Genomics and Bioinformatics Project 1

Hamd Bilal Tahir

Gang Fang

Introduction

Our project was to propose and answer the question that how is an extremophile bacteria different from its closely related species that lack the extremophile characteristics. For this reason, we wanted to see that was there any dinucleotide or dipeptide usage bias in the extremophile genomes. By that, we wanted to see that were there and dinucleotide or dipeptide usage differences between extremophile species and non-extremophile species. Based on the idea that there may be certain proteins found in extremophiles that non-extremophiles lacked, and vice versa, we hypothesized that there may be dinucleotide or dipeptide usage bias between extremophiles and non-extremophiles.

Materials and Methods

We started off by taking an extremophile bacteria Thermotoga Maritima MSB8 (*T.* Maritima MSB8) as our species of focus. This bacterium is hyperthermophilic, with an ideal living environment being water at a temperature of 80^0C. It is the only bacterium known to grow at this high a temperature.

We downloaded *T*. Maritima MSB8 genome from the NCBI genome database website. The proteome sequence was downloaded from the Uniprot database from the Embl website. The genome and proteome sequences were parsed to form dinucleotides and dipeptides from the data. First, the downloaded Fasta file for the genomic data was opened and the ID, the first line, was deleted. Then, the genomic data was read in to R

Studio by the read.table() function. Then, the strings in the data were converted into substrings of two. Then, the new data had a lot of letter pairs of dinucleotides. The letter pairs were extracted according to the pairs and tabulated into a list, that also kept track of the number of pairs of same type. After this was done for every pair, the numbers of dinucleotides were revealed in the list by the concatenation of a control vector which had the general strings for the 16 dinucleotides. Now, we had the list with each key as the name of the dinucleotide and each value as the frequency of occurrence of that particular dinucleotide in our initial data. Then, the list was stored into a matrix by using the unlist() function. That matrix was saved as a csv (Excel) file for further retrieval. The csv file contained the names of the dinucleotides, the frequencies and the percentages of the frequencies in three different columns.

The same thing was done for the dipeptides, except, that the dipeptide file was a little harder to deal with because it had a lot of IDs in the Fasta file, which had to be removed through coding. From there everything was carried out in the same way except that the general vector that, in the case of dinucleotides, had the 16 general dinucleotide strings, this time, had 400 general dipeptide strings that were to be retrieved from the data.

The data was again saved as a csv (Excel) file, containing the names of the dinucleotides, the frequencies and the percentages of the frequencies in three different columns.

Once both the dinucleotide and dipeptide data for *T*. Maritima MSB8 were made, there were two approaches to follow. The first approach was to get the expected values for the Dinucleotide and Dipeptide frequencies for *T*. Maritima MSB8, by calculation of the original genome data and proteome data. The second approach allowed us to compare the genome and proteome of *T*. Maritima MSB8 with 4 other closely related species found on the species phylogenetic tree, which were not extremophiles.

Approach 1

First, the expected dinucleotide frequencies were to be calculated for *T*. Maritima MSB8. This was done by coding in R Studio, where we parsed through the HEP Genome file of the

bacterium and found the given values for the total number of base pairs and the total number of separate bases A, G, C and T. Once we had the separate correct numbers for A, G, C and T, we got the proportions of A, G, C and T by dividing each base's quantity by the total: for example, if A=501834 and total=1860725, then the proportion of A=501834/1860725=0.27. Then, in order to find the expected dinucleotide, the proportions of dinucleotides were calculated through the following logic: if A=0.27 and G=0.33, then, AG=GA=0.27x0.33=0.09. Then, the proportional value for the dinucleotide was multiplied by the total to get the expected quantity of that dinucleotide: Quantity of AG=0.09x1860725=167465. (The values have been rounded off for simplification). This calculation was repeated for all the 16 dinucleotides and the data was saved in a csv file.

For comparison between the expected and observed dinucleotides, the previously stored expected and observed dinucleotide frequencies were imported from the csv files into two vectors. Then, these vectors were plotted and compared, with frequency on y-axis and the names of dinucleotides on the x-axis. The plots were used to find any outliers. Then, a proportionality test was done on the values which seemed as outliers in the plot, in order to see if the expected and observed values had a significant difference. If the p-value was very high, then the two values had a significant difference, implying that those specific expected and observed dinucleotides greatly vary.

For the comparison of dipeptides, a similar thing was done, but it was more challenging because we did not have access to the correct quantities of peptides in the proteome. So, in R studio, the proteome data was parsed to provide the quantities (hence proportions) of each of the 20 peptides that compose the proteome. The parsing was done in a similar way to the expected dinucleotide parsing but required other coding algorithms. The expected proportions of dipeptides were found by using the same logic as that for expected dinucleotides. The expected dipeptides were saved in a csv file.

For comparison between the expected and observed dipeptides, the previously stored expected and observed dipeptide frequencies were imported from the csv files into two vectors. Then, these vectors were plotted and compared, with frequency on y-axis and the names of dipeptides on the x-axis. The plots were used to find any outliers. The plots

were used to find any outliers. Then, a proportionality test was done on the values which seemed as outliers in the plot, in order to see if the expected and observed values had a significant difference. If the p-value was very high, then the two values had a significant difference, implying that those specific expected and observed dinucleotides greatly vary.

Approach 2

Four closely related species that were not extremophiles, were taken and their genomes and proteomes were parsed to obtain their dinucleotide and dipeptide frequencies. The data was saved in csv files. Then, the dinucleotides of all the five species were plotted against each other, with frequencies of y-axis and the dipeptide names on the x-axis. Then, our personally chosen thresholds were applied to the plots, with normalization for the bid dipeptide plots. We gradually simplified the huge plots into smaller and more specific plots, by using the normalization and thresholds methods and identifying the outliers. Then, we just plotted the graphs with the outliers. Consequently, the plots were read and trends and patterns were picked up in order to decide that which dinucleotides and dipeptides were used in different quantities between our species of focus and the rest of the four species.

Results

For the sake of keeping it precise, only the important results have been included here. All the other plots, data and codes can be found as supplements attached to the report.

The results of our expected and observed dinucleotides are shown below in Fig.1 and Fig.2.

Figure 1: Expected vs Observed Comparison (Red=Observed, Blue=Expected)

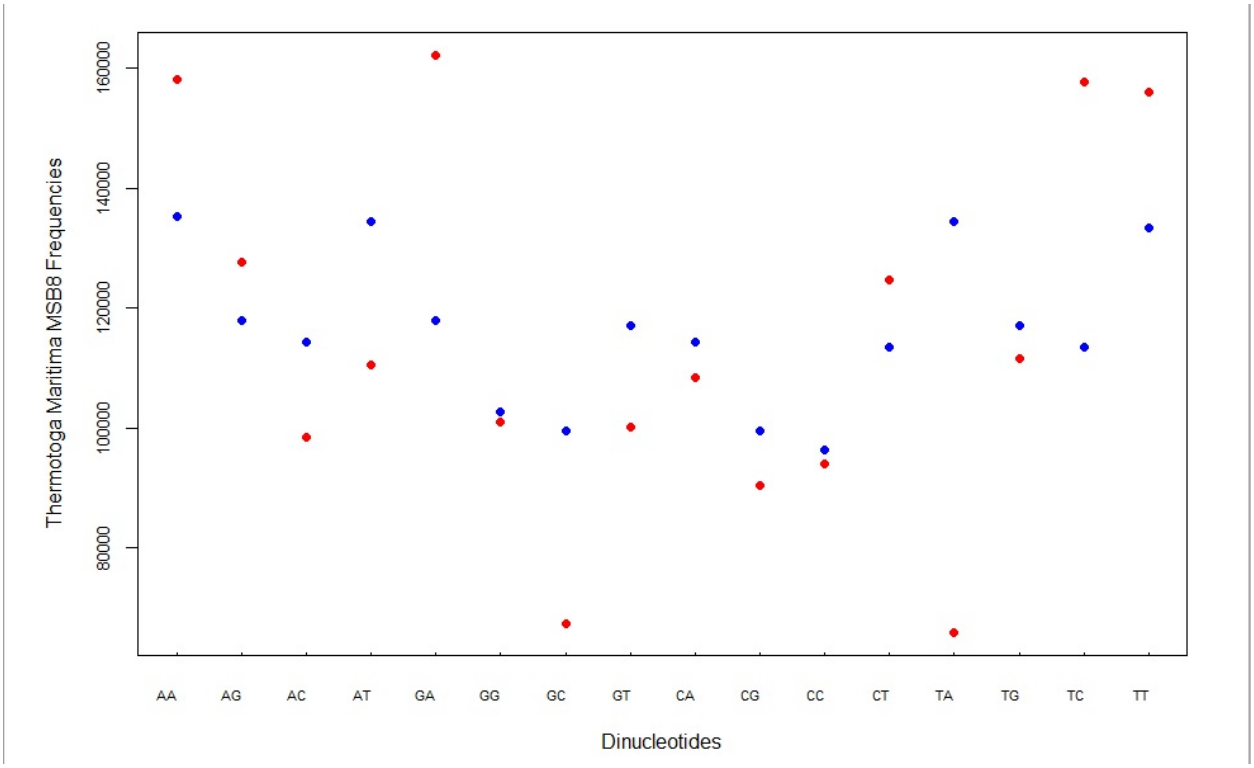
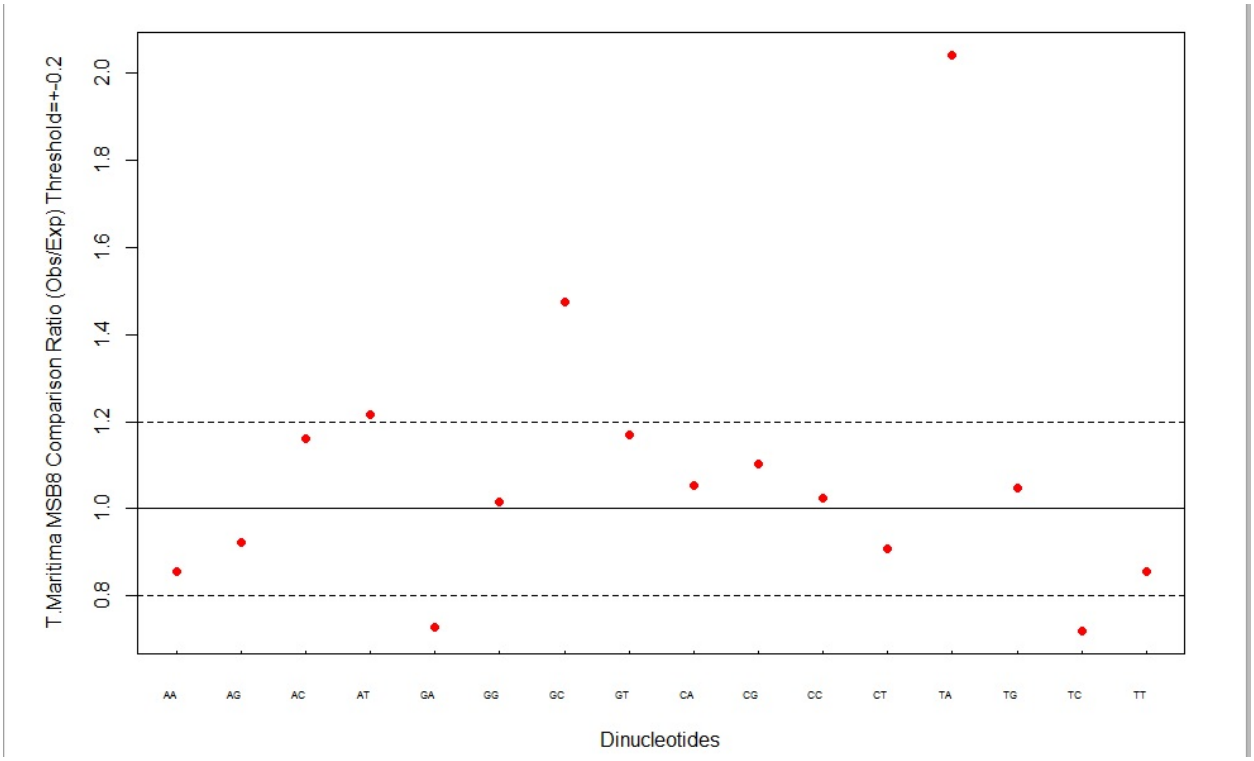


Figure 2: Expected/Observed Ratio (Red=Observed, Blue=Expected)



The Fig. 1 shows the plots of the expected and observed dinucleotides together at the x-axis, with observed shown as Red and expected shown as Blue plots. The y-axis shows the frequencies. We can see that at AA, GA, GC, TA, TC and TT, we can see outliers for the observed points. To make the differences more understandable, we plotted the ratio of expected over observed (Fig. 2). We also set a threshold of 1+-0.2. We saw that there were 5 outliers at AT, GA, GC, TA and TC. So, we took these 5 outliers and did proportionality tests on them. Fig. 3.1 and 3.2 show the results of the proportionality tests.

Figure 3.1: Proportionality Tests

```
[1] "AT"

        2-sample test for equality of proportions with continuity correction

data:  c(va[i], vb[i]) out of c((total - 1), (total - 1))
X-squared = 2511.9, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.01338649 -0.01237823
sample estimates:
    prop 1     prop 2
0.05934948 0.07223184

[1] "GA"

        2-sample test for equality of proportions with continuity correction

data:  c(va[i], vb[i]) out of c((total - 1), (total - 1))
X-squared = 7585.9, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.02328355 0.02435554
sample estimates:
    prop 1     prop 2
0.08715264 0.06333309

[1] "GC"

        2-sample test for equality of proportions with continuity correction

data:  c(va[i], vb[i]) out of c((total - 1), (total - 1))
X-squared = 6455, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.01765936 -0.01681797
sample estimates:
    prop 1     prop 2
0.03621977 0.05345844
```

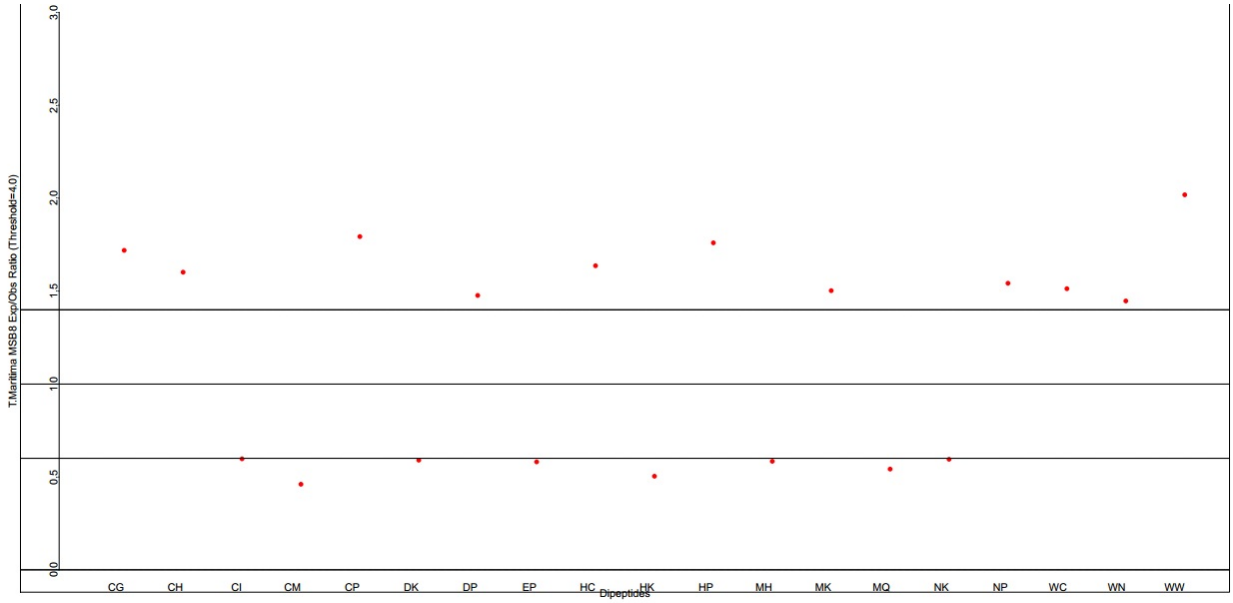
Figure 3.2: Proportionality Tests

```
[1] "TA"

        2-sample test for equality of proportions with continuity correction

data:  c(va[i], vb[i]) out of c((total - 1), (total - 1))
X-squared = 24794, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.03729453 -0.03637949
sample estimates:
    prop 1     prop 2
0.03539482 0.07223184

[1] "TC"

        2-sample test for equality of proportions with continuity correction

data:  c(va[i], vb[i]) out of c((total - 1), (total - 1))
X-squared = 7780.5, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.02323963 0.02429581
sample estimates:
    prop 1     prop 2
0.08473745 0.06096973
```

In fig. 3.1 and 3.2, we saw that the p-value for each outlier was very small. This concluded that there was not a significant difference between the observed and expected values of dinucleotides for these outliers, hence justifying that there was not a significant difference between the two data for *T*. Maritima MSB8.

Moving onto the dipeptides, the results of the expected and observed dipeptides are shown in the figure 4 below.

Figure 4: Extracted Outliers Plot



This figure is the result of a direct bypass of the dipeptide plotting and picking the obvious outliers in order to have a small range of dipeptides for better analysis. The x-axis shows the dipeptide names and the y-axis shows the ratio of expected and observed dipeptide values for *T*. Maritima MSB8. The threshold value is set to 1+-0.4. Based on this plot, we can see a graphical representation of the ratios of expected over observe. We further statistically enquired if these dipeptides had significant differences between the observed and expected values by doing proportional tests on this data. The following figures show the results of the proportional tests.

Figure 5.1

```
[1] "AT"

        2-sample test for equality of proportions with continuity correction

data:  c(va[i], vb[i]) out of c((total - 1), (total - 1))
X-squared = 2511.9, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.01338649 -0.01237823
sample estimates:
    prop 1     prop 2
0.05934948 0.07223184

[1] "GA"

        2-sample test for equality of proportions with continuity correction

data:  c(va[i], vb[i]) out of c((total - 1), (total - 1))
X-squared = 7585.9, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.02328355 0.02435554
sample estimates:
    prop 1     prop 2
0.08715264 0.06333309

[1] "GC"

        2-sample test for equality of proportions with continuity correction

data:  c(va[i], vb[i]) out of c((total - 1), (total - 1))
X-squared = 6455, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.01765936 -0.01681797
sample estimates:
    prop 1     prop 2
0.03621977 0.05345844
```

Figure 5.2

```
[1] "TA"

        2-sample test for equality of proportions with continuity correction

data:  c(va[i], vb[i]) out of c((total - 1), (total - 1))
X-squared = 24794, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.03729453 -0.03637949
sample estimates:
    prop 1     prop 2
0.03539482 0.07223184

[1] "TC"

        2-sample test for equality of proportions with continuity correction

data:  c(va[i], vb[i]) out of c((total - 1), (total - 1))
X-squared = 7780.5, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.02323963 0.02429581
sample estimates:
    prop 1     prop 2
0.08473745 0.06096973
```

Figure 5.3

```
[1] "CG"

        2-sample test for equality of proportions with continuity correction

data:  c(va1[i], vb1[i]) out of c((total - 1), (total - 1))
X-squared = 22.836, df = 1, p-value = 1.764e-06
alternative hypothesis: two.sided
95 percent confidence interval:
 0.0002028535 0.0004928235
sample estimates:
      prop 1       prop 2
0.0008314681 0.0004836296

[1] "CH"

        2-sample test for equality of proportions with continuity correction

data:  c(va1[i], vb1[i]) out of c((total - 1), (total - 1))
X-squared = 3.5478, df = 1, p-value = 0.05962
alternative hypothesis: two.sided
95 percent confidence interval:
 -2.602198e-06  1.387924e-04
sample estimates:
      prop 1       prop 2
0.0001812679 0.0001131728
```

Figure 5.4

```
[1] "CI"

        2-sample test for equality of proportions with continuity correction

data:  c(va1[i], vb1[i]) out of c((total - 1), (total - 1))
X-squared = 12.064, df = 1, p-value = 0.0005142
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.0003089716 -0.0000838637
sample estimates:
      prop 1       prop 2
0.0002916049 0.0004880226

[1] "CM"

        2-sample test for equality of proportions with continuity correction

data:  c(va1[i], vb1[i]) out of c((total - 1), (total - 1))
X-squared = 7.4864, df = 1, p-value = 0.006217
alternative hypothesis: two.sided
95 percent confidence interval:
 -1.514938e-04 -2.373651e-05
sample estimates:
      prop 1       prop 2
7.487154e-05 1.624867e-04

[1] "CP"

        2-sample test for equality of proportions with continuity correction

data:  c(va1[i], vb1[i]) out of c((total - 1), (total - 1))
X-squared = 15.022, df = 1, p-value = 0.0001062
alternative hypothesis: two.sided
95 percent confidence interval:
 0.0001049071 0.0003274197
sample estimates:
      prop 1       prop 2
0.0004886353 0.0002724719
```

Figure 5.5

```
[1] "DK"

        2-sample test for equality of proportions with continuity correction

data:  c(va1[i], vb1[i]) out of c((total - 1), (total - 1))
X-squared = 108.08, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.001971859 -0.001340959
sample estimates:
     prop 1      prop 2
0.002388008 0.004044417

[1] "DP"

        2-sample test for equality of proportions with continuity correction

data:  c(va1[i], vb1[i]) out of c((total - 1), (total - 1))
X-squared = 47.572, df = 1, p-value = 5.302e-12
alternative hypothesis: two.sided
95 percent confidence interval:
 0.0006983045 0.0012605643
sample estimates:
     prop 1      prop 2
0.003034268 0.002054833
```

Figure 5.6

```
[1] "EP"

        2-sample test for equality of proportions with continuity correction

data:  c(va1[i], vb1[i]) out of c((total - 1), (total - 1))
X-squared = 105.51, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.001876375 -0.001269759
sample estimates:
     prop 1      prop 2
0.002183096 0.003756163

[1] "HC"

        2-sample test for equality of proportions with continuity correction

data:  c(va1[i], vb1[i]) out of c((total - 1), (total - 1))
X-squared = 3.9442, df = 1, p-value = 0.04703
alternative hypothesis: two.sided
95 percent confidence interval:
 8.932728e-07 1.431782e-04
sample estimates:
      prop 1       prop 2
0.0001852085 0.0001131728

[1] "HK"

        2-sample test for equality of proportions with continuity correction

data:  c(va1[i], vb1[i]) out of c((total - 1), (total - 1))
X-squared = 53.919, df = 1, p-value = 2.089e-13
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.0008282814 -0.0004747271
sample estimates:
     prop 1      prop 2
0.000662022 0.001313526
```

Figure 5.7

```
[1] "HP"

        2-sample test for equality of proportions with continuity correction

data:  c(va1[i], vb1[i]) out of c((total - 1), (total - 1))
X-squared = 34.895, df = 1, p-value = 3.479e-09
alternative hypothesis: two.sided
95 percent confidence interval:
 0.0003361159 0.0006777685
sample estimates:
      prop 1       prop 2
0.0011743009 0.0006673587

[1] "MH"

        2-sample test for equality of proportions with continuity correction

data:  c(va1[i], vb1[i]) out of c((total - 1), (total - 1))
X-squared = 10.507, df = 1, p-value = 0.00119
alternative hypothesis: two.sided
95 percent confidence interval:
 -2.670957e-04 -6.386199e-05
sample estimates:
      prop 1       prop 2
0.0002324958 0.0003979746
```

Figure 5.8

```
[1] "MK"

        2-sample test for equality of proportions with continuity correction

data:  c(va1[i], vb1[i]) out of c((total - 1), (total - 1))
X-squared = 47.98, df = 1, p-value = 4.307e-12
alternative hypothesis: two.sided
95 percent confidence interval:
 0.0006765247 0.0012183049
sample estimates:
     prop 1      prop 2
0.002833297 0.001885882

[1] "MQ"

        2-sample test for equality of proportions with continuity correction

data:  c(va1[i], vb1[i]) out of c((total - 1), (total - 1))
X-squared = 15.944, df = 1, p-value = 6.526e-05
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.0003291992 -0.0001097360
sample estimates:
      prop 1       prop 2
0.0002600801 0.0004795477

[1] "NK"

        2-sample test for equality of proportions with continuity correction

data:  c(va1[i], vb1[i]) out of c((total - 1), (total - 1))
X-squared = 72.205, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.0013889791 -0.0008634227
sample estimates:
     prop 1      prop 2
0.001655055 0.002781256
```

Figure 5.9

```
[1] "NP"

        2-sample test for equality of proportions with continuity correction

data:  c(va1[i], vb1[i]) out of c((total - 1), (total - 1))
X-squared = 41.109, df = 1, p-value = 1.44e-10
alternative hypothesis: two.sided
95 percent confidence interval:
 0.0005291806 0.0010030049
sample estimates:
     prop 1      prop 2
0.002179156 0.001413063

[1] "WC"

        2-sample test for equality of proportions with continuity correction

data:  c(va1[i], vb1[i]) out of c((total - 1), (total - 1))
X-squared = 1.3458, df = 1, p-value = 0.246
alternative hypothesis: two.sided
95 percent confidence interval:
 -2.031092e-05  8.710549e-05
sample estimates:
      prop 1       prop 2
9.851518e-05 6.511789e-05

[1] "WN"

        2-sample test for equality of proportions with continuity correction

data:  c(va1[i], vb1[i]) out of c((total - 1), (total - 1))
X-squared = 6.6377, df = 1, p-value = 0.009984
alternative hypothesis: two.sided
95 percent confidence interval:
 3.516828e-05 2.666882e-04
sample estimates:
      prop 1       prop 2
0.0004886353 0.0003377071
```

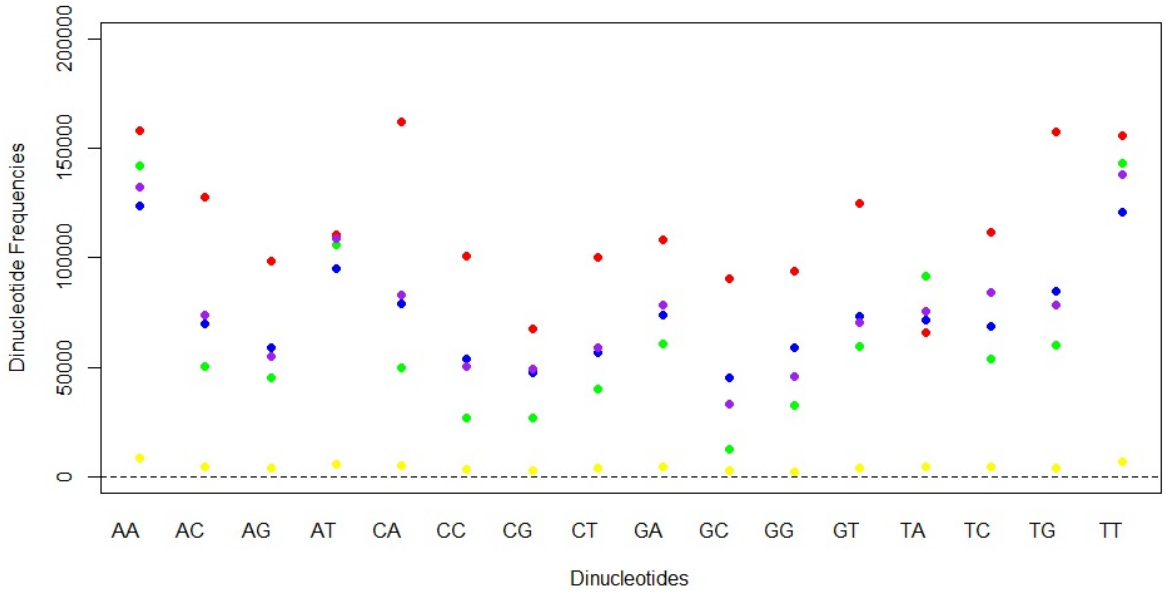
Figure 5.10

```
[1] "WW"

        2-sample test for equality of proportions with continuity correction

data:  c(va1[i], vb1[i]) out of c((total - 1), (total - 1))
X-squared = 7.3399, df = 1, p-value = 0.006744
alternative hypothesis: two.sided
95 percent confidence interval:
 2.475206e-05 1.621269e-04
sample estimates:
      prop 1       prop 2
1.852085e-04 9.176904e-05
```

All the Figures 5.1-10 show very small p-values for the outliers, hence they statistically prove that these outliers are not significantly different than their corresponding pairs in the observed proteome sequence. So, this statistically concludes that there is no significant difference between the observed and expected dipeptides for *T*. Maritima MSB8.

For the second approach, we plotted the dinucleotide frequencies of *T*. Maritima MSB8 with four other closely related non-extremophile species: Kosmotoga Olearia TBF, Thermosipho Africanus TCF52B, Pseudothermotoga Lettingae TMO and Fervidobacterium Pennivorans DSM. The results of the plot are as follows.

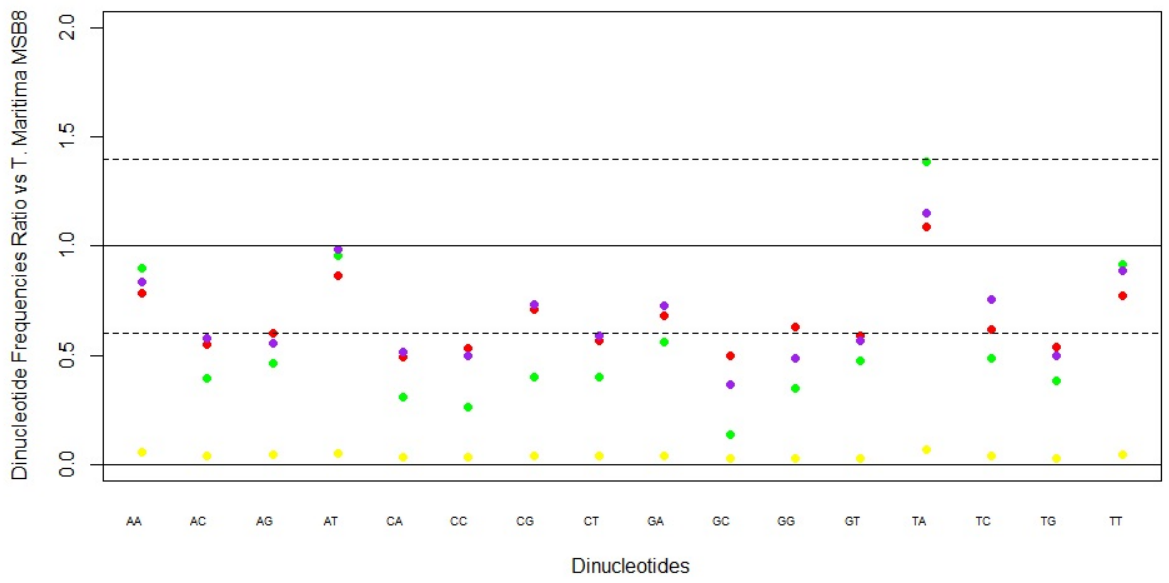
Figure 6



The Fig. 6 shows the plots of the dinucleotides of the 5 species together at the x-axis, with each species shown with a different color. The y-axis shows the dinucleotide frequencies. In the plot, red=*T*. Maritima MSB8, blue=*K*. Olearia, green=*T*. Africanus, purple=*P*. Lettingae and yellow=F. Pennivorans. From the plot we can see that apart from the plot of TA, *T*. Maritima MSB8 has the most dinucleotides as compared to all the plots. So, we might have a smaller hypothesis that maybe loss of TA might be the factor for the extremophilic characteristics of *T*. Maritima MSB8 because none of the other 4 species have lesser TA

than *T*. Maritima MSB8. To make a better comparison and analysis, we plotted the ratios of the 4 species against *T*. Maritima MSB8. The results are shown in Fig. 7 below.

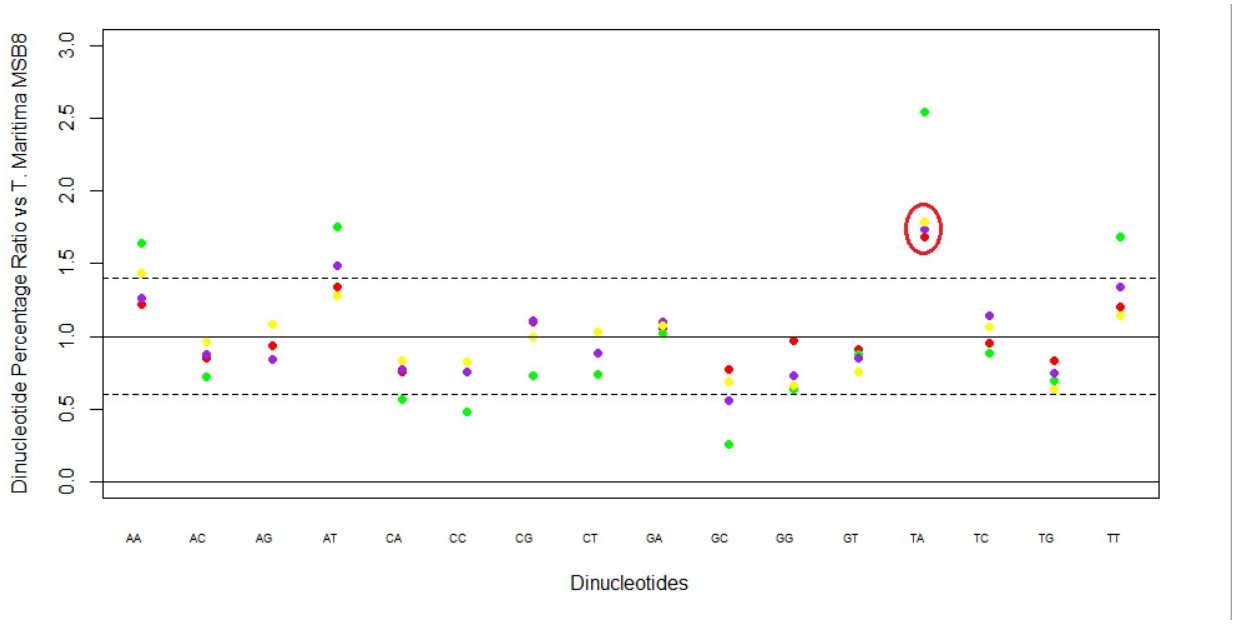
Figure 7



In Fig. 7, we can see that the threshold is set to 1+-0.4. and we can see a lot of plots below the 0.6 mark. But, still our main focus was to identify those points which were common for all the 4 species but not for our species in focus. Such as the points TA and TG. So, based on this, we might conclude that TG and TA may have something to do with the non-extremophilic characteristics of a species. But, as seen in both the Figures 6 and 7, the plots of F. Pennivorans (yellow) are very low, primarily because the genome of F. Pennivorans is relatively minute as compared to the rest of the species. This introduces a huge bias in our plots because F. Pennivorans has contributive characteristics that may affect an organism's ability to be non-extremophile.

In order to get rid of this bias, we got percentage ratio plots for the dinucleotides of each species. In this way, even if F. Pennivorans dinucleotides is relatively very less in quantity, their percentage may be very high, relative to their genome. So, by percentage analysis, we can see that how much does the percentage dinucleotide quantity affect the genome of the particular species. The results are shown in Fig. 8.

Figure 8



In Fig. 8, we see that that now F. Pennivorans is relatively very much more contributive to the overall analysis between the species. In the plots, we can see that TA is a very strong candidate for being a contributor for non-extremophilic characteristics because it is the most common between the 4 non-extremophilic species as well as being outside the threshold, hence, ending as an outlier.

Moving onto the dipeptide comparison, we made similar plots for all the 5 species. The results are shown below in Fig. 9. It should be noted that we have bypassed the overly time-consuming plots that contain all the 400 dipeptides for all the 5 species, and have narrowed down the most important aspects in the following figures. (All the other plots can be seen in the supplemental attached with the report).

Figure 9: Dipeptide Frequency Ratios vs T. Maritima, Threshold=4

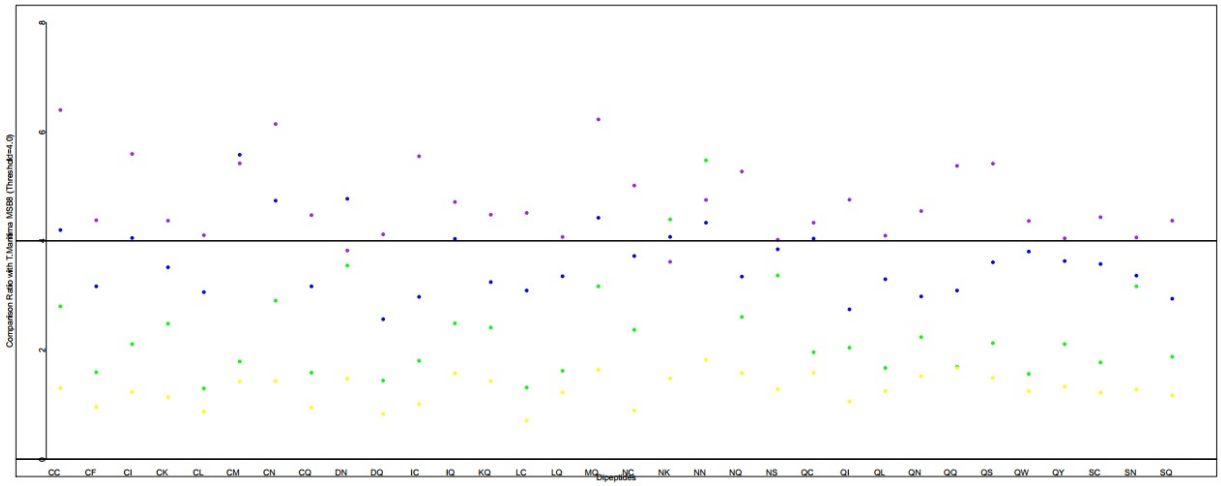
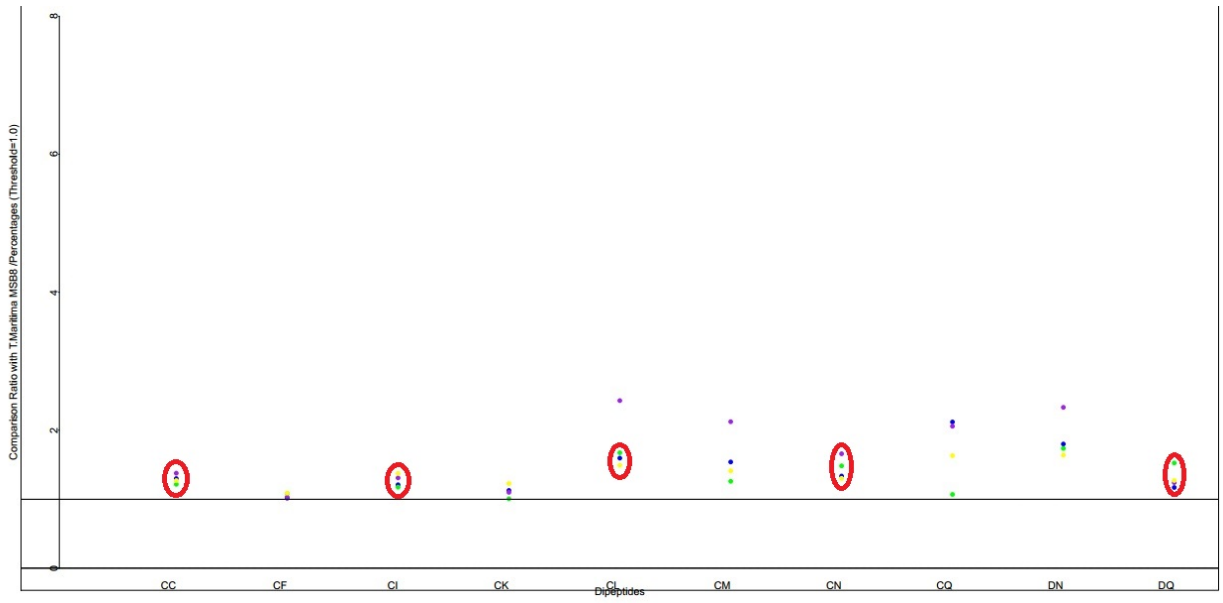


Fig. 9 shows the outliers already identified in initial comparisons of 400 dipeptides (shown in supplemental). We can see a lot of dipeptides above the threshold (y=4), but no specific relation or trend can be seen in this scatterplot. Also, once again we have the same bias caused by the low quantity of F. Pennivorans dipeptides. So, in order to have a more contribution of F. Pennivorans and to show a better analyzable data, we decided to plot the frequency percentages of the dipeptides for the 5 species. The result is shown in Fig. 10 below.



In Fig. 10, we can see a very good contribution of F. Pennivorans in the plots, and also, we can see a very analyzable data here. As the plots show here, there are 5 very common

dipeptides found in the non-extremophilic species in approximately similar quantities. Moreover, they are also above the threshold (y=1) set by us in the plots. The 5 dipeptides CC, CI, CL, CN and DQ are very strong candidates that (if not with alone expression, then combined expression) might be responsible for non-extremophilic characteristics among the four species.

Discussion

Our results decisively show a very good relationship of some particular dinucleotides and dipeptides that are generally in more quantities in non-extremophiles, as compared to extremophiles such as the bacteria species Thermotoga Maritima MSB8. We picked the 4 control species based on their closeness to *T*. Maritima in the phylogenetic tree, so that means that they have very good evolutionary relationships with common phylogenies among each other that can make our analysis very easy to form and compare between these species. From our analysis, we conclude that there indeed is a dinucleotide and dipeptide bias in our species in focus, and hence, generally, extremophiles. But, what we have seen in our analysis is a decreasing bias instead of an increasing bias. Apart from the idea that our bacteria species had the largest quantity of dinucleotides instead TA, might also suggest that maybe extremophilic characteristics are not a result of a gain in function of a protein, but a loss in function of a protein. So far, our analysis could only detect dinucleotides and dipeptides that were less in *T*. Maritima, but it begs the question that maybe, a different way of analysis may discover some dinucleotides and dipeptides that are found in relatively more percentages in *T*. Maritima as compared to the non-extremophilic species. An easy way to achieve that might be to inverse the ratios for the plots and make new plots, to identify the dinucleotides and dipeptides that are more than the threshold (y=1), meaning that these dinucleotides and dipeptides are to be found in more quantity (percentage) in *T*. Maritima, as compared to the rest of the 4 non-extremophilic species. From the expected vs observed dinucleotide and dipeptide analysis, we can also see that *T*. Maritima's genome is supposed to be in this way because the dinucleotides that we

observed were not significantly different than the expected values of those of *T*. Maritima. This also stands to prove that our basic data had a minimal error due to minimal (insignificant difference between the observed and expected), leading to the conclusion that our analysis was not significantly erroneous, even though there might be some random or systematic errors not taken into consideration here.

The analysis we did in this project can further be polished by more statistical tests and other bioinformatics approaches such as ancestral state estimation where, based on the common phylogenies among species, their common ancestor can be found. This can help in a way that we can analyze the common ancestor and siblings, and based on their common phylogenies and evolutionary relationships, we can make an estimated model for our species in focus. Then, we can compare the estimated model with our original species to see that how far is the species from the estimate. Probably, any differences found may contribute to the uncommon traits between the siblings (neighbors), which is a very good idea when we have neighbors with different characteristics, such as extremophilicity in our case. Apart from this, since Bioinformatics is a very largely growing field, we can expect to have many better ways of interspecies characteristics and anomaly analysis with a wide and better variety of tools, that will not only make the tasks simpler but also less time consuming and more effective.