Statistical Analysis of CMV DNA Data

2024-11-10

Contribution

1. Introduction

Cytomegalovirus (CMV) is a complex DNA virus with a genome comprising 229,354 base pairs. DNA sequences contain many patterns, and some of these patterns may flag important sites on the DNA, such as the origin of replication. The origin of replication is special patterns in the virus' DNA that contains instructions for its reproduction. To develop strategies for combating the virus, scientists want study the way in which the virus replicates. Our analysis uses data from the DNA sequence of CMV that was published in 1990 (Chee etal.) that is 229,354 letters long. Leung et al. (1991) implemented search algorithms to screen the sequence for many types of patterns. Altogether, 296 palindromes were found that were at least 10 letters long, with the longest being 18 letters.

The primary goal of this statistical analysis is to assess how the distribution of palindromes deviates from a uniform scatter across the DNA sequence and if any clusters are due to random chance. We found that [results of section 2.1 and 2.2]. Analysis of palindrome counts revealed that shorter intervals had greater variability than larger intervals and that certain region lengths (500, 20000, 50000) showed clustering, suggesting that these areas are less likely to occur by random chance. Further exploration of these clusters showed statistically significant clustering at region length 500.

The report is structured as follows: Section 1 introduces this study's objectives, background, and key questions, Section 2 analyzes the randomness of palindrome distribution, spatial patterns of palindromic sequences, palindrome counts in different genome regions, and the presence of significant clusters. Section 3 explores [advanced analysis topic]. Section 4 summarizes the findings and discusses possible methods for searching for the origin of replication.

2. Analysis

2.1

2.1.1 Methods

2.1.2 Analysis

2.1.3 Conclusion

2.2

2.2.1 Methods

2.2.2 Analysis

2.2.3 Conclusion

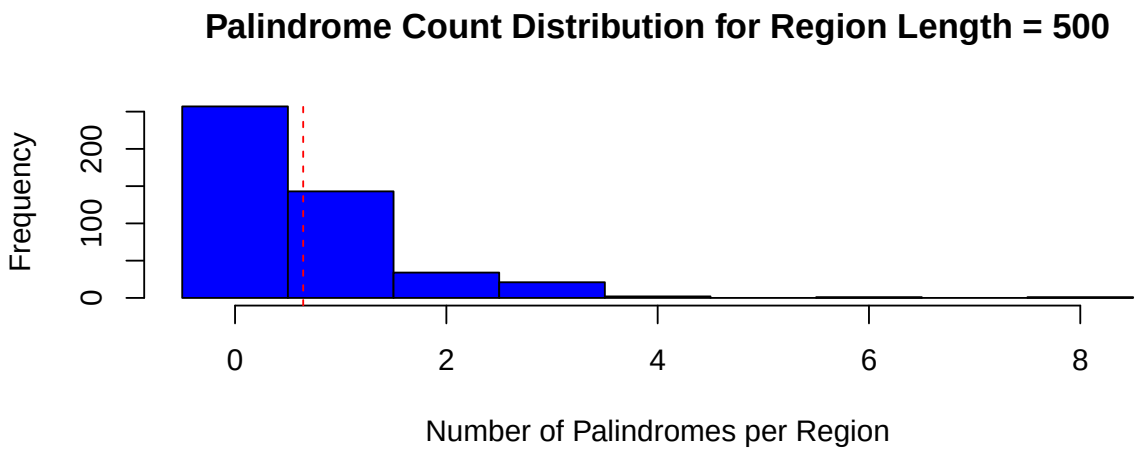
2.3 Palindrome Counts

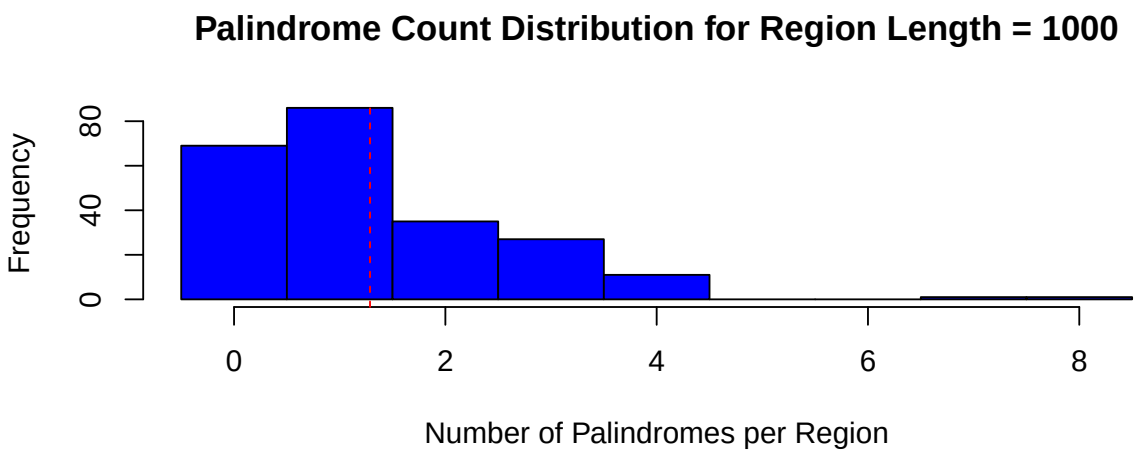
2.3.1 Methods

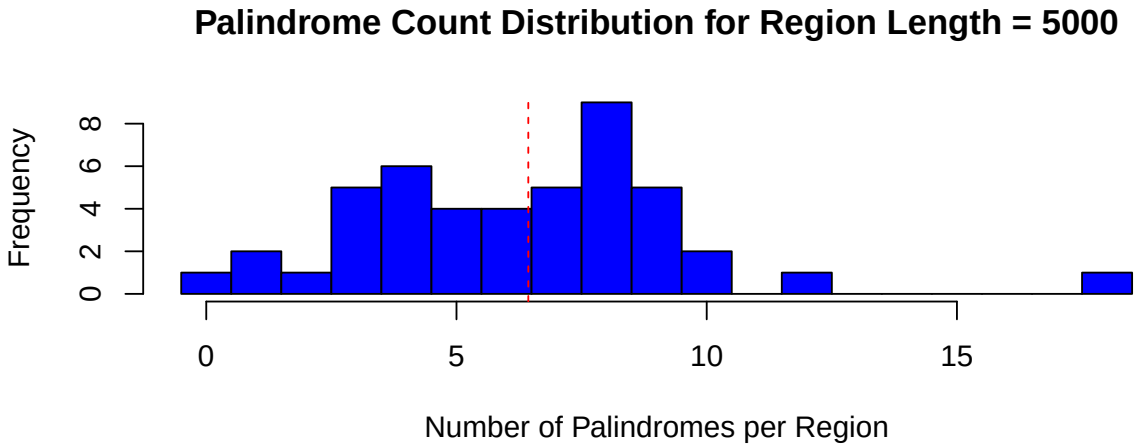
To examine palindrome counts in different regions of DNA, we split the DNA sequence into equal-length, non-overlapping regions, counted the palindromic sequences in each, and compared these counts to those expected from a uniform random scatter. To visualize the results, we graphed histograms to visualize the distribution of palindrome counts across regions. We then used Chi-Square Goodness of Fit Tests to compare observed palindrome counts to expected counts, testing the hypothesis of a uniform random distribution of palindromic sequences across regions.

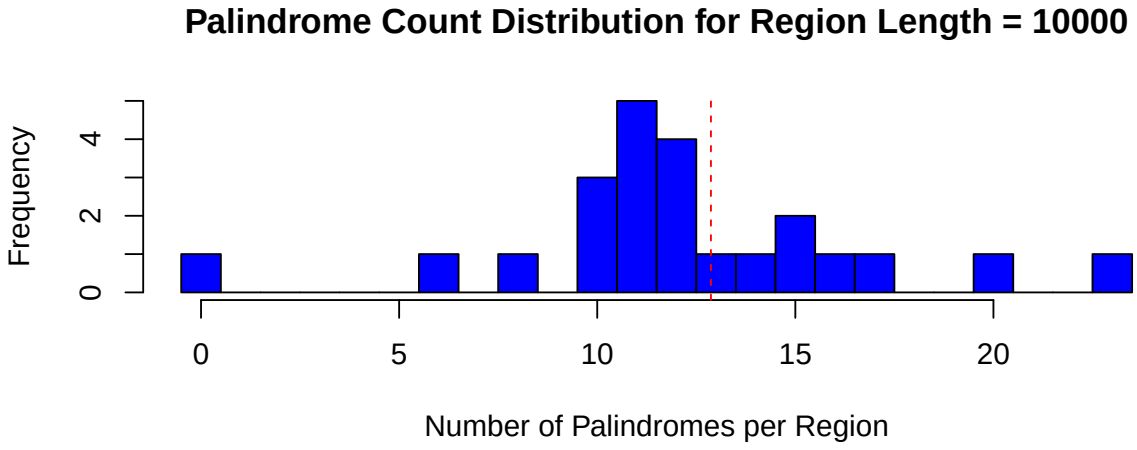
2.3.2 Analysis

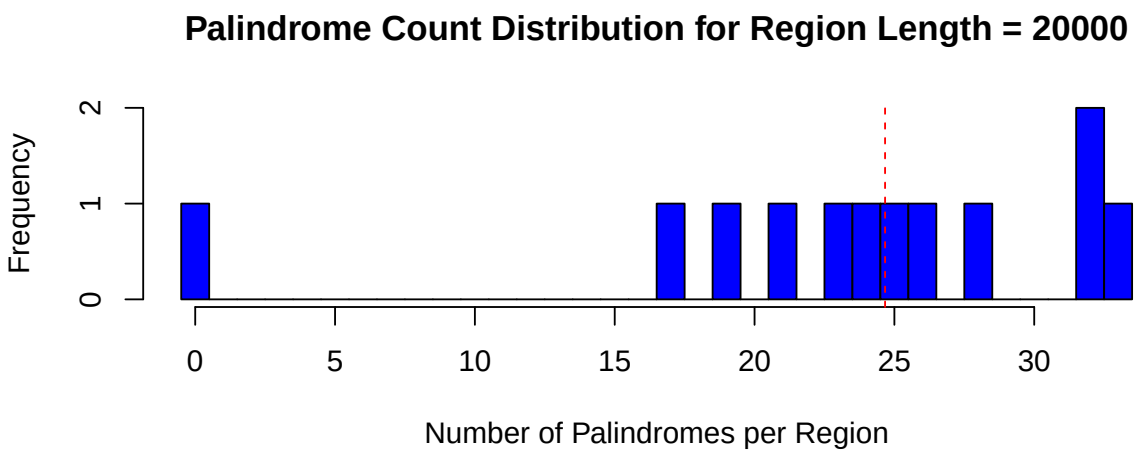
The histograms and Chi-Square Test results are shown below.




Palindrome Count Distribution for Region Length = 1000
Frequency
0
40
80
0
2
4
6
8
Number of Palindromes per Region


Palindrome Count Distribution for Region Length = 5000
Frequency
0
2
4
6
8
0
5
10
15
Number of Palindromes per Region


Palindrome Count Distribution for Region Length = 10000
Frequency
0
2
4
0
5
10
15
20
Number of Palindromes per Region



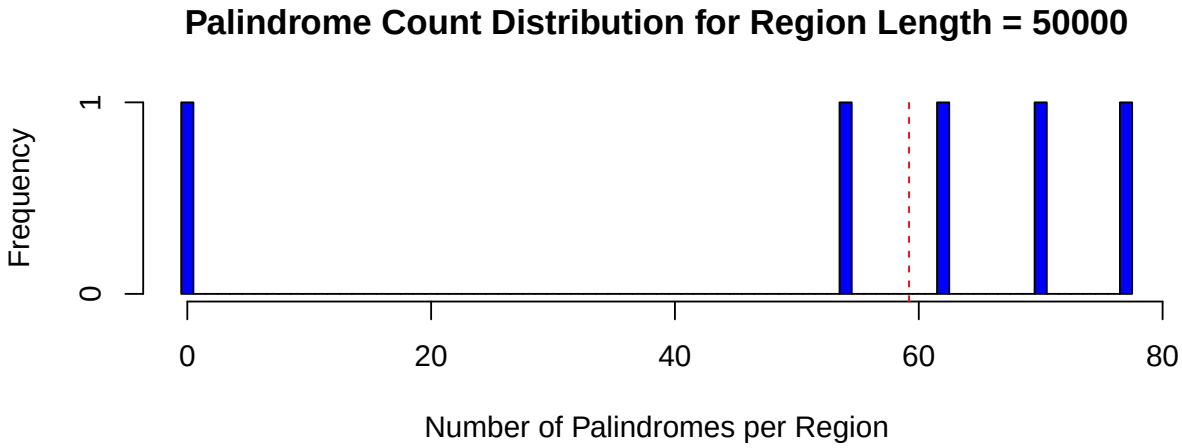


Table 1: Chi-Square Test Results

	Region.Length	Chi.Square.Statistic	p.value	Interpretation
X-squared	500	634.40541	0.0004998	Significant
X-squared1	1000	292.98649	0.0039980	Not Significant
X-squared2	5000	74.70242	0.0024988	Not Significant
X-squared3	10000	37.89286	0.0239880	Not Significant
X-squared4	20000	37.91429	0.0004998	Significant
X-squared5	50000	71.39163	0.0004998	Significant

2.3.3 Conclusion

The histograms of palindrome counts revealed that the smaller the region length, the more right skewed the count histogram was, which indicates that shorter regions have a higher frequency of intervals with low or zero palindrome counts, while a few regions have unusually high counts. This is consistent with the idea that smaller regions are more susceptible to variability in palindrome counts. Using a cut-off value of 0.001, the chi-square tests yield significant p-values for very small and very high region lengths, but insignificant

p-values for middle region lengths. The significant result at region length 500 suggests that palindromic sequences are clustered in certain small regions rather than being uniformly scattered across the DNA. The non-significant results at the intermediate scales (region lengths 1,000, 5,000, and 10,000) imply that, when viewed at these sizes, the distribution of palindromic sequences appears more uniform. At these lengths, local clusters of palindromic sequences are likely averaged out, resulting in counts that are more consistent with a uniform random distribution. The significant results at region lengths 20,000 and 50,000 indicate that there are large-scale patterns in the distribution of palindromic sequences across broad DNA segments.

2.4 The Biggest Cluster

2.4.1 Methods

We now want to investigate whether the interval with the highest number of palindromic sequences in a DNA sequence could suggest a potential origin of replication. To do this, we determined if a high-count palindrome cluster deviates significantly from a random scatter and whether such clusters align with features characteristic of origins of replication. Based on prior analysis, region lengths were selected to avoid both overly large and overly small intervals. Small intervals risk splitting clusters across adjacent regions, while large intervals may obscure tight clusters. For each interval, the number of palindromic sequences was counted, yielding a distribution of palindrome counts across intervals and the interval with the highest count of palindromic sequences was identified. Finally, Chi-square goodness of fit tests were applied to determine if the palindrome count in the highest-count interval deviates significantly from expected counts under a random distribution. Histograms were generated to visualize palindrome counts across intervals.

2.4.2 Analysis

The results of the Chi-Square Tests across different region lengths is shown below.

Table 2: Chi-Square Test Results

Region.Length	Max.Count	Max.Interval	Expected.Count	Chi.Square.p.value	Interpretation
500	8	186	0.6462882	0.0004998	Significant
1000	8	93	1.2925764	0.0074963	Not Significant
5000	18	19	6.5777778	0.0154923	Not Significant
10000	23	10	13.4545455	0.2838581	Not Significant
20000	33	5	26.9090909	0.3028486	Not Significant
50000	77	2	74.0000000	0.2088956	Not Significant

2.4.3 Conclusion

As shown in the table above, only the smallest region length (500) showed a significant p-value, while the larger region lengths (1000, 5000, 10000, 20000, and 50000) did not. This suggests that clustering of palindromic sequences exists on a small scale but not over broader intervals. The presence of a high number of palindromic sequences within an interval suggests non-random clustering, which may signal biological importance. Since the statistical analysis shows that the count in this interval is significantly higher than expected by chance, this could imply functional relevance, such as a replication origin.

2.5 Discussion

Our statistical analysis can provide insight on how to start experimentally searching for the origin of replication. The results showed a significant clustering of palindromic sequences in small regions (length 500), while

larger intervals showed a more uniform distribution. These findings suggest that the origin of replication may lie in one of these dense clusters of palindromic sequences within shorter segments of DNA. Given the statistically significant clustering at region length 500, biologists could begin the experimental search within these smaller regions.

3. Advanced Analysis

3.1 Methods

3.2 Analysis

3.3 Conclusion

4. Discussion