Department of Electronic & Telecommunication Engineering University of Moratuwa



BM4321 – Genomic Signal Processing

**Assignment 1**

**Promoter Discovery in Bacteria**

Palihakkara A.T. 170418E

This report is submitted in partial fulfillment of the requirements for the module BM4321 Genomic Signal Processing.

17 October 2021

**Primary Data**

Organism Name : Salmonella enterica

Accession : NZ\_CP075108

Version : NZ\_CP075108.1

**Question 1**

Table 1: Number of Valid gene distribution

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **From Protein Table** | **Valid Genes** | **Proportion of selected genes** | **Proportion of not selected genes** |
| **Sense Strand** | 2188 | 1299 | 59.37 % | 40.63 % |
| **Anti–Sense Strand** | 2089 | 1251 | 59.89 % | 40.11 % |
| **Total** | 4277 | 2550 | 59.62 % | 40.38 % |

Valid genes are the genes which are filtered using the 50 bases upstream, threshold and the Methionine start codon check.

According to Table 1, we can see that nearly 40% of the genes from the protein table are not considered as valid genes.

Table 2: Bases distribution of the valid genes

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **A** | | **C** | | **T** | | **G** | |
| **count** | **%** | **count** | **%** | **count** | **%** | **count** | **%** |
| **Sense Strand** | 288007 | 23.46 | 310962 | 25.33 | 341286 | 27.79 | 287495 | 23.42 |
| **Anti–Sense Strand** | 271790 | 23.39 | 291070 | 25.06 | 325307 | 28.00 | 273438 | 23.54 |
| **Total** | 559797 | 23.43 | 602032 | 25.19 | 666593 | 27.89 | 560933 | 23.48 |

According to the Table 2, we can see that there are almost equal occurrences of bases in the valid genes.

Table 3: Total number of Bases in the sense strand of the DNA

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A** | **C** | **T** | **G** | Total Base pairs in the DNA |
| **Count** | 1105779 | 1217852 | 1215258 | 1110015 | 4648904 |

**Question 2**

First 1000 genes of the valid genes from the sense strand are taken as the training set data to obtain the Position Probability Matrix (PPM). From those 1000 genes, the upstream bases of 30 to 5 (length of 25) are extracted as the possible sequence of locating the promotor. Then those sequences are aligned with W matching (intact query) with the Pribnow Box (‘TATAAT’) and obtain the aligning start position of the sequence. Since we take a 10-position long PPM, 10 bases from the aligned position were taken as the promotor sequence. Aligning positions where we can’t take 10 bases from the 25 bases long sequence, those aligning positions are neglected. The obtained PPM is shown in the Table 4.

Table 4: 10-Position Probability Matrix with 1000 Sense Strand Genes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** |
| **A** | 0.5011199 | 0.4321185 | 0.435512 | 0.4185444 | 0.3721664 |
| **C** | 0.0022737 | 0.0079295 | 0.0452581 | 0.0791932 | 0.157244 |
| **G** | 0.0000113 | 0.0000113 | 0.0339464 | 0.0803244 | 0.1131284 |
| **T** | 0.4965952 | 0.5599407 | 0.4852835 | 0.4219379 | 0.3574612 |
|  | | | | | |
|  | **6** | **7** | **8** | **9** | **10** |
| **A** | 0.3303131 | 0.3178702 | 0.3291819 | 0.3280508 | 0.2941157 |
| **C** | 0.1923103 | 0.2409506 | 0.2149337 | 0.190048 | 0.2171961 |
| **G** | 0.1628999 | 0.2058843 | 0.2330324 | 0.2590494 | 0.2873286 |
| **T** | 0.3144767 | 0.2352948 | 0.2228519 | 0.2228519 | 0.2013597 |

By taking the maximum probabilities of each column we can obtain the initial sequence which has the maximum consensus score. The obtained **Consensus Sequence** is **“ATTTAAAAAA”**

**Question 3**

Chart

Description automatically generated

Figure 1: Entropy vs PPM Position | n = 1000

Entropy values were calculated along each column and plotted it with respect to the position. According to the Figure 1, we can see that the positions after 6 are having very low entropy values. It happens because in those positions all A, C, T, G bases are likely to be occurred in equiprobable manner.

By looking into the PPM and the Entropy values, I selected **0.04** as the **Entropy Threshold** for the reduced PPM.

**Question 4**

Test data set included the remaining 299 genes from the sense strand and all 1251 genes from the anti – sense strand. So, the Test data set had 1550 genes for the test.

At first the Benchmark Consensus Score and the Reduced Benchmark Consensus Score for the Consensus Sequence obtained by the PPM, were calculated and the obtained values are in Table 5.

Table 5: Benchmark Consensus Scores | n = 1000

|  |  |
| --- | --- |
| **Consensus Score for the 10-PPM** | -9.54851 |
| **Consensus Score for the reduced PPM** | -4.95288 |

When obtaining the sequences from the sense strand gene, base sequence of 5 to 30 upstream from the start location of the respective gene is selected. And when selecting base sequence from the anti – sense strand, base sequence of 5 to 30 downstream from the end location of the respective gene is selected and the sequence is reversed when aligning with the PPM in the promotor search.

After selecting the sequences they were aligned with the PPM and the reduced PPM. Then the respective scores are compared with the benchmark values in Table 5. And classified using the thresholds which were changed from -1 to -5. The number of classified valid genes for each threshold can be seen in Table 6, Table 7 & Table 8.

Table 6: Selected percentages of genes from the Sense Strand | n = 1000

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Threshold** | **Sense Strand Genes (299)** | | | |
| **With Initial PPM** | | **With Reduced PPM** | |
| **Valid Count** | **Valid %** | **Valid Count** | **Valid %** |
| **-1** | 59 | 19.73 | 182 | 60.87 |
| **-2** | 179 | 59.87 | 232 | 77.59 |
| **-3** | 237 | 79.26 | 267 | 89.30 |
| **-4** | 264 | 88.29 | 283 | 94.65 |
| **-5** | 282 | 94.31 | 290 | 96.99 |

Table 7: Selected percentages of genes from the Anti- Sense Strand | n = 1000

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Threshold** | **Anti - Sense Strand Genes(1251)** | | | |
| **With Initial PPM** | | **With Reduced PPM** | |
| **Valid Count** | **Valid %** | **Valid Count** | **Valid %** |
| **-1** | 261 | 20.86 | 694 | 55.48 |
| **-2** | 701 | 56.04 | 955 | 76.34 |
| **-3** | 942 | 75.30 | 1110 | 88.73 |
| **-4** | 1099 | 87.85 | 1185 | 94.72 |
| **-5** | 1176 | 94.0 | 1217 | 97.28 |

Table 8: Total Selected & Not Selected percentages of genes | n = 1000

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Threshold** | **Total Test set Genes(1550)** | | | | | |
| **With Initial PPM** | | | **With Reduced PPM** | | |
| **Valid Count** | **Valid %** | **Not Valid %** | **Valid Count** | **Valid %** | **Not Valid %** |
| **-1** | 320 | 20.65 | 79.35 | 876 | 56.72 | 43.48 |
| **-2** | 880 | 56.77 | 43.23 | 1187 | 76.58 | 23.42 |
| **-3** | 1179 | 76.06 | 23.94 | 1377 | 88.84 | 11.16 |
| **-4** | 1363 | 87.94 | 12.06 | 1468 | 94.71 | 5.29 |
| **-5** | 1458 | 94.06 | 5.94 | 1507 | 97.23 | 2.77 |

By looking into above Table 6, Table 7 and Table 8, we can see that the selected number of genes count increases with changing the threshold from -1 to -5. When taking the upper threshold and the lower threshold values, the genes which are not having valid promotors are very low with the -5 threshold compared to -1 threshold, according to Table 8.

**Question 5**

**PPM using random 10 sequences.**

Table 9: 10-Position Probability Matrix with 10 Sense Strand Genes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** |
| **A** | 0.2998008 | 0.5986056 | 0.499004 | 0.2998008 | 0.3994024 |
| **C** | 0.000996 | 0.000996 | 0.1005976 | 0.1005976 | 0.1005976 |
| **G** | 0.000996 | 0.000996 | 0.000996 | 0.1005976 | 0.3994024 |
| **T** | 0.6982072 | 0.3994024 | 0.3994024 | 0.499004 | 0.1005976 |
|  | **6** | **7** | **8** | **9** | **10** |
| **A** | 0.5986056 | 0.1005976 | 0.2001992 | 0.1005976 | 0.3994024 |
| **C** | 0.1005976 | 0.2998008 | 0.2998008 | 0.2998008 | 0.2001992 |
| **G** | 0.1005976 | 0.2998008 | 0.2998008 | 0.2998008 | 0.3994024 |
| **T** | 0.2001992 | 0.2998008 | 0.2001992 | 0.2998008 | 0.000996 |

Chart, line chart

Description automatically generated

Figure 2: Entropy vs PPM Position | n = 10

Table 10: Benchmark Consensus Scores | n = 10

|  |  |
| --- | --- |
| **Consensus Sequence** | TAATGATGTG |
| **Entropy Threshold** | 0.2 |
| **Consensus Score for the 10-PPM** | -8.22531 |
| **Consensus Score for the reduced PPM** | -10.60536 |

Table 11: Total Selected & Not Selected percentages of genes | n = 10

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Threshold** | **Total Test set Genes (1550)** | | | | | |
| **With Initial PPM** | | | **With Reduced PPM** | | |
| **Valid Count** | **Valid %** | **Not Valid %** | **Valid Count** | **Valid %** | **Not Valid %** |
| **-1** | 76 | 4.90 | 95.1 | 1471 | 94.90 | 5.10 |
| **-2** | 291 | 18.77 | 81.23 | 1483 | 95.68 | 4.32 |
| **-3** | 691 | 44.58 | 55.42 | 1520 | 98.06 | 1.94 |
| **-4** | 1040 | 67.10 | 32.90 | 1540 | 99.35 | 0.65 |
| **-5** | 1286 | 82.97 | 17.03 | 1545 | 99.68 | 0.32 |

To obtain the PPM, a random 10 samples beginning from the 611th index of the selected valid genes were selected.

Form Table 10, we can see that the reduced PPM has got a higher consensus score compared to the initial PPM. This can happen because with 10 samples, the probabilities are not very reliable. By reducing the columns with the threshold, the consensus score can become higher than the initial consensus score. And Looking into the Table 11, we can say that the more than 94% of samples are classified as having valid promotors with the reduced PPM regardless of the threshold.

**PPM using random 100 sequences.**

Table 12: 10-Position Probability Matrix with 100 Sense Strand Genes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** |
| **A** | 0.5109951 | 0.3555087 | 0.377721 | 0.3999334 | 0.377721 |
| **C** | 0.0001111 | 0.0223234 | 0.0667481 | 0.0889605 | 0.1667037 |
| **G** | 0.0001111 | 0.0001111 | 0.0334296 | 0.1000666 | 0.1000666 |
| **T** | 0.4887828 | 0.6220569 | 0.5221013 | 0.4110395 | 0.3555087 |
|  | **6** | **7** | **8** | **9** | **10** |
| **A** | 0.3888272 | 0.3221901 | 0.3444025 | 0.3444025 | 0.3332963 |
| **C** | 0.1555975 | 0.2333407 | 0.2333407 | 0.1555975 | 0.1778099 |
| **G** | 0.1444913 | 0.2000222 | 0.1444913 | 0.2777654 | 0.3221901 |
| **T** | 0.311084 | 0.2444469 | 0.2777654 | 0.2222346 | 0.1667037 |

Chart

Description automatically generated

Figure 3: Entropy vs PPM Position | n = 100

Table 12: Benchmark Consensus Scores | n = 100

|  |  |
| --- | --- |
| **Consensus Sequence** | ATTTAAAAAA |
| **Entropy Threshold** | 0.08 |
| **Consensus Score for the 10-PPM** | -8.96652 |
| **Consensus Score for the reduced PPM** | -4.60330 |

Table 13: Total Selected & Not Selected percentages of genes | n = 100

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Threshold** | **Total Test set Genes (1550)** | | | | | |
| **With Initial PPM** | | | **With Reduced PPM** | | |
| **Valid Count** | **Valid %** | **Not Valid %** | **Valid Count** | **Valid %** | **Not Valid %** |
| **-1** | 138 | 8.90 | 91.10 | 600 | 38.71 | 61.29 |
| **-2** | 548 | 35.35 | 64.65 | 1073 | 69.23 | 30.77 |
| **-3** | 989 | 63.81 | 36.19 | 1315 | 84.84 | 15.16 |
| **-4** | 1244 | 80.26 | 19.74 | 1455 | 93.87 | 6.13 |
| **-5** | 1403 | 90.52 | 9.48 | 1502 | 96.90 | 3.10 |

By comparing with the 10 samples PPM classification, in Table 13 we can see that the samples are not classified with higher percentage with lower thresholds. Also, the initial consensus sequence includes only A & T bases while the 10 sample PPM initial consensus sequence includes G base apart from the A & T bases.

**Question 6**

Here all the percentages are calculated with respect to the number of valid genes in the respective DNA.

Valid genes are the genes which are filtered using the 50 bases upstream, threshold and the Methionine start codon check.

Detectable promotor check was done with the reduced PPM used in for the NZ\_CP075108.1 file.

Table 14: Used parameters for the test

|  |  |
| --- | --- |
| **Consensus Sequence** | ATTTAAAAAA |
| **Entropy Threshold** | 0.04 |
| **Consensus Score for the reduced PPM** | -4.95288 |

Table 15: Proportion of genes with detectable promotors w.r.t. threshold

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Accession** |  |  | **Thresholds** | | | | |
| **Valid** | **-1** | **-2** | **-3** | **-4** | **-5** |
| **NZ\_CP066047.1** | **Count** | 2597 | 1532 | 2036 | 2358 | 2504 | 2551 |
| **%** | - | 58.99 | 78.40 | 90.80 | 96.42 | 98.23 |
| **NZ\_CP028172.1** | **Count** | 2541 | 1480 | 1988 | 2306 | 2455 | 2494 |
| **%** | - | 58.24 | 78.24 | 90.75 | 96.62 | 98.15 |
| **NZ\_CP030194.1** | **Count** | 2459 | 1439 | 1929 | 2231 | 2366 | 2415 |
| **%** | - | 58.52 | 78.45 | 90.73 | 96.22 | 98.21 |
| **NZ\_CP030231.1** | **Count** | 2514 | 1472 | 1978 | 2292 | 2433 | 2472 |
| **%** | - | 58.55 | 78.68 | 91.17 | 96.78 | 98.33 |
| **NZ\_CP030238.1** | **Count** | 2493 | 1471 | 1962 | 2269 | 2401 | 2448 |
| **%** | - | 59.01 | 78.70 | 91.01 | 96.31 | 98.19 |
| **NZ\_CP037891.1** | **Count** | 2537 | 1496 | 1992 | 2307 | 2452 | 2495 |
| **%** | - | 58.97 | 78.52 | 90.93 | 96.65 | 98.34 |
| **NZ\_CP040380.1** | **Count** | 2512 | 1453 | 1961 | 2269 | 2405 | 2465 |
| **%** | - | 57.84 | 78.07 | 90.33 | 95.74 | 98.13 |
| **NZ\_CP046277.1** | **Count** | 2534 | 1483 | 1965 | 2293 | 2431 | 2487 |
| **%** | - | 58.52 | 77.55 | 90.49 | 95.94 | 98.15 |
| **NZ\_CP046279.1** | **Count** | 2516 | 1468 | 1979 | 2287 | 2429 | 2469 |
| **%** | - | 58.35 | 78.66 | 90.90 | 96.54 | 98.13 |
| **NZ\_CP046280.1** | **Count** | 2530 | 1477 | 1994 | 2293 | 2443 | 2488 |
| **%** | - | 58.38 | 78.81 | 90.63 | 96.56 | 98.34 |
| **NZ\_CP046291.1** | **Count** | 2546 | 1472 | 1987 | 2299 | 2435 | 2495 |
| **%** | - | 57.82 | 78.04 | 90.30 | 95.64 | 98.00 |
| **NZ\_CP053581.1** | **Count** | 2513 | 1474 | 1960 | 2291 | 2433 | 2476 |
| **%** | - | 58.65 | 77.99 | 91.17 | 96.82 | 98.53 |
| **NZ\_CP060508.1** | **Count** | 2543 | 1487 | 1993 | 2301 | 2432 | 2487 |
| **%** | - | 58.47 | 78.37 | 90.48 | 95.64 | 97.80 |
| **NZ\_CP069518.1** | **Count** | 2572 | 1516 | 2030 | 2342 | 2491 | 2530 |
| **%** | - | 58.94 | 78.93 | 91.06 | 96.85 | 98.37 |

By looking into the Table 15, we can see that nearly 60% promotors were detectable with the -1 threshold, nearly 78% promotors were detectable with the -2 threshold, nearly 90% promotors were detectable with the -3 threshold, nearly 96% promotors were detectable with the -4 threshold and nearly 98% promotors were detectable with the -5 threshold.