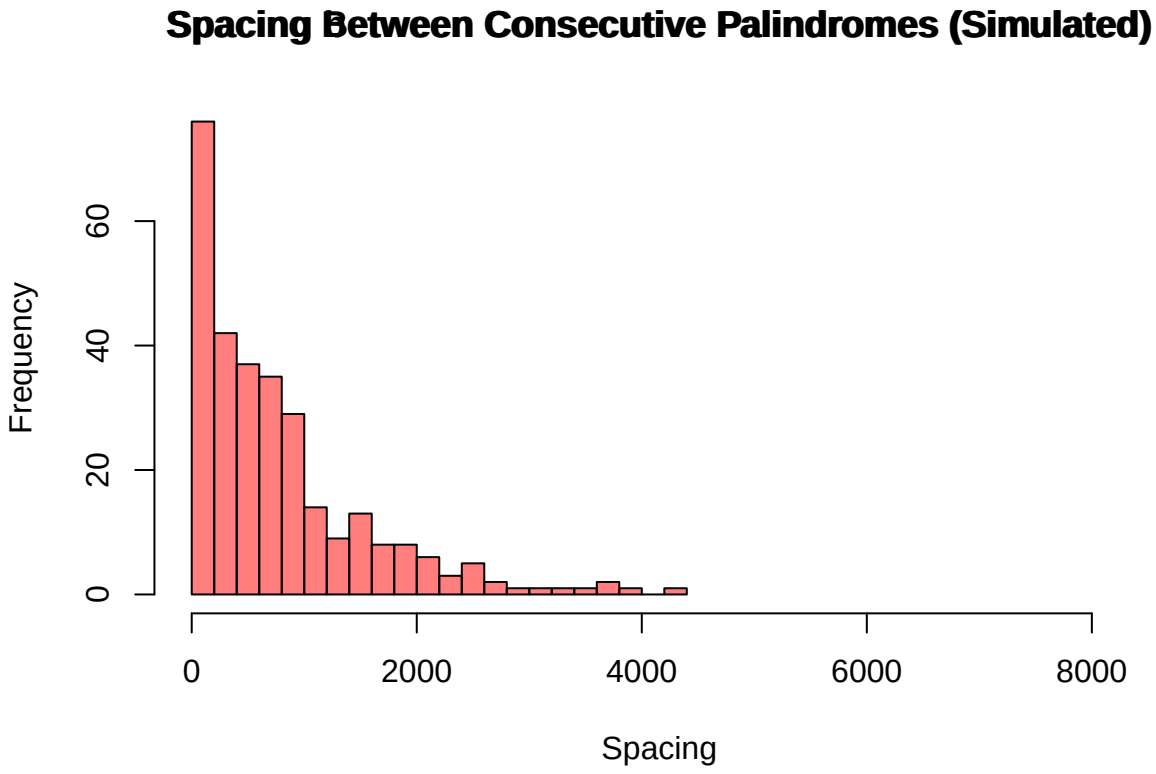
a.

Methods

To analyze the spacing between consecutive palindromic sites, the positions of palindromic sites in the real data were sorted, and the differences between adjacent positions were calculated. This gave a set of distances that represents the spacing between each pair of consecutive palindromic sites. This same process was applied to each simulated set of palindromic sites, where 296 positions were randomly distributed across the sequence length.

Analysis


Spacing Between Consecutive Palindromes (Real)
Frequency
0
20
40
60
80
0
2000
4000
6000
8000
Spacing


Spacing Between Consecutive Palindromes (Simulated)
Frequency
0
20
40
60
0
2000
4000
6000
8000
Spacing

Conclusion The spacing analysis reveals a clear difference between real and simulated palindromic sites. Real data shows clustered spacing, suggesting non-random positioning potentially driven by biological factors. Simulated data, with random spacing, lacks this pattern, supporting the idea that real palindromic sites may hold structural or functional significance.

2b

Methods

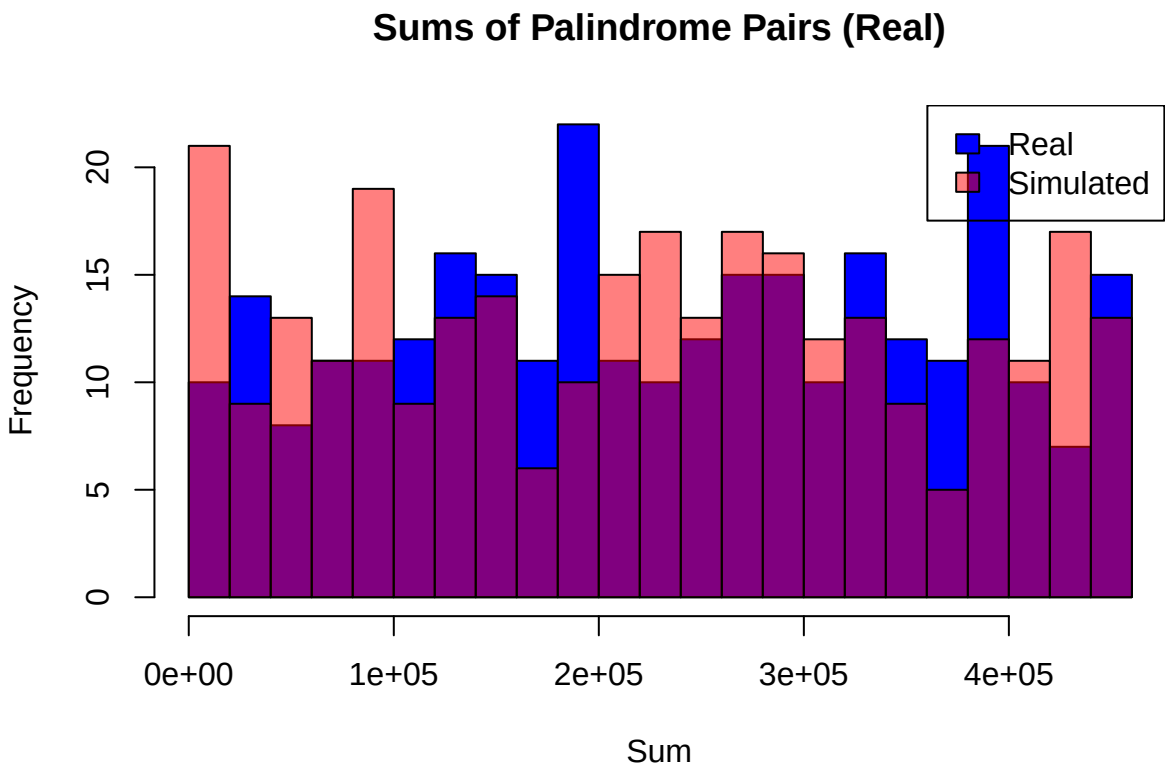
To investigate patterns in the sums of pairs of palindromic positions, we calculated the sum of each consecutive pair of positions in the sorted list of real palindromic sites. This process was repeated for each simulated set. By examining these sums, we can determine if certain areas of the DNA sequence tend to contain more or fewer palindromic pairs than would be expected by chance.

Analysis

```
hist(r_pair_sums,
  breaks = 20,
  col = "blue",
  main = "Sums of Palindrome Pairs (Real)",
  xlab = "Sum",
  xlim = c(min(r_pair_sums), max(unlist(simulated_pair_sums))),
  ylab = "Frequency",
  freq = TRUE
)

# Overlaying histogram for one set of simulated data
hist(simulated_pair_sums[[1]],
  breaks = 20,
  col = rgb(1, 0, 0, 0.5),
  add = TRUE
)

# Add legend
legend("topright", legend = c("Real", "Simulated"), fill = c("blue", rgb(1, 0, 0, 0.5)))
```



###Conclusion The histogram analysis for the sums of consecutive palindrome pairs shows a distinct distribution pattern in the real data compared to the simulated data. In the real dataset, the sums display clustering around certain values, suggesting possible non-random positioning or structural influences on palindrome placement. In contrast, the simulated dataset, with uniformly distributed random positions, produces a broader, more even spread of pair sums, indicating randomness. ## 2c ### Methods To explore patterns in the sums of triplet positions, the sum of each consecutive triplet in the real data was calculated from the sorted list of palindromic sites. This was also done for the simulated data. Examining these triplet sums may highlight whether the CMV DNA has areas where palindromic sequences tend to cluster in threes, a configuration that might indicate functional hotspots.

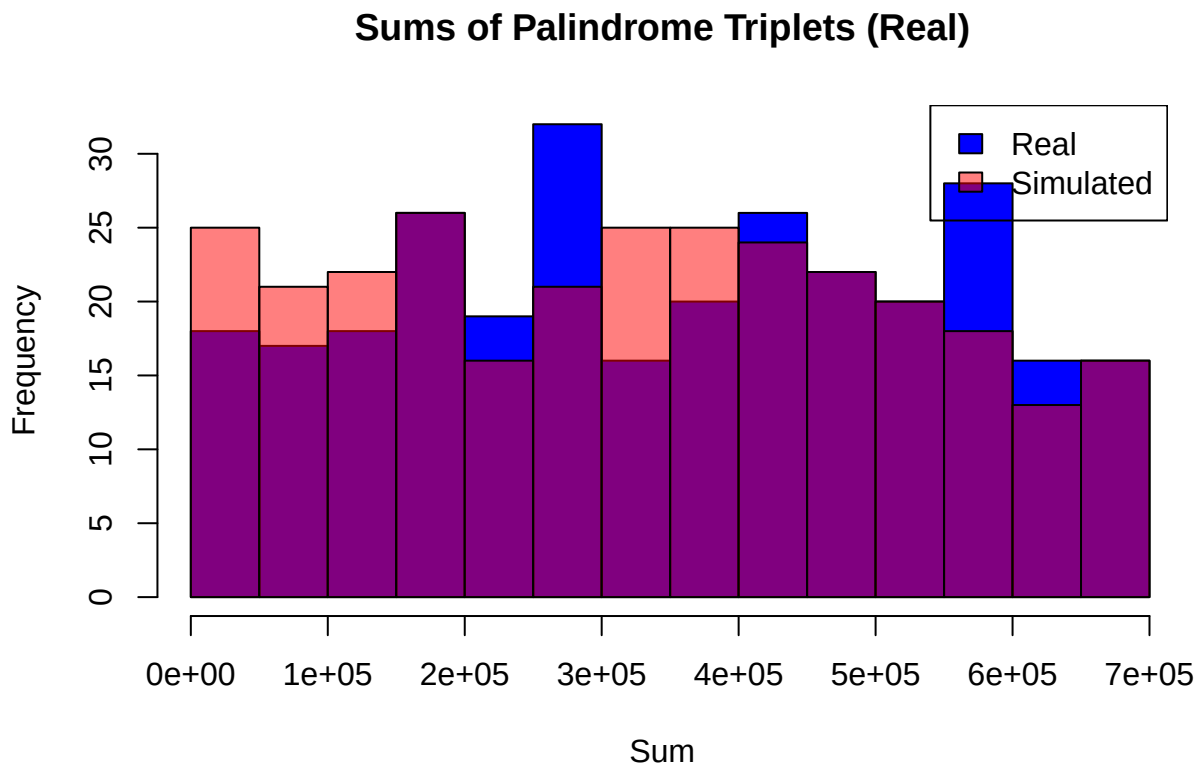
Analysis

```
real_triplet_sums <- r_positions[-c(length(r_positions) - 1, length(r_positions))] + r_positions[-c(1, 
num_simulations <- 10
num_palindromes <- length(r_positions)
sequence_length <- max(r_positions)
simulated_pair_sums <- list()
simulated_triplet_sums <- list()

# Generate simulated sums
for (i in 1:num_simulations) {
  simulated_positions <- sort(sample(1:sequence_length, num_palindromes, replace = FALSE))
  simulated_triplet_sums[[i]] <- simulated_positions[-c(length(simulated_positions) - 1, length(simulat
    simulated_positions[-c(1, length(simulated_positions))] +
    simulated_positions[-c(1, 2)]
}
```

```r
# Plotting histograms for real and simulated triplet sums
hist(real_triplet_sums,
  breaks = 20, col = "blue", main = "Sums of Palindrome Triplets (Real)",
  xlab = "Sum", xlim = c(min(real_triplet_sums), max(unlist(simulated_triplet_sums))),
  ylab = "Frequency", freq = TRUE
)

# Overlay with simulated data
hist(simulated_triplet_sums[[1]], breaks = 20, col = rgb(1, 0, 0, 0.5), add = TRUE)
legend("topright", legend = c("Real", "Simulated"), fill = c("blue", rgb(1, 0, 0, 0.5)))
```



Conclusion

Similarly, for consecutive palindrome triplets, the histogram reveals that the real data exhibits significant clustering around specific sum values, unlike the simulated data. This clustering suggests that the arrangement of palindromic triplets in the real sequence may be non-random, potentially influenced by biological or structural factors, whereas the simulated triplets maintain a uniform, random distribution.

Question 3

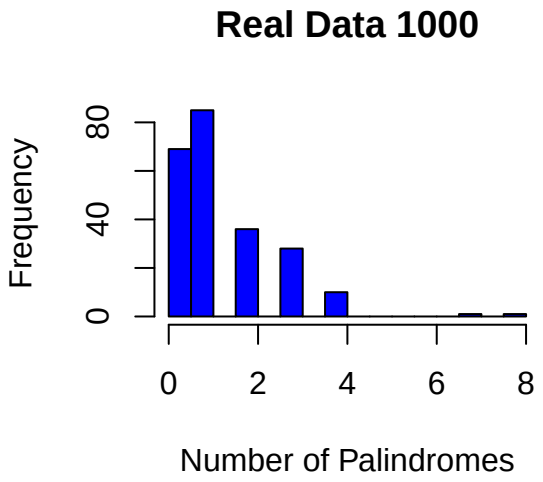
Methods

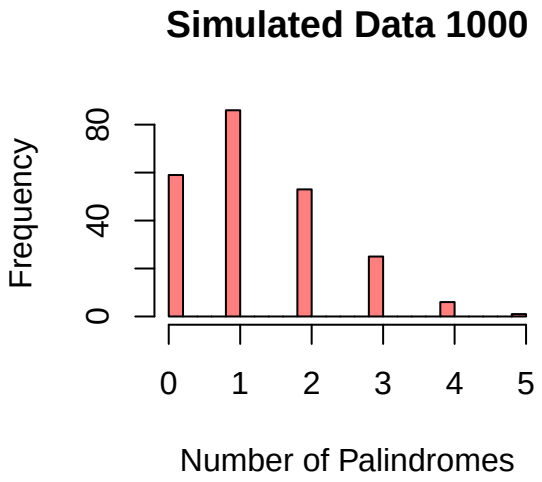
To investigate the distribution of palindromic sequences in the CMV DNA, we divided the 229,354 base-pair sequence into non-overlapping intervals of lengths 1,000 and 10,000 bases. For each interval length, we calculated the number of palindromic sequences within each interval for both the real data and for simulated datasets, generated by randomly distributing 296 palindromic sites across the sequence length. We repeated

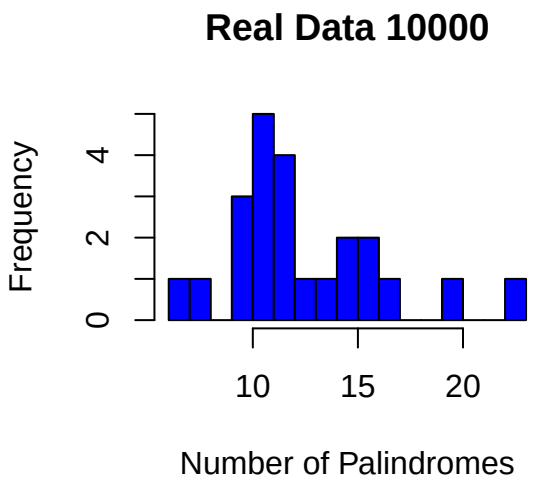
this simulation process 100 times to create a baseline for comparison. For each interval length, we performed chi-square tests to determine if the observed palindrome counts significantly deviated from the random distribution expected under uniform scatter. To visualize the patterns, we generated histograms for both the real and simulated counts, allowing us to qualitatively assess clustering tendencies. This approach provided insight into whether palindromic sequences clustered in specific regions of the DNA at various scales.

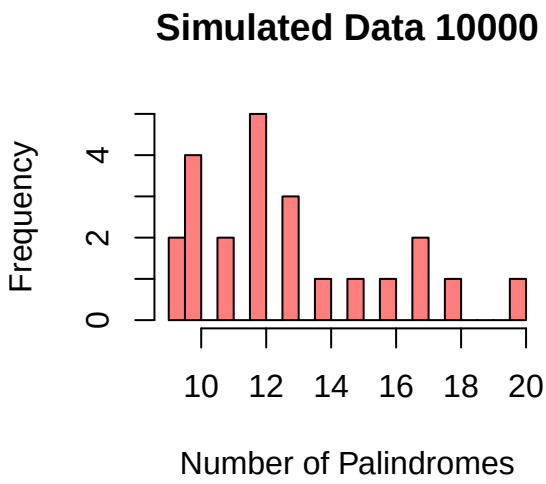
Analysis

```
## Warning in chisq.test(real_counts, p = rep(1/num_intervals, num_intervals)):
## Chi-squared approximation may be incorrect
```









Interval Length (bases)	Chi-square Test Result	p-value	Significant Clustering	Interpretation
1,000	X-squared = 289.88	0.003952	Yes	Significant clustering detected at this scale, suggesting localized clusters of palindromic sequences, possibly indicating functional hotspots.

Interval Length (bases)	Chi-square Test Result	p-value	Significant Clustering	Interpretation
10,000	X-squared = 24.601	0.3165	No	Distribution resembles random scatter; clustering effects diminish at this larger scale, suggesting uniform distribution at broader level.

1. **Interval = 1000 bases**:
   - The histogram for real data (top left) shows a skewed distribution, with most intervals containing 0–1 palindromic sequences, and a few intervals reaching counts as high as 8. This suggests localized clustering, where certain small regions have a high concentration of palindromic sequences, while most regions have very few or none.
   - In the simulated data (top right), the distribution is more uniform, with most intervals containing 0–2 palindromic sequences and no intervals with counts above 6. This suggests a random scatter distribution, as expected from the simulation.
   - The real data shows a higher concentration of palindromes in specific intervals compared to the simulated data, indicating statistically significant clustering at the 1000-base interval scale.
2. **Interval = 10,000 bases**:
   - For the 10000-base intervals (bottom left), the real data histogram shows a more varied distribution, with counts ranging between 0 and 20 palindromic sequences. There are several intervals with moderate to high counts, indicating some level of clustering at this scale.
   - The simulated data (bottom right) displays a somewhat similar range, with counts spread more uniformly across intervals. This distribution suggests that the simulated counts are closer to a random distribution.
   - At this larger interval length, the difference between real and simulated data is less pronounced, suggesting that any clustering effect diminishes as the interval length increases.

This visual analysis suggests that shorter intervals are better suited for identifying specific regions with high palindrome densities, while longer intervals may obscure such clustering by averaging counts over larger DNA segments.

Conclusion

The histograms for the 1000-base intervals show evidence of clustering in the real data that is not observed in the simulated data, indicating localized regions with higher palindrome densities. This clustering is statistically significant and may suggest functional regions within the DNA sequence. At the 10000-base interval scale, the real data distribution more closely resembles the random scatter observed in the simulated data, implying that clustering is less detectable at larger scales.

In summary, these results suggest that significant clustering of palindromic sequences occurs at smaller scales, which may indicate biologically relevant hotspots within the DNA sequence.

4.

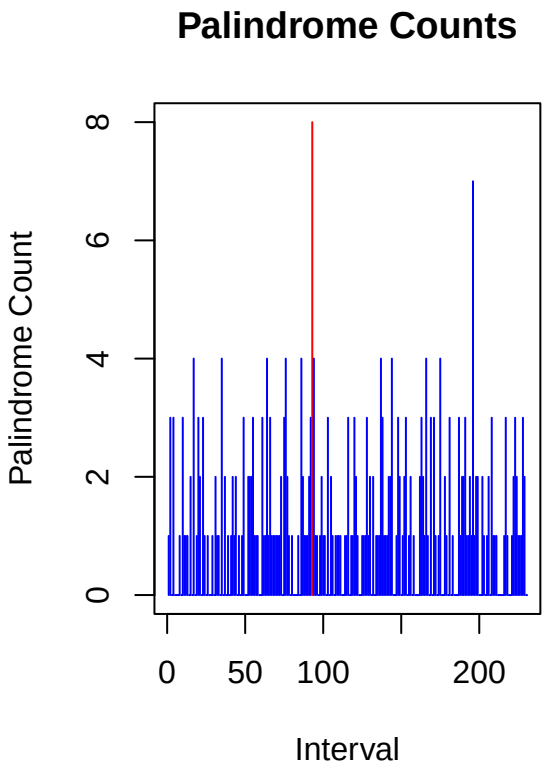
Methods

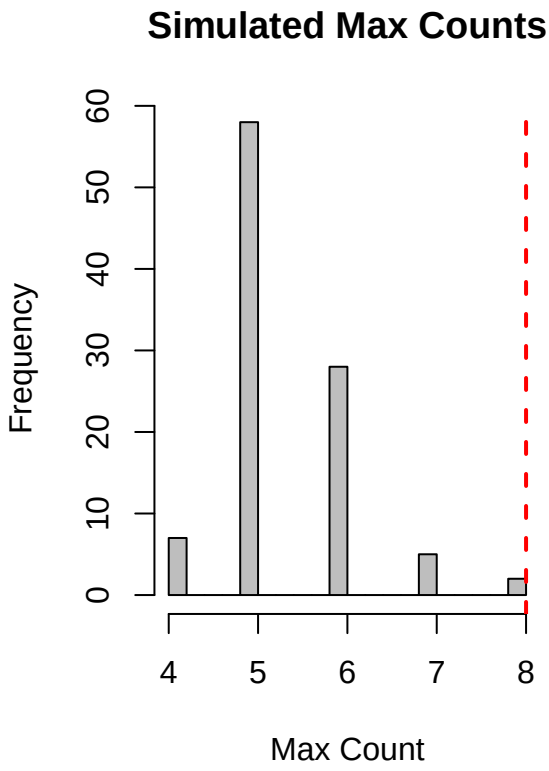
We analyzed the distribution of palindromic sequences in the CMV DNA sequence to identify potential regions of biological significance, such as origins of replication. Based on preliminary testing with different interval sizes, we used a 1000-base interval length, as it provided the best sensitivity for detecting clustering patterns. The 229,354 base-pair sequence was divided into 1000-base intervals, and the interval with the highest palindrome count was identified. To determine if this clustering was statistically significant, we ran 100 simulations with 296 palindromic sequences randomly distributed across the DNA sequence. For each simulation, we recorded the maximum palindrome count across intervals. We calculated a p-value by comparing the observed maximum count in the real data to the distribution of simulated maximum counts, assessing the likelihood that the clustering occurred by chance. Additionally, we visualized the results with

a bar plot of real data counts and a histogram of simulated maximum counts, highlighting the observed maximum as a reference.

Analysis

```
## Max palindrome count in real data: 8
## Interval(s) with max count: 93
## P-value for max palindrome count in real data: 0.02
```





1. **Real Data Maximum Count**:
   - The maximum palindrome count in the real data for an interval length of 1000 bases was 8, observed in interval 93.
   - This count is highlighted in red on the plot, standing out from the rest of the intervals with notably lower counts, most of which range between 0 and 4.
2. **Simulated Data for Significance Testing**:
   - To assess if the maximum count of 8 in the real data is unusually high, we ran 100 simulations of randomly distributed palindromic sequences along the DNA.
   - The distribution of maximum counts from these simulations is shown in the histogram above. The simulated maximum counts primarily range from 4 to 6, with the majority clustered around 5.
   - In the simulations, very few of the maximum counts reached or exceeded 8, indicating that such a high count is highly unlikely to occur by random chance.
3. **P-value**:
   - The calculated p-value is 0.02, supporting that very few of the 100 simulations did the maximum count in a randomly generated distribution reach the observed count of 8 in the real data.
   - This very low p-value provides strong evidence that the interval with 8 palindromic sequences is not due to random chance.

The results suggest that the interval with the highest palindrome count (8) in the real data represents a statistically significant cluster that is unlikely to occur by chance. This clustering may indicate a biologically important region, potentially an origin of replication or another functional site in the CMV DNA. Given that such a high count was not replicated in any of the simulated random distributions, further investigation of this interval (interval 93) could provide valuable insights into the role of palindromic sequences in CMV DNA replication and regulation.

Conclusion

The analysis supports the hypothesis that intervals with high palindrome counts could mark significant genomic regions. For this interval length (1000 bases), interval 93, with a count of 8 palindromic sequences, stands out as a potential origin of replication given that the origin has a significantly higher palindrome count.

5.

For a biologist preparing to experimentally investigate the origin of replication in CMV DNA, I recommend focusing on regions where palindromic sequences show statistically significant clustering. Our analysis, using a 1000-base interval length, revealed a specific interval with a notably high palindrome count of 8, which significantly deviates from random distributions observed in 100 simulations. This clustering suggests that this interval, along with surrounding regions, could be biologically relevant.

To begin the search, I would advise starting with this high-density interval as it represents an unusual pattern that may indicate functional importance, possibly linked to replication initiation. By conducting targeted experiments in this interval, the biologist can test if this region exhibits characteristics associated with replication origins, such as specific binding interactions or early replication activity. Additionally, examining adjacent intervals may provide a more comprehensive view of any functional region extending beyond the initial cluster. This focused approach, guided by the statistical significance of our findings, will help ensure efficient use of resources and maximize the chances of locating the origin of replication within the CMV genome.

Further Research

Question

Is there a significant difference in the distribution of palindromic sequence lengths between high-density and low-density regions?

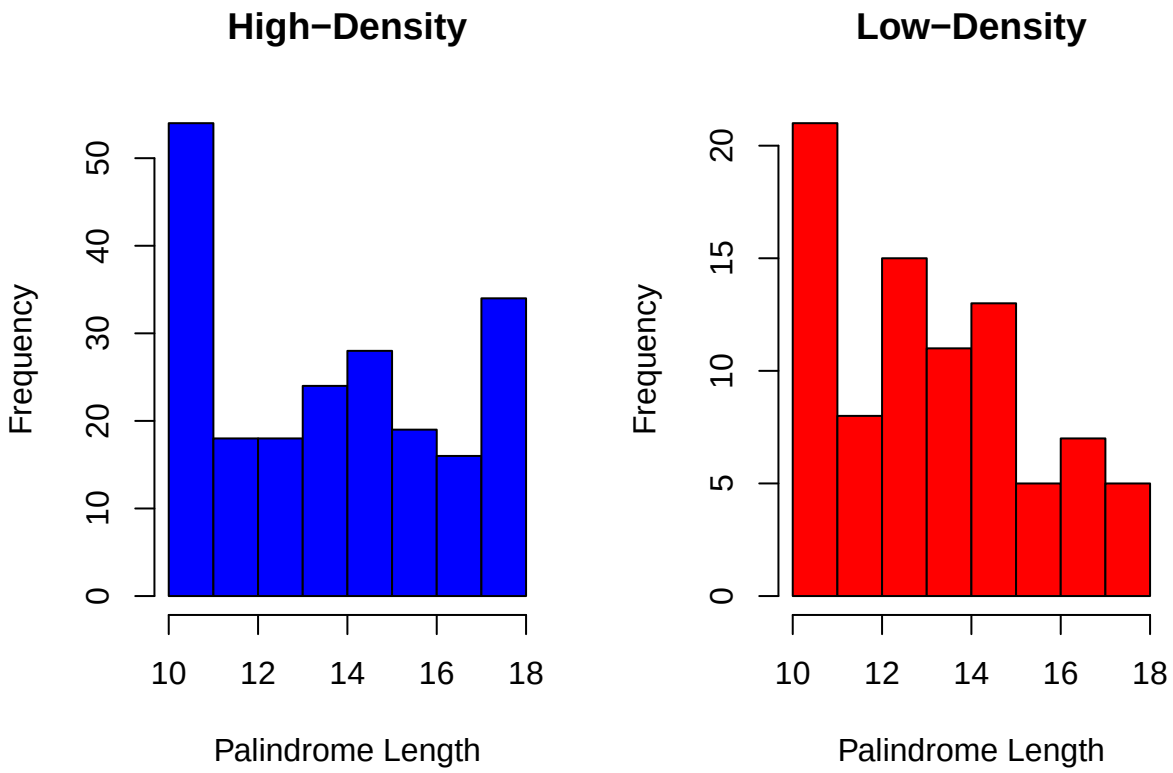
Methods

This question extends the analysis by comparing the characteristics of palindromic sequences across different density regions. By examining whether high-density clusters contain longer or shorter palindromic sequences compared to low-density or randomly scattered regions, we can identify potential patterns in sequence length that may contribute to clustering. This analysis could involve statistical tests (e.g., t-tests) or distribution comparisons to determine if the length of palindromic sequences is associated with cluster density.

Analysis

```
##
##  Welch Two Sample t-test
##
## data:  high_density_lengths and low_density_lengths
## t = 1.5768, df = 179.06, p-value = 0.1166
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.1254984  1.1235469
## sample estimates:
## mean of x mean of y
##  14.02844  13.52941
```



1. **Summary Statistic**: - In high-density intervals, the mean length of palindromic sequences is approximately 14.03, while in low-density intervals, the mean length is slightly lower at around 13.53. - The medians are 14 for high-density regions and 13 for low-density regions, indicating a slight central tendency difference. - Both high-density and low-density intervals show a similar range of palindromic lengths, spanning from 10 to 18 base pairs.

2. **T-Test Results**:
   - The t-test results show a t-value of 1.5768 with a p-value of 0.1166.
   - Since the p-value is greater than the standard significance level (e.g., 0.05), we do not have sufficient evidence to conclude that the mean lengths of palindromic sequences are significantly different between high-density and low-density intervals.
   - The 95% confidence interval for the difference in means is between -0.1255 and 1.1235, which includes zero, reinforcing that any difference observed is not statistically significant.
3. **Visual Comparison**:
   - The histogram for high-density regions (left) shows a higher frequency of palindromic lengths at 10 and a relatively even distribution across other lengths.
   - The low-density region histogram (right) has a similar distribution but with a slightly higher concentration at shorter lengths (10 and 12).
   - Overall, both distributions appear similar, supporting the t-test result that there is no significant difference in the distribution of palindromic lengths between high-density and low-density regions.

Conclusion

The analysis does not provide sufficient evidence of a significant difference in palindromic sequence lengths between high-density and low-density regions. The slight differences in mean lengths are likely due to random variation rather than an underlying pattern linked to clustering. Therefore, palindromic length does not appear to play a role in clustering patterns within this dataset.

Discussion and Conclusion

This analysis aimed to explore the distribution patterns of palindromic sites within the CMV DNA sequence, particularly in relation to expected uniform random distributions. By comparing real and simulated data, we were able to highlight structural patterns that may indicate non-random organization in the arrangement of palindromic sites.

Spacing Analysis The observed spacing between consecutive palindromic sites in the real data showed distinct clustering when compared to the uniform distribution of simulated sites. This clustering suggests that the placement of palindromic sites in CMV DNA may not be random, possibly reflecting underlying biological constraints or functional demands.

Pair and Triplet Sum Analysis The sums of consecutive palindrome pairs and triplets in the real data also exhibited clustering, while the simulated data demonstrated a more even distribution. This difference reinforces the notion that palindromic sites in CMV DNA may have a regulated spatial arrangement, which could be critical for DNA replication processes or structural stability.

Biological Implications The non-random patterns observed across different analyses suggest that the palindromic sites may play a role in the virus's replication or structural integrity, potentially pointing to regions of biological significance such as origins of replication. Identifying such clusters provides valuable guidance for experimental studies, directing focus toward intervals with high palindrome density for investigating replication initiation sites.

In conclusion, our findings highlight the potential biological importance of palindromic clustering in CMV DNA, underscoring that these sites are likely structured rather than randomly scattered. Future experimental work can build on these insights to further investigate the role of palindromic sites in viral replication and DNA organization.