Xueru Xie, Yichi Zhang, Jinsong Yang

Professor Schwartzman

MATH 189 Report 3

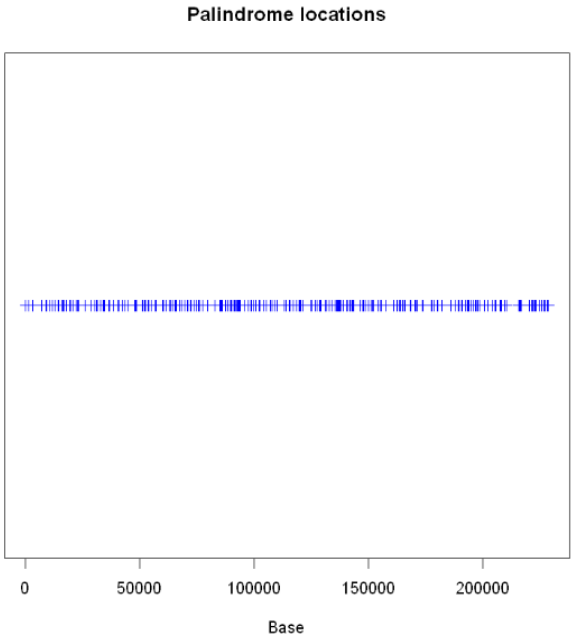
16 May 2020

Find DNA Pattern Using Clustering

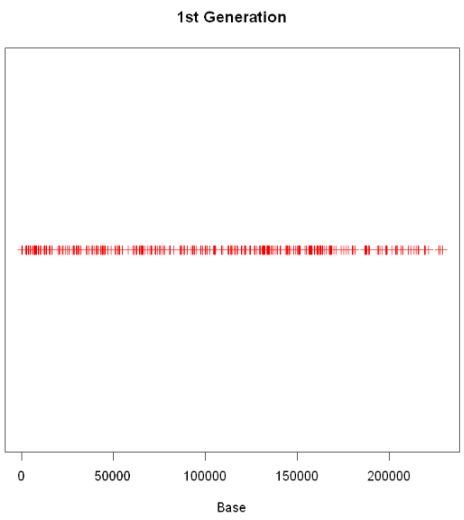
In lecture 3, the class learned about the DNA pattern of the disease human cytomegalovirus (CMV). CMV is a disease that can be life-threatening. In order to attack the disease successfully, scientists study the virus's DNA in order to find out how it reproduces. CMV DNA is 229,354 letters long, and it contains palindromes. Palindrome is a sequence of letters that reads the same forward and backward, such as 1001001, taco cat, or ACGTGCA. The scientist's goal is to find the origin of replication, then test to see if it can replicate. Our task as data scientists is to find an efficient way for scientists to find the origin of replication as quick as possible. Given the dataset hcmv.txt, which contained 296 locations of the palindromes of CMV, we wanted to see if the dataset came from a random uniform scatter. If not, then we could search for unusual clusters in the strand of DNA, hoping to find the best location for scientists to look into.

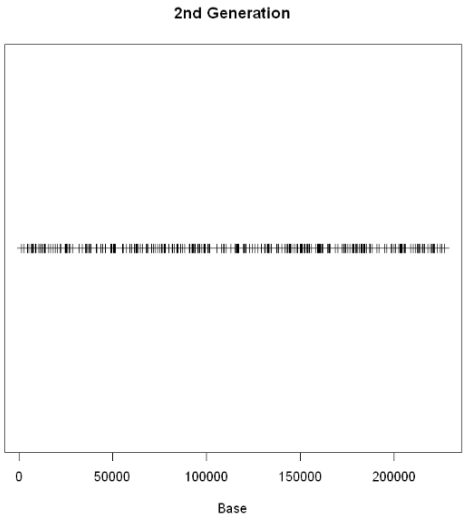
Random Scatter

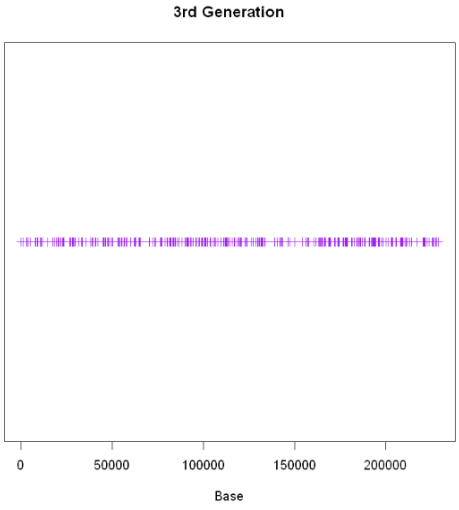
To look for the structure of the locations of palindromes, we opened the given data in R. The data consisted of a table of 1 column, which were numbers that represent the location of the palindrome inside the CMV DNA. To begin, we imported the dataset, then we used stripchart() in R. This was the result.

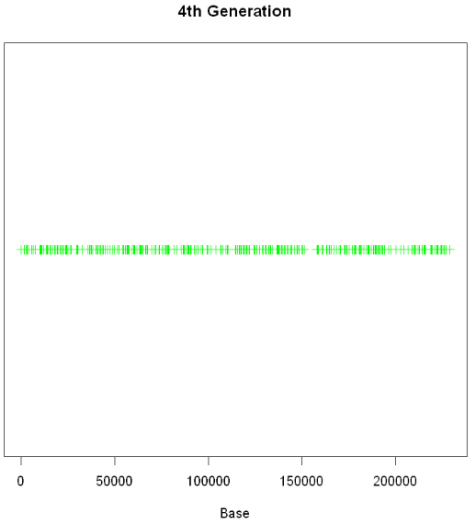


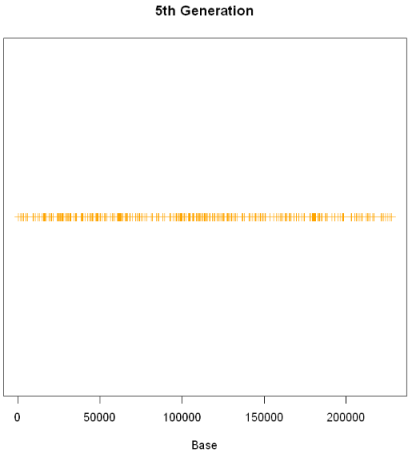
Then, we sampled from the palindrome locations 5 and used strip plot again to see the distributions. Our method of doing that was the runif() function in r. That way we could compare the sample palindrome locations to random uniform scatter. Based on the graphs, we could see that the given data palindrome locations were nearly distributed evenly. But we could see that there were more palindromes near the 100,000th location. However, this could also be due to random chance, we could not make an inference based on this visualization.


1st Generation
0
50000
100000
150000
200000
Base


2nd Generation
0
50000
100000
150000
200000
Base

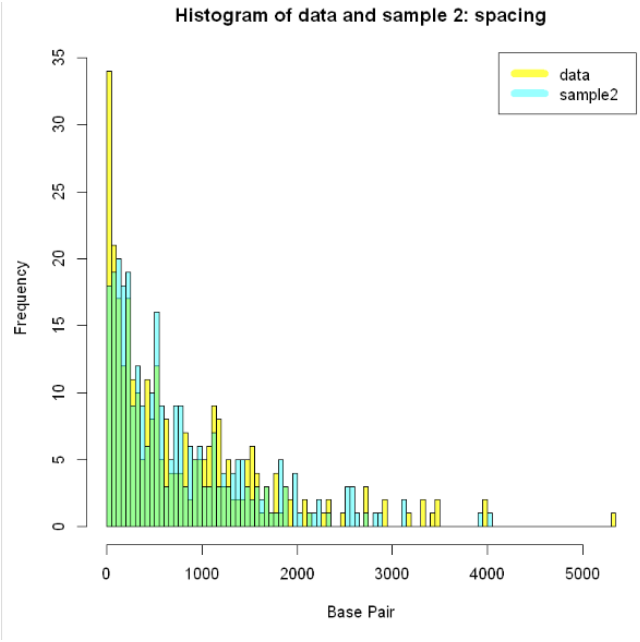

3rd Generation
0
50000
100000
150000
200000
Base

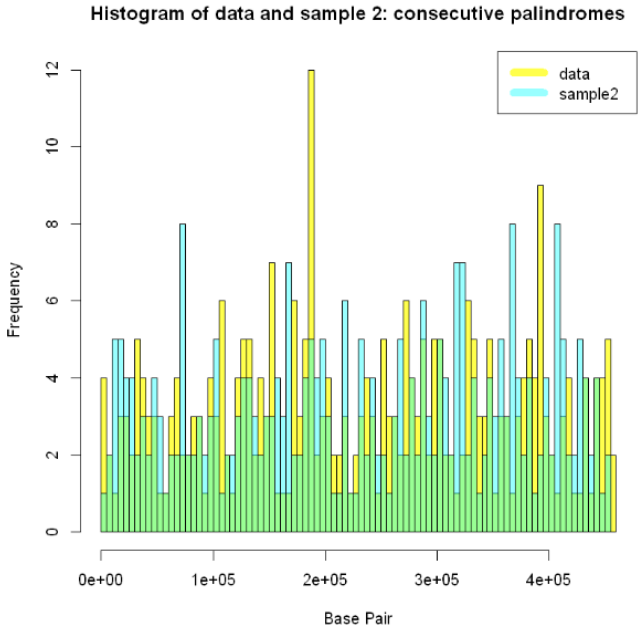

4th Generation
0
50000
100000
150000
200000
Base

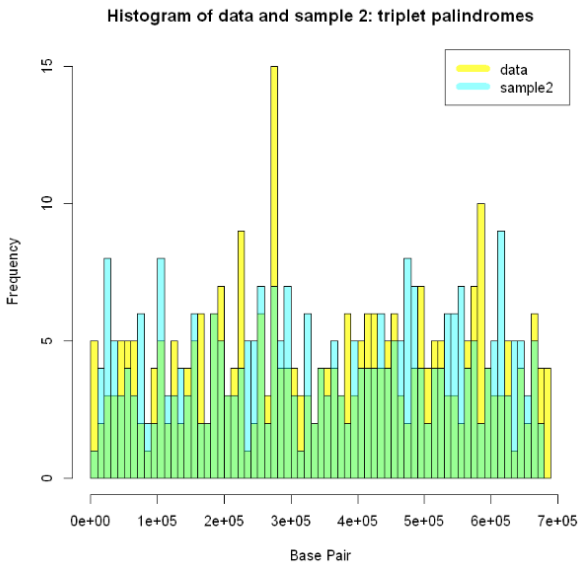

5th Generation
0
50000
100000
150000
200000
Base

Location and Spacing

There were 3 types of spacings that we used to examine: spacings between consecutive palindromes, spacings between palindromes with one in between, and spacings between palindromes with two in between. We graphed all of the spacings above, with the given location of palindromes. We also picked the second sample from the 5 samples above (random). Here were the graphs below. We overlaid both output together in every spacing.


Histogram of data and sample 2: spacing
data
sample2
Frequency
0
5
10
15
20
25
30
35
0
1000
2000
3000
4000
5000
Base Pair

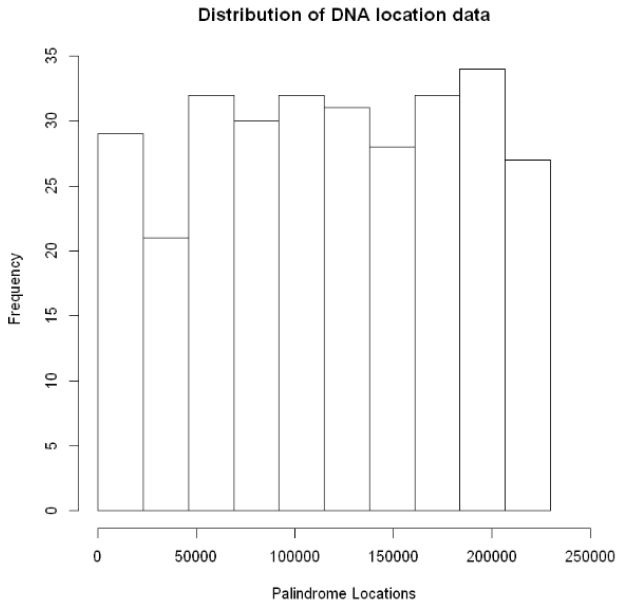

Histogram of data and sample 2: consecutive palindromes
data
sample2
Frequency
0
2
4
6
8
10
12
0e+00
1e+05
2e+05
3e+05
4e+05
Base Pair



When we examined the spacings between consecutive palindromes for both data, they looked pretty similar in terms of the frequency, although the spacings of the given data had a lot of frequency values in the beginning of the location of palindromes. For the sum of pairs of consecutive spacings and triplet spacings graphs, the given data had a lot more frequency near the 200,000 and 300,000, while the sample distributions were fairly spread out. Based on the result, there could be a group of clustering from the given CMV DNA, which their sum of pairs of consecutive spacings and triplet spacings were mostly around 200,000 and 300,000. This phenomenon could be validated with the previous strip plots because we noticed there were clusters of palindromes around the 100,000 location. So the consecutive sum of the cluster palindromes would be about 200,000 and their triplets would be around 300,000.

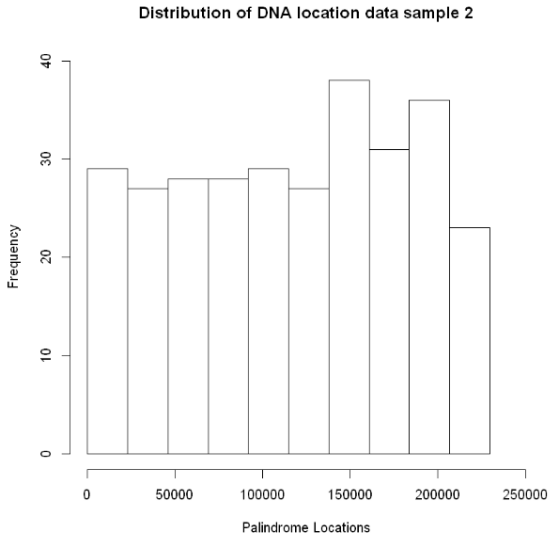
Counts of Palindromes in Regions of the DNA

In addition to examining the spacings between palindromes, we also observed the counts of palindromes between different intervals. This could be important for searching the clustering because we could see how the observed counts could differ than the expected counts. Since there were 229,354 codes of DNA, we divided the total into 10 bins, then we examined the number of palindromes in each bin.



observed	expected
29	29.6
21	29.6
32	29.6
30	29.6
32	29.6
31	29.6
28	29.6
32	29.6
34	29.6
27	29.6

We also looked at the distribution of sample 2 by counts.



observed	expected2
29	29.6
27	29.6
28	29.6
28	29.6
29	29.6
27	29.6
38	29.6
31	29.6
36	29.6
23	29.6

From both histograms, we could see that the counts were fairly distributed within the intervals for both the given data and uniform sample. Then we applied the Poisson model to make a comparison because this model is for uniform random scatter test. The lambda in this Poisson model was calculated by total palindromes divided by the number of regions. Here were the results for intervals 3000, 4000, and 6000 of the given dataset.

```
homog_pois_process(3000,dna)[3]
```

	Palindrome_Counts	Number_Observed	Number_Expected
1	1	5	6.1420509
2	2	2	11.8800196
3	3	8	15.3189726
4	4	10	14.8150590
5	5	9	11.4621772
6	6	8	7.3900880
7	7	5	4.0839960
8	8	4	1.9748270
9	9	4	0.8488292
11	11	1	0.3283629
14	14	1	0.1154769

```
homog_pois_process(4000,dna)[3]
```

	Palindrome_Counts	Number_Observed	Number_Expected
1	1	5	1.6916173
2	2	2	4.3625919
3	3	8	7.5005966
4	4	10	9.6718219
5	5	9	9.9772478
6	6	8	8.5769324
7	7	5	6.3198449
8	8	4	4.0746368
9	9	4	2.3351720
11	11	1	1.2044571
14	14	1	0.5647694

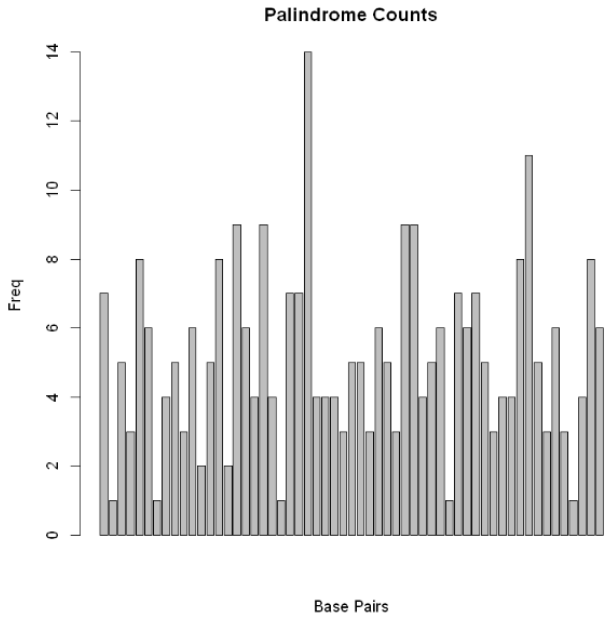
```
homog_pois_process(6000,dna)[3]
```

	Palindrome_Counts	Number_Observed	Number_Expected
1	1	5	0.1283156
2	2	2	0.4963788
3	3	8	1.2801348
4	4	10	2.4760503
5	5	9	3.8313620
6	6	8	4.9404405
7	7	5	5.4604868
8	8	4	5.2808655
9	9	4	4.5396914
11	11	1	3.5122876
14	14	1	2.4703649

Based on the result, there were differences between observed values and expected values. Then, we could proceed to use Chi-squared test to test our null hypothesis: the given DNA data came from random uniform scatter.

Largest Cluster of Palindromes in a Sub-Interval

Using Chi-squared test, we got P-values less than 0.05 for all of the intervals. Since all of the results were favoring against the null hypothesis, we only graphed one of the results.

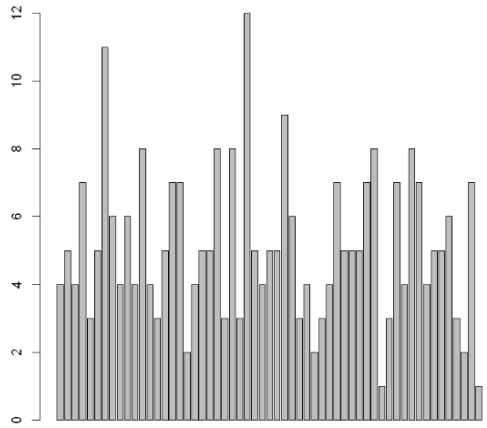


Looking at the graph above, we could clearly see a big spike in the middle. Referring back to the table, we could see that the spike happened from around 92,000 to 96,000, with frequency 14.

(9.2e+04,9.6e+04]	14

In addition, for all tested base lengths(3000,4000,4500,6000), the secondary largest frequency all falls in(1.92e+05,1.96e+05], which also suggests a likely replication site in addition to the one discussed above.

Then we wanted to perform the same procedure to one of our random scatters. We chose sample 3 as our input.



As you can see, although there was a spike at the similar location as the given DNA data, the frequency in this one was not as many as the given data. The overall shape of the distribution was even. This could mean that the given DNA data might not be a random uniform scatter, and the unusual spike could be the possible origin of replication.

Limitations

While performing the chi square test, we grouped occurrence numbers of 9 and more into one group. During the test, we levitate such a threshold to 11 and get a vastly different p value, which we eventually abandon. This implies that a swift change in one of the sites may drastically influence our test result.

Conclusion

In the beginning, we did end up getting a close distribution of palindromes from the scatter plots first. For spacing and location, our sampled graph indicated that the CMV DNA data was similar to our sampled data but there were some parts (near region 100,000) of the graph that the DNA data had more frequency than the sample ones.. After we applied the Poisson model and Chi Squared test, we were able to conclude to see an unusual frequency at around region

from 92,000 to 96,000. Lastly, the largest cluster of palindromes all had similar p-values which indicates that the experiment favored in rejecting the null hypothesis. Based on this result, scientists could look into a region from 92,000 to 96,000 to detect possible replication sites. For the future, we could apply this similar procedure on finding other clusters inside a DNA strand. That way scientists would be able to find the origin of replication in a more efficient way.

Appendix

In this report, student Xueru Xie wrote and coded scenarios 1 to 2. Student Yichi Zhang had the responsibility to write all the R codes from scenarios 3 to 4. Student Jinsong Yang completed scenarios 3 to 4 and the conclusion section. The coding PDF contained all the work from scenarios 1 to 4. During our procedure, we encountered a problem when we approached scenario 3, counting. We were not able to compute the chi-square test as quickly as we thought.