Darren Chang
Shiyin Liang
Anthony Martinez

Homework 3

Introduction

Now, more than ever, it is important to find the best ways to combat new viruses. With many viruses that pose a huge threat, it is important to figure out how viruses replicate in order to stop this process. This is why many scientists are in search of the origin of replication in a virus' DNA. In this paper, we will search for complimentary palindromes in cytomegalovirus' (CMV) DNA in order to advise a biologist who is searching for the origin of replication. CMV is a potentially life-threatening disease to people with weakened immune systems, so it is extremely important to figure out how to prevent the virus from replicating.

In order to advise biologists on this problem, we will examine palindrome locations of CMV DNA (hcmv.txt). The CMV DNA is 229,354 letters long and altogether there are 296 palindromes between 10 and 18 base pairs long. CMV is in the same family as Herpes simplex and Epstein-Barr virus. These other viruses are marked by a long palindrome and repeat clustered at the origin of replication. We will look for a similar pattern in CMV in hopes that is where the origin of replication is located.

We determined that a large amount of palindromes are found in the beginning of the DNA sequence, so this area could possibly be the origin of replication for CMV. The information that we found from these conclusions would be given to biologists in order to help them find the origin of replication in CMV.

The rest of this paper will go over four different scenarios of analysis that allowed us to reach this conclusion. Similarly, there will be a discussion focused on the meaning of these results and future proposals for this project.

Analysis

1. Locations

To make it easy to see the distribution of palindromes across the DNA sequence, we created three histograms. Figure 1 shows the distribution of palindromes of our sample and we find signs of an unusually high amount of palindromes around the 90,000 and 19,000 mark. Compared to Figure 3, which shows what a uniform distribution of palindromes would look like, Figure 1 doesn't look like a uniform distribution. Figure 2 shows the distributions of 5 instances of randomly generated palindrome sites and there are no clear outliers like in Figure 2. Note that Figure 2 shows no locations with more than 12 palindromes whereas Figure 1 shows two locations with more than 12 palindromes. This indicates that the clusters at the 90,000 and 19,000 mark is a departure from the usual structure of the DNA chain and may not be due to random chance as Figure 2 shows us instances of what random uniform scatters look like. These two locations in our sample may be a potential replication site.

Figure 1:

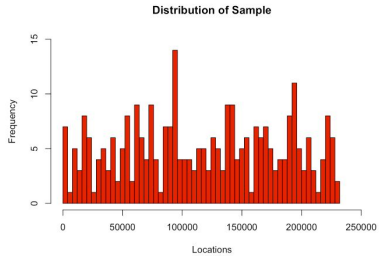


Figure 2:

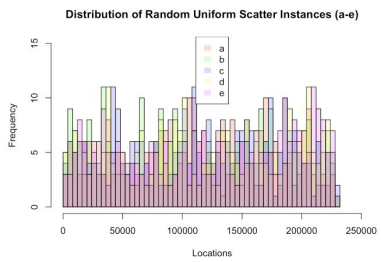
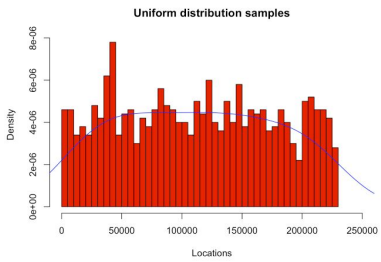


Figure 3:



Conclusion

The distribution of palindromes in our sample seems to be unique and not due to random chance. When compared to the 5 different random distributions and the uniform distribution, we can see different peak locations where there are a high number of palindromes. Importantly, we also found the locations of interest in our sample that we should look more into--the 90,000 and 190,000 marks. There is also a strong repetition of clusters of palindromes in the first 100,000 marks. By looking at the location of palindromes in our sample, we would advise biologists to investigate the sites of unusual activity as well as narrow their search in the first 100,000 instances of our sample.

2. Spacings

To graphically examine the distribution of our sample spacing, we created Figure 4 to show 5 randomly sampled distributions of spacings between consecutive palindromes, Figure 5 to show 5 randomly sampled distributions of spacings with one palindrome in between other palindromes and Figure 6 to show 5 randomly sampled distributions of spacings with two palindromes in between other palindromes. Figure 4 shows the frequency of palindromes reaching almost 150 counts at the 500 mark location. Figure 5 shows the frequency of palindromes at 60 counts at the 500 mark and has a more uniform distribution compared to Figure 4. Figure 6 shows the frequency of palindromes at around 50 counts for all the locations and has the most uniform distribution of all three scenarios. Note that in all three scenarios, the highest amount of palindromes all occur at the smaller values of x, or the left part of the graphs. Since we know that distances between successive hits follow an exponential distribution, and distances between hits that are two apart follows a gamma distribution with parameters 2, λ and distances between hits that are three apart follows a gamma distribution with parameters 3, λ, this makes sense because smaller value inputs for exponential and gamma distributions give larger outputs than larger value outputs.

Figure 4:

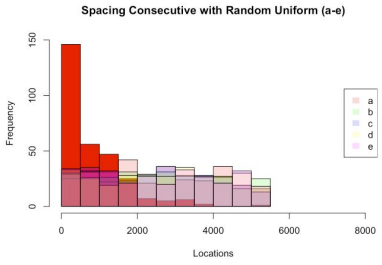


Figure 5:

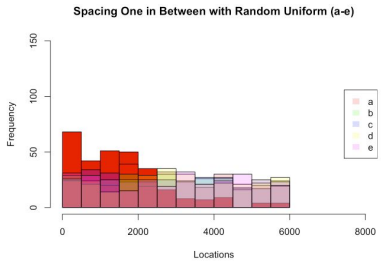
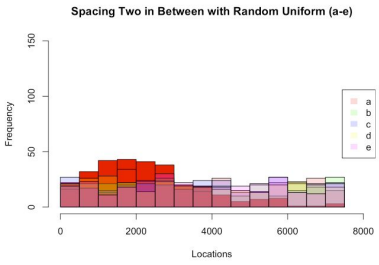


Figure 6:



Conclusion

When the spacing between palindromes is consecutive, then there is a large frequency of palindromes around the 500 mark. When the spacing between palindromes is one or two, then the frequency at each location becomes very similar. When the spacing between palindromes is consecutive, it helps us focus on one location where there is a large number of palindromes. This location where there is a high number of palindromes could be useful in determining the location of the origin of replication for CMV.

3. Counts

Here we will be using a statistical method to examine counts of palindromes in different regions. Instead of looking for a trend by using all 296 palindrome locations we will be using 4 different numbers of regions to divide all the genes and we will be using 20, 30, and 40 regions. For each number of regions we will be plotting histograms to compare the distributions to a random uniform scatter distribution. We will also perform chi-squared tests to determine if the distributions of the palindromes are randomly scattered. For all 3 regions, our null hypotheses are that the site locations we observe follow random scatter/Poisson distribution and any unusual activity is due to chance. Our alternative hypotheses are that what we observe is not due to random chance.

For 20 regions, our chi square test resulted in a p value of close to 1 which is higher than our alpha level of 0.05. Thus, we fail to reject the null hypothesis for 20 regions. For 30 regions, our chi square test resulted in a p value of close to 0.086 which is higher than our alpha level of 0.05. Thus, we fail to reject the null hypothesis for 30 regions. For 40 regions, our chi square test resulted in a p value of close to 0.0049 which is lower than our alpha level of 0.05. Thus, we can reject the null hypothesis for 40 regions. In 40 regions, the count of palindromes don't follow a Poisson distribution, while we cannot say definitively that 40 regions are the site for replication, there is an unknown reason or variable causing the palindromes to occur more in 40 regions.

Conclusion

Based on our chi-squared test, we can determine that our data sample is random or more specifically, follows the Poisson distribution for the regions 20 and 30. However, in our results for 40 regions, it seems that our sample doesn't follow random scatter. Researchers should look more deeply into 40 regions rather than 20 and 30 regions to have a better chance of finding the origin of replication.

4. Biggest Cluster

Using the similar process for the analysis in part 3 (Counts) we are able to analyze different sized clusters of palindromes in our data. Earlier we found that our data follows the Poisson distribution and we wanted to test and see how this trend does with more or less regions. We see that from 40 regions up the P-Value is far below the 0.05 significance level and it seems to stop decreasing at 0.0004998. This further supports our idea that the trend is towards a Poisson distribution. On the other hand, when we decreased the number of regions to 10 the P-Value score was the same as 20 regions: 1. This process we used can be applicable for biologists who need to examine distribution trends within their genomic datasets. Giving them a process to go by to find interesting trends in their data.

Number of regions	**Lambda**	**P-Value**
10	**29.6**	**1**
20	**14.8**	**1**
30	**9.8666**	**0.08796**
40	**7.4**	**0.0004998**
50	**5.92**	**0.0004998**
60	**4.9333**	**0.0004998**

Conclusion

The overall trend of the sample is Poisson distributed so the palindromes are randomly scattered throughout the samples. We were able to prove this by decreasing and increasing the number of regions in order to find a trend in the P-Values. This will help biologists be able to find trends in their clustering for further data.

Advanced Analysis: Cumulative Distribution Function (CDF) Plot

We decided to use cumulative distribution function (CDF) plots in order to get better insight to our data. A CDF plot shows the empirical cumulative distribution function of the data. The empirical CDF shows the proportion of values less than or equal to X, where X is the location of palindromes. We chose CDF plots for our advanced analysis because they are useful for comparing the distribution of different sets of data. We created CDF plots for our data sample, a random distribution, and a uniform distribution. Figure 10 shows the CDF plot for our data sample. In Figure 10, we can see that the distribution of palindromes is spread linearly across the locations, however, there are many steps in the plot. While in Figure 11, we can see the distribution of palindromes has abnormal steps in it. In other words, the locations where palindromes can have a large count is not normal. In Figure 12, we see the distribution of palindromes is linear. There aren't any steps in the plot, so the data is similar to a normal distribution.

Figure 10:

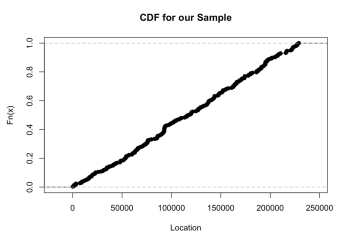


Figure 11:

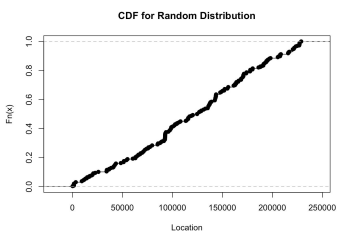
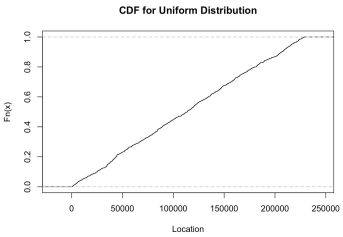


Figure 12:



Conclusion

When comparing the distributions of our data sample, a random distribution, and a uniform distribution we found out that the distribution of our sample is slightly different than a true random distribution. This shows that the distribution of our sample is unique. Knowing that our sample does not follow random distribution or uniform distribution, we can assume that there is an abnormal location that could possibly contain the origin of replication in CMV.

Conclusion

Our objective for this report was to aid a biologist who is about to start searching for the origin of replication in CMV. In order to do this, we looked for information in the data that would be useful to find a particular location in CMV's DNA where the origin of replication could likely be. From our analysis, we found out that our data does not follow a random distribution or a uniform distribution. This is important for a biologist to know because it shows us that we can narrow our search to a particular location. Any abnormal location where there is a high count of palindromes, or consecutive locations with palindromes could indicate the location for the origin of replication. From our location analysis, there is a large count of palindromes at the 90,000 mark, so this would be a good first place for a biologist to check. From our spacing analysis, there is a large count of palindromes at the 500 mark, when the spacings between palindromes are consecutive. So a biologist should look at palindromes with consecutive spacings in order to optimize finding a location with a large count. From our count analysis, we found out our data follows a poisson distribution at 20 regions. However, when the data is separated into 30 regions, the data is abnormal and has a higher chance of having a large count of palindromes. From our biggest cluster analysis, it is best to separate the data into 40 or 50 regions in order to find the largest count of palindromes. A biologist should follow our procedures to have the best chance at finding a location with a high count of palindromes. A high count of palindromes would likely lead to the location of the origin of replication in CMV.

In the future we can improve this project by gathering more data since we only had 296 data points to work off of. We can also continue this project by verifying whether the results the biologists have found after taking our advice are truly the origins of replication by using similar statistical methods. We can also expand this project to help pinpoint areas that are most likely to be origins of replications for other diseases as well. We can also expand on this project by creating a program that finds areas of unusually high counts of palindromes that generalizes to all diseases and genome sequences.

Appendix

1. hcmv.txt
2. https://www.ncbi.nlm.nih.gov/pubmed/1666311
3. https://analyse-it.com/docs/user-guide/distribution/continuous/cdf-plot

Contributions

Darren: Wrote code for locations, spacings, count, and biggest cluster, wrote analysis for biggest cluster
Shiyin: wrote analysis for part 1, part 2 and part 3 and conclusion
Anthony: Wrote code for advanced analysis, wrote introduction, worte analysis for locations, spacings, and advanced analysis, and part of conclusion