Unusual Cluster in the Palindrome

Benson Wu

February 19, 2019

Abstract

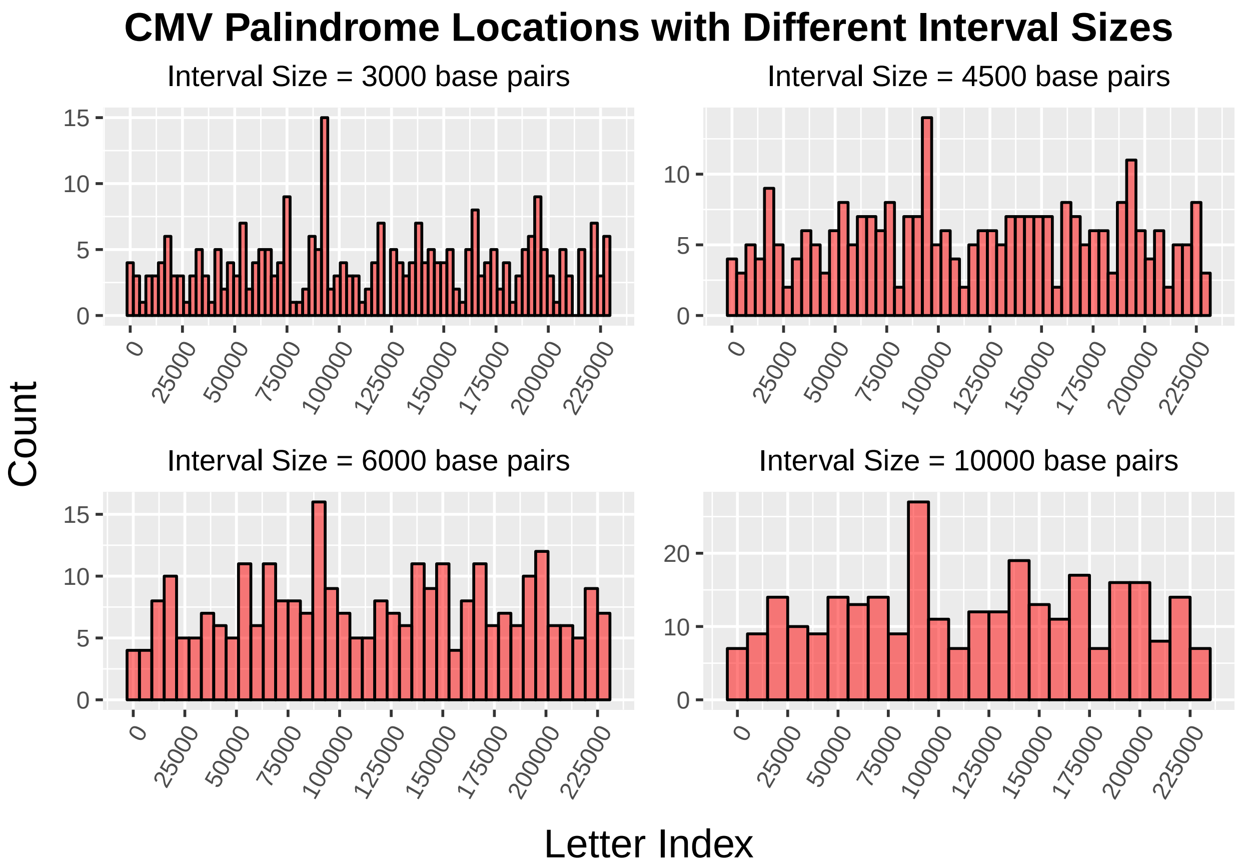
Method

Result

Conclusion

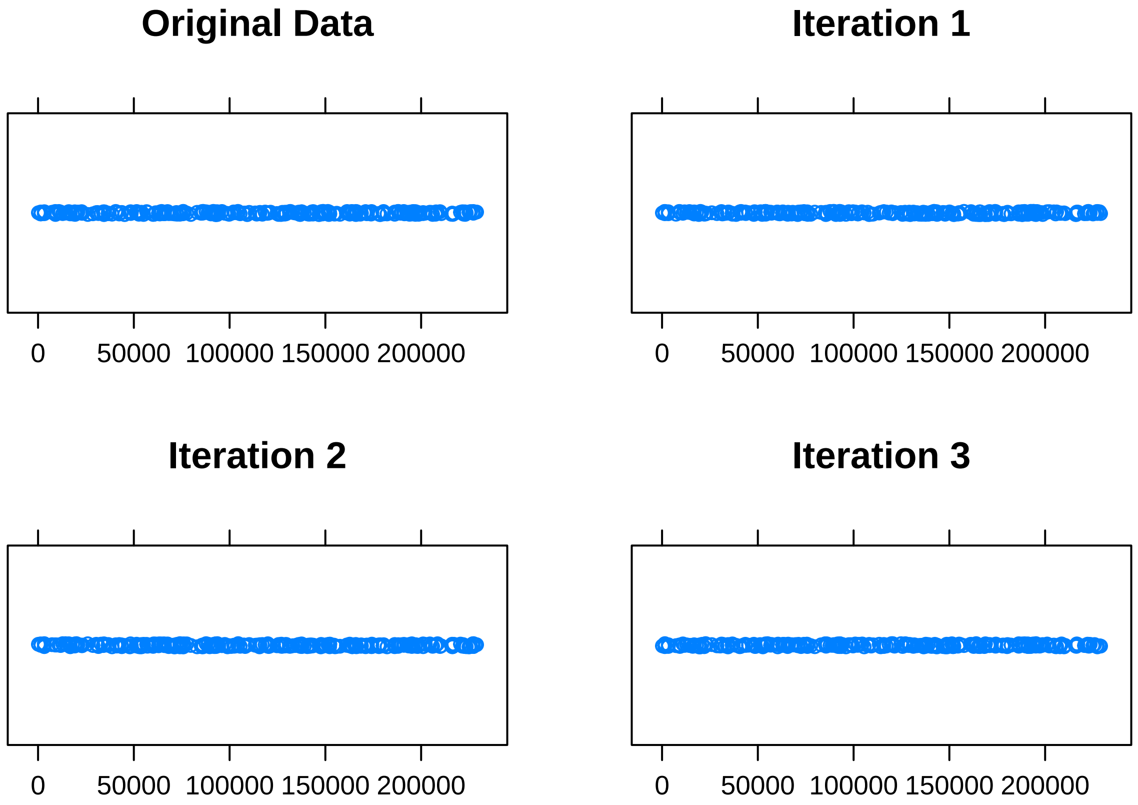
Background

DNA has a double helix structure made of two long chains of nucleotides. Each nucleotide is composed of a sugar, a phosphate and a base. The bases have four types: adenine (A), cytosine (C) , guanine (G), and thymine (T). In this analysis, we look particularly at the DNA sequence of Human cytomegalovirus (CMV). CMV is a common virus that infects people of all ages. When CMV is in a person's body, it stays there for life, but most people infected with CMV show no signs or symptoms. However, CMV infection can cause grave health problems for people with weakened immune systems and for unborn babies. Our data set contains the locations of 296 palindromes of length greater than 10 found in a particular CMV DNA sequence that is 229,354 letters long. It is hypothesized that being able to find origins of replication through The following figures show the distribution of palindromes across the DNA sequence with respect to different interval widths. We can see that the clusters of palindromes change as we increase the interval widths.



Random Scatter

We first begin by performing a random scatter through Monte Carlo uniform sampling to scatter 296 palindromes across a DNA sequence of 229,354 base pairs. Three simulations were done in order to ensure randomness. The simulations are compared visually compared to the location of palindromes in our original data. The following plot shows the 2D scatter of palindromes across the original DNA sequence and the simulated palindrome locations.



Original Data

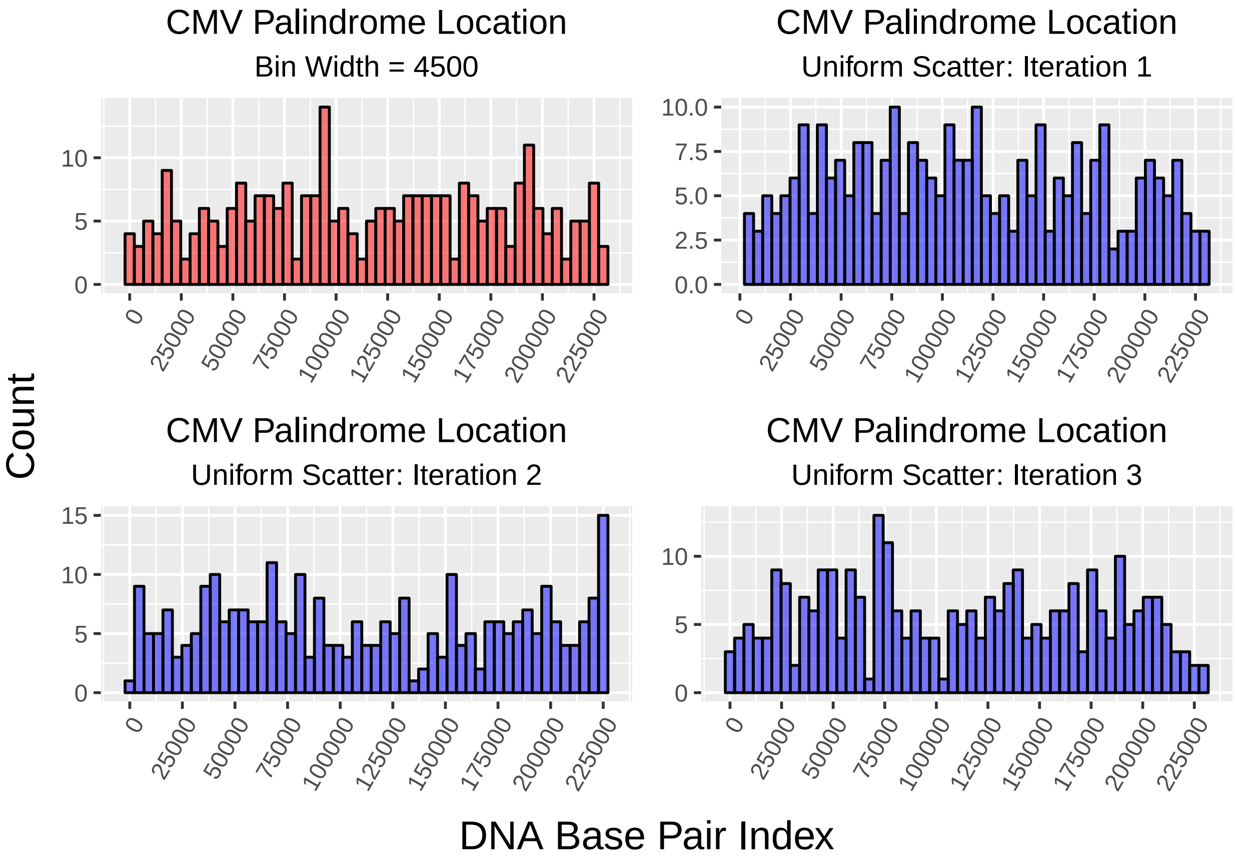
Iteration 1

Iteration 2

Iteration 3

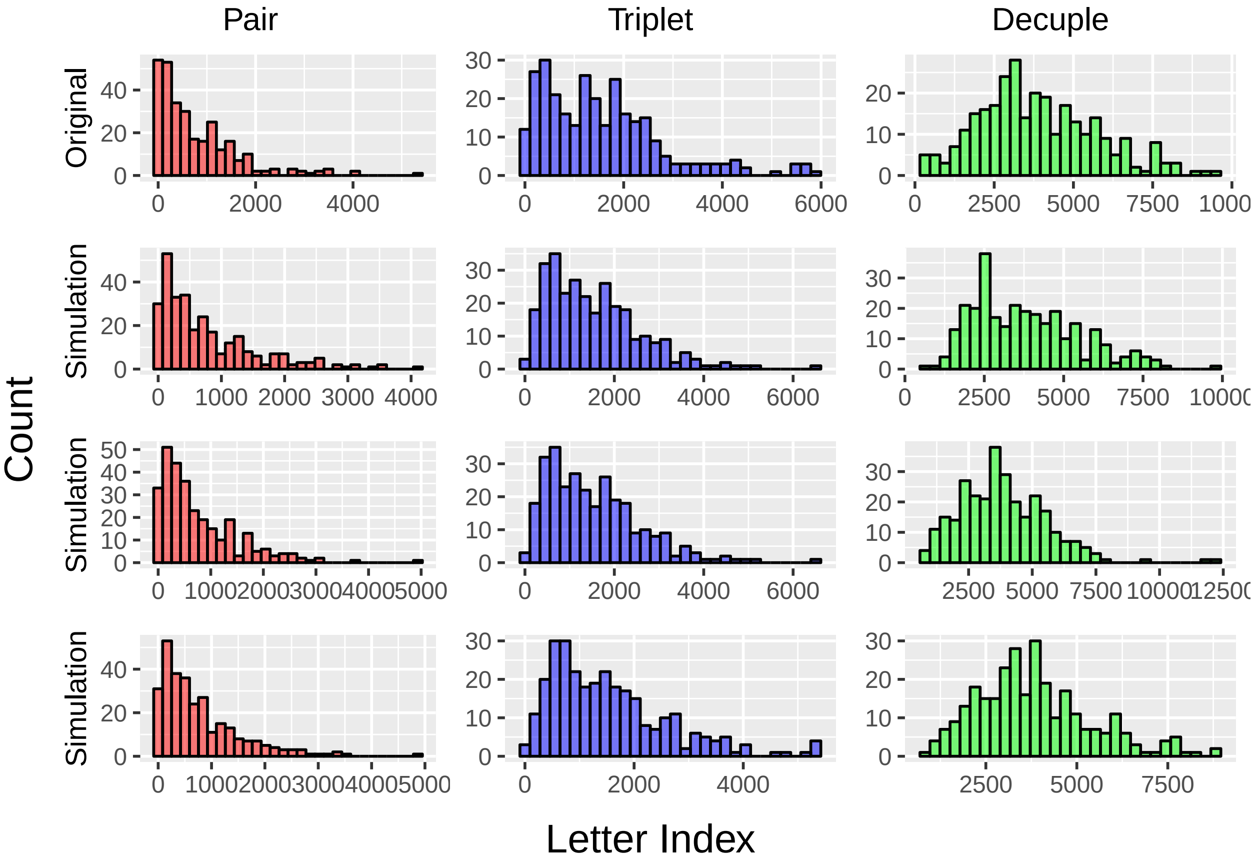
As we can observe, it is hard to tell if there are significant clusters because the simulated palindrome locations seem to be spread evenly across the DNA sequence, indicating that the original data departs from a uniform scatter of palindromes across the DNA sequence.

The following figure shows the distribution of palindromes across the DNA sequence with the interval size set to 4,500. Again, we see that there is no significant cluster of palindromes in the simulated data unlike the palindrome locations in the original data. This leads us to investigate the location and spacing of palindromes next.



Location and Spacings

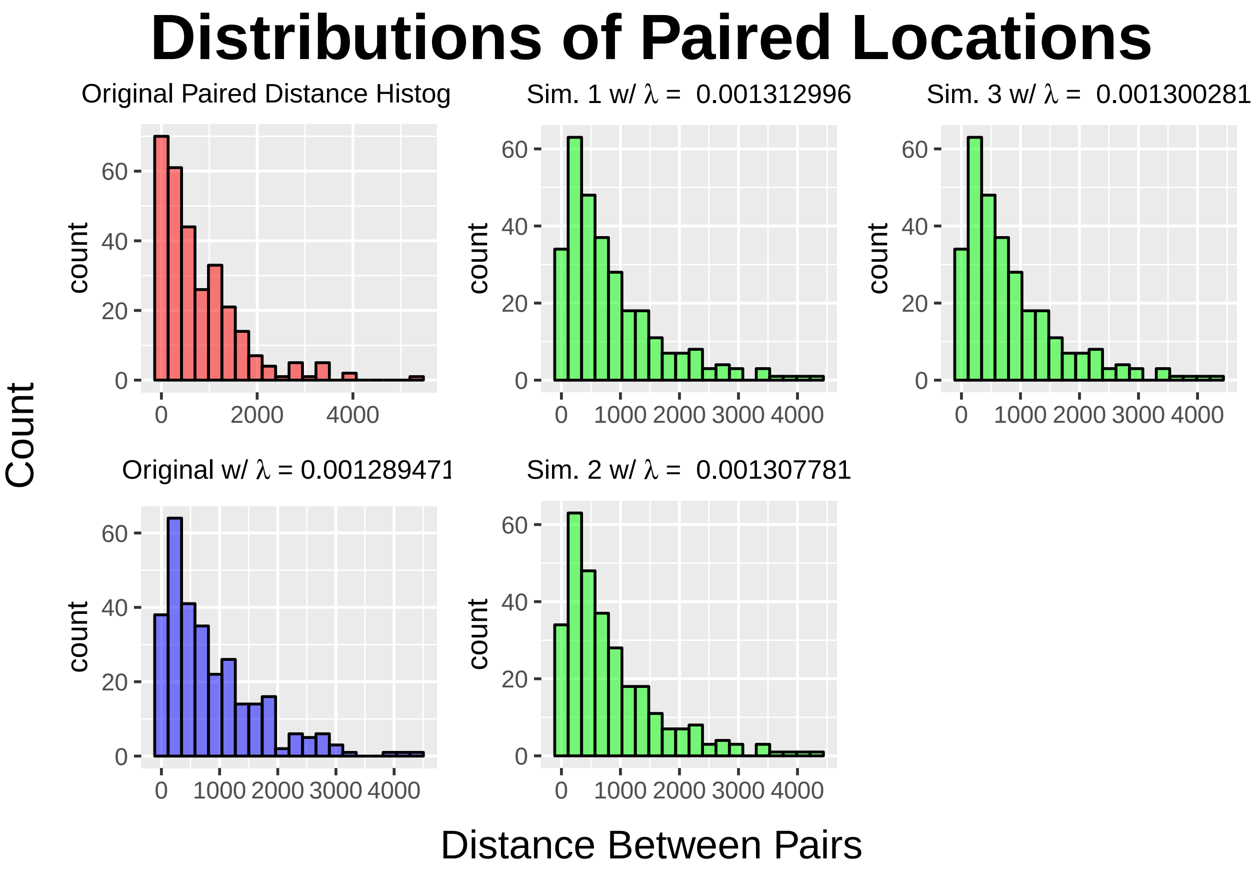
We examined the spacings between various intervals of palindromes to see if they follow any particular distributions. To ensure replicability, we examined the spacings not only in the original palindrome data, but also in our three palindrome simulations as well. We made the choice to perform this part of the analysis using an interval size of 4,500 DNA base pairs. We compared the distances between pairs, triplets, and decuples. The following figure shows the distributions of the different groupings.



As we can see in the figure, we see the distribution of spacing between pairs of palindrome appears to follow an exponential distribution. We use Maximum Likelihood Estimation to estimate the lambda parameter of the paired spacing in our original data and three simulations, such that we solve for $\hat{\lambda}$ in the first order condition of the log-likelihood function of the exponential distribution and obtain the following values.

Table 1: Exponential MLE of $\hat{\lambda}$

Data	$\hat{\lambda}$
Original	0.0012895
Simulation 1	0.0013130
Simulation 2	0.0013078
Simulation 3	0.0013003



In the figure above, we can see how a generation of random deviates with the respective estimated $\hat{\lambda}$ compares to the distribution of paired spacing in the original data. Interestingly, we see that for random deviates of the various $\hat{\lambda}$, the behavior of the distributions don't contain as many values in the first interval as expected for an exponential distribution. This leads us to investigate whether the palindrome locations from the original data actually follows a Poisson distribution and whether the paired spacing in the original data actually follows an exponential distribution.

Counts

The Biggest Cluster

Theory

Maximum Likelihood Estimation

Poisson Process

Chi-Square Test

Poisson

$P\left(x\right)=\frac{e^{-\lambda}\lambda^{x}}{x!}$

$\mathcal{L}(\lambda\,;x_1\,,..,x_n)=\prod_{i=1}^{\infty}\frac{e^{-\lambda}\lambda^{x_j}}{x_j!}$

$\hat{\lambda}$