UNIVERSITY OF MORATUWA

Faculty of Engineering

Department of Electronic and Telecommunication Engineering

BM4321: Genomic Signal Processing

Assignment 1

Promoter Discovery in Bacteria

K.G. Abeywardena – 160005C

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

June 2020

OVERVIEW OF ASSIGNED GENBANK ACCESSION OF THE BACTERIA

01	**Organism**	Escherichia coli
02	**Family**	Enterobacteriaceae
03	**Genus**	Escherichia
04	**Description**	Typically present at the lower intestines of humans, where it is the dominant facultative anaerobe present, but it is only one minor constituent of the complete intestinal microflora.
05	**Accession**	NZ_AP018808.1
06	**Summary of Genome**	
	Base Pairs	5678205
	Coding Genes/Protein Count	(5478) 5868
	Gene Count	(5990) 6129
	Sense strand Genes	2786
	Antisense strand Genes	2692
	GC content (%)	50.54

QUESTION 1

a) Standard Intact Query Local Search (IQLS) to locate *Pribnow box promotor (TATAAT)* within upstream position from 5 to 30

Results: By accessing the NCIB Genome Website, under *genome assemblies*, the relevant accession number chromosome *NZ_AP018808.1* was located under the column *Replicons* and its corresponding protein table under the column *CDS* and downloaded the full chromosome as a *fasta file* and protein table as a *csv file.*

Based on the *protein table,* sequences having a length of *50 bases upstream* and *3 bases downstream* were extracted from each coding gene on both sense and antisense strand. To make sure the extracted sequences contain 53 bases after the removal of *EOL*, a *safety of 3 bases* was added and then adjusted for 53 bases. The upstream direction had to be adjusted based on from which strand the sequences are extracted.

First the presence of *Methionine* site was checked for each of the obtained sequences by checking whether the last three bases correspond to *ATG.* By performing *Standard IQLS, Pribnow box promoter (TATAAT)* was located within a search region of upstream positions from 5 to 30 in the selected sequences. For intact query, scores were assigned as; *match = 1, mismatch = -1, gap penalty = 2.*

# seq with mutated *Methionine*	633	11.56%
# seq in sense strand with *Pribnow box* located	2112	75.81%
# seq in antisense strand with *Pribnow box* located	2042	75.85%
Total seq with *Pribnow box* located	4154	75.83%
# seq with *Pribnow box* unlocated	691	12.61%

Discussion: Based on the analysis for the presence of *Methionine site*, *633 sequences* were detected with *mutated methionine sites* which were discarded as it corresponded to a lower percentage compared to the available genes. Since *Pribnow Box* can easily mutate *(A -> T or T -> A),* using *exact alignment* in traditional S*tandard IQLS* to locate *Pribnow Box* is not ideal. Instead, *W-search* was used to locate *Pribnow Box* where both *Adenine (A) and Thymine (T)* were considered as a common base *W.* Based on the search, *Pribnow Box* was located in more than 75% sequences within search region.

b) Obtaining *Position Probability Matrix (PPM)* using first 1000 sequences with 10 positions for the Pribnow box.

Results: Out of the selected 4154 sequences first 1000 sequences were used to obtain the *Position Frequency Table* which was then converted to the *PPM* using the below equation. As the *k value,* 0.01 was used.

$$p_{j,N} = \frac{f_{j,N} + k}{\sum_N (f_{j,N} + k)}$$

Where $f_{j,N}$ *– frequency of a base at position j* $(N \in \{A, C, G, T\})$

$p_{j,N}$ *– probability of a base at position j* $(N \in \{A, C, G, T\})$

Table 1: PPM obtained from the first 1000 sequences for Pribnow Box with Entropy measure

Base	**Position**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A	0.554	0.438	0.456	0.423	0.343	0.294	0.273	0.290	0.359	0.305
C	0.001	0.019	0.052	0.096	0.185	0.231	0.222	0.224	0.190	0.180
G	0.001	0.003	0.029	0.102	0.116	0.195	0.265	0.268	0.272	0.328
T	0.444	0.540	0.463	0.379	0.356	0.280	0.240	0.218	0.179	0.187
S	0.988	0.864	0.599	0.284	0.129	0.018	0.005	0.010	0.059	0.052

Discussion: To prevent the *alignment scores* being 0.0, a small constant *k = 0.01* has been added prior to calculating the probabilities for each cell in PPM. Based on the *PPM*, consensus sequence was identified as *ATTATAAAAG* and the consensus score as *-9.77.* It can be seen that *Guanine (G)* has been identified as one of the bases in the consensus sequence that indicates possible mutations of *G or C* is also possible within *Pribnow Box* regions.

QUESTION 02

a) Obtaining the entropy measures for each position and using suitable entropy measure eliminating the redundant position of the obtained PPM

Results: Using the following equation, the entropy values for the 10 positions were calculated using $p_{0,N} = 0.25$ for each base. (as shown in Table 1)

$$I_j = \frac{1}{\ln(2)} \sum_N p_{j,N} \ln\left(\frac{p_{j,N}}{p_{0,N}}\right)$$

I_j *– entropy of the column j*

$p_{j,N}$ *– probability of a base at position j* $(N \in \{A, C, G, T\})$

$p_{0,N}$ – initial *probability of a base at position j* $(N \in \{A, C, G, T\})$

To eliminate the redundant positions of the PPM obtained for the Pribnow Box, *entropy threshold = 0.02* was selected. Based on the column entropies shown in the Table 1, the *columns 6,7,8* are identified as redundant as they contain lesser certainty compared to the threshold selected.

Table 2: Reduced PPM after removing redundant positions

Base	**Position**						
	1	**2**	**3**	**4**	**5**	**6 (9)**	**7 (10)**
A	0.554	0.438	0.456	0.423	0.343	0.359	0.305
C	0.001	0.019	0.052	0.096	0.185	0.190	0.180
G	0.001	0.003	0.029	0.102	0.116	0.272	0.328
T	0.444	0.540	0.463	0.379	0.356	0.179	0.187

Discussion: Based on Table 1 the probabilities of occurrence of one of the four bases at *column 6,7,8* is equi-probable which reduces the entropies as it increases the uncertainty. Further for *column 9,10* the probability distributions are not uniform for the 4 bases and are skewed. Thus the information from *column 9, 10* is more certain compared to *column 6,7,8.* Hence based on the nature of the probability distributions at each position and their entropy values, entropy threshold was set to *0.02* and obtained Table 2. Based on the reduced PPM, the consensus sequence was identified as *ATTATAG* and consensus score was *-6.01.* Again we can see that *Guanine (G)* at the last position is having a considerable probability and an entropy value meaning it is likely to locate mutations of *G or C* for the Pribnow Bow within this genome.

QUESTION 03

a) Statistical alignment of sequences in *test set* with the initial PPM

Results: The remaining 3154 sequences were statistically aligned with the initial PPM shown in Table 1 using a moving window of size of 10. For a given sequence (of 25 bases long), 16 windows were aligned against the Initial PPM and from the resulted array of 16 alignment scores subtracted the consensus score for the initial PPM to obtain the relative scores. If at least one of the 16 relative scores were less than or equal to the threshold for which it was tested, the particular test sequence was considered to have Pribnow Box promoter. The promoter search was done for 5 threshold values (i.e. -1, -2, -3, -4, -5) and the results were as follows.

Table 3: Results of Statistical Alignment for Pribnow Box search using Initial PPM

	Threshold				
	-1	**-2**	**-3**	**-4**	**-5**
# Genes positive for Pribnow Box	998	2179	2757	3004	3092
% (out of 3154)	31.6	69.1	87.4	95.2	98.0

b) Statistical alignment of sequences in *test set* with the reduced PPM

Results: Similar to *part (a),* the 3154 sequences were statistically aligned with the *reduced PPM* shown in Table 2. Since the last column having significant entropy value corresponds to position 10, we still use the same window size of 10. By subtracting consensus score for reduced PPM, 16 relative scores were obtained. If at least one of the 16 relative scores were less than or equal to the threshold for which it was tested, the particular test sequence was considered to have Pribnow Box Promoter.

Table 4: Results of Statistical Alignment for Pribnow Box search using Reduced PPM

	Threshold				
	-1	**-2**	**-3**	**-4**	**-5**
# Genes positive for Pribnow Box	1664	2465	2903	3047	3115
% (out of 3154)	52.8	78.2	92.0	96.6	98.8

c) Comparison between the Pribnow Box search using Initial PPM and Reduced PPM

		-1	**-2**	**-3**	**-4**	**-5**
With Initial PPM	**# Genes *Negative* for *Pribnow Box***	2156	975	397	150	62
	% (out of 3154)	68.4%	30.9%	12.6%	4.8%	2.0%
With Reduced PPM	**# Genes *Negative* for *Pribnow Box***	1490	689	251	107	39
	% (out of 3154)	47.2%	21.8%	8.0%	3.4%	1.2%

Discussion: By comparing the success rates for statistical alignment methods, alignment with *reduced PPM* shows better results compared to the alignment with initial PPM. Clear distinction can be seen for the threshold vales from -1 to -3 where the alignment with reduced PPM misses only *47.2%, 21.8% and 8%* respectively as opposed to the *68.4%, 30.9% and 12.6%* for the alignment with initial PPM. When the threshold w.r.t. to consensus scores increases, more than 95% of the sequences have been aligned for both the methods. The results with reduced PPM may be better as while the effect due to the redundant 3 positions of the PPM is small, it is not negligible. Due to the reduction of redundant columns having high uncertainty compared to others, alignment with reduced PPM has decreased the number of the calculations and improved the alignments by discarding uncommon mutations of *Guanine (G) or Cytosine (C)* at the redundant positions that where discarded.

As discussed earlier, since both the consensus scores contain *Guanine (G)* at the end position and that it has a considerate probability of 0.328, the success rates can be affected by the distribution differences of the test sequences and the train sequences selected. The results may improve if the selected sequences are shuffled prior to splitting them to train and test instead of selecting the first 1000 sequences.

QUESTION 04

a) Detection of possible mutated Pribnow Box promoters using Non-Intact Query Local Search (NIQLS) for genes that returned negative of the presence of Pribnow Box from both alignment methods

Results: It is highly likely that the Pribnow Box promoter to have mutations including insertions and point mutations due to substitution. Due to these possible mutations, the sequences may not have aligned statistically with either the initial PPM or reduced PPM or with both of them. In this section, we use a NIQLS with *W-search* to detect promoters with insertions possible. The resultant alignments are further analyzed for point mutations specifically for point mutations of *G or C.* The analysis was done for all the threshold values in terms of insertions and point mutations. For convenience, the contingency tables between number of insertion and number of point mutations at the aligned positions of the promoter for threshold -1 is shown as follows. For NIQLS, scores were assigned as; *match = 3, mismatch = -3, gap penalty = 2.*

Table 5: Contingency table for mutated Pribnow Boxes for threshold -1 related to Initial PPM

		Number of substitutions										Total	
		0		1		2		3		3>			
Number of insertions	**0**	347	0.17	92	0.05	86	0.04	34	0.02	10	0.01	**569**	**0.25**
	1	655	0.33	78	0.04	48	0.02	10	0.01	1	0.00	**792**	**0.41**
	2	402	0.20	47	0.02	14	0.01	1	0.00	0	0.00	**464**	**0.24**
	3	116	0.06	13	0.01	1	0.00	0	0.00	0	0.00	**130**	**0.07**
	3>	42	0.02	1	0.00	0	0.00	0	0.00	0	0.00	**43**	**0.02**
Total		**1562**	**0.78**	**231**	**0.12**	**149**	**0.07**	**45**	**0.02**	**11**	**0.01**	**1998**	**1.00**

Table 6: Contingency table for mutated Pribnow Boxes for threshold -1 related to Reduced PPM

		Number of substitutions										Total	
		0		1		2		3		3>			
Number of insertions	**0**	78	0.06	37	0.03	74	0.05	34	0.02	10	0.01	233	**0.12**
	1	415	0.30	70	0.05	48	0.03	10	0.01	1	0.00	544	**0.27**
	2	387	0.28	47	0.03	14	0.01	1	0.00	0	0.00	**449**	**0.22**
	3	116	0.08	13	0.01	1	0.00	0	0.00	0	0.00	**130**	**0.07**
	3>	42	0.03	1	0.00	0	0.00	0	0.00	0	0.00	**43**	**0.02**
Total		**1038**	**0.74**	**168**	**0.12**	**137**	**0.10**	**45**	**0.03**	**11**	**0.01**	**1399**	**1.00**

Considering a maximum of 2 insertions and 2 substitutions (i.e. up to 4 maximum mutations), the detection rates for mutated Pribnow Box promoters were calculated as follows for all the thresholds.

Table 7: Summary of the mutated Pribnow Box detection for both the cases

		-1	**-2**	**-3**	**-4**	**-5**
With Initial PPM	Total Number of Detections (% of negative genes from Question 3)	1998 (92.7)	921 (94.5)	369 (92.9)	133 (88.7)	51 (82.3)
	# Genes with possible mutated Pribnow Boxes	1422	687	196	40	14
	% mutations from Total Number of Detections	**71.2**	**74.6**	**53.1**	**30.1**	**27.5**
With Reduced PPM	Total Number of Detections (% of negative genes from Question 3)	1399 (93.9)	656 (95.2)	229 (91.2)	92 (86.0)	32 (82.1)
	# Genes with possible mutated Pribnow Boxes	1092	440	97	25	5
	% mutations from Total Number of Detections	**78.1**	**67.1**	**42.4**	**27.2**	**15.6**

Discussion: We searched mainly for point mutations and insertions to the promoter. Since Pribnow Boxes are likely to have mutations of *A or T* without compromising its functionality, we used *W-search* with NIQLS to detect possible Pribnow Boxes with insertions present. Based on the analysis shown in Table 5 and Table 6 for threshold -1, we can see that majority of the aligned sequences have either 1 or

2 insertions present. Since we did the alignment using *W-search,* these insertions represent either an insertion of *Guanine(G)* or *Cytosine(C).* Further it shows that number of insertions in aligned sequences rarely exceed 3. (with 2% for both the cases) Since NIQLS considers the alignment of individual bases rather than the alignment of a distribution, a 17% and 6% of genes have been detected with Pribnow boxes with no mutations which reduces to 0% when the thresholds become -4 and -5. Further by analyzing the percentages of point mutations of *G or C,* probability of having such mutations is below 25% in both cases. These point mutations hinder the functionality of Pribnow Box promoter as it changes the number of H-bonds present in the promoter. As can be seen, the point mutations more than 3 is highly unlikely having a 1% of total detected genes for both cases.

Based on the Table 7, out of the negatively detected genes in Question 03, more than 90% have been aligned using NIQLS for first three thresholds. We impose a minimum of one mutation and maximum of 4 mutations criteria to identify the number of possible mutated Pribnow Box detections from the total detections. It can be seen that for the first three thresholds more than 50% out of the total aligned genes have been detected as positive for having a mutated Pribnow Box in both the cases. With a threshold of -4 and -5, it is unlikely that statistical alignment missed many of the possible Pribnow boxes intact even with point mutations present hence they have a lower percentile in both the cases.

QUESTION 05

a) Standard IQLS to locate *Sigma Binding site (TTGACA)* within upstream position from 30 to 50

Results: As described in *Question 01*, after following similar procedure and performing Standard IQLS, within a search region of upstream positions from 30 to 50 we selected the genes with TTGACA Box present.

# seq with mutated *Methionine*	633	11.56%
# seq in sense strand with *TTGACA box* located	1976	70.93%
# seq in antisense strand with *TTGACA box* located	1931	71.73%
Total seq with *TTGACA box* located	3907	71.32%
# seq with *TTGACA box* unlocated	938	17.12%

Discussion: It can be seen that using the *Standard IQLS* more than 70% are identified with a *TTGACA Box* within the search region of 30 to 50 bases upstream. Since it is not possible to use *W-search* to detect the presence of the TTGACA box with the presence of *Guanine (G)* and *Cytosine(C)*, the only method suitable is by performing a traditional *Standard IQLS* that looks for *exact alignment.*

b) Obtaining *Position Probability Matrix (PPM)* and the entropy measures for each position

Results: Using the first 1000 sequences of the selected sequences, the PPM was obtained for 10 positions following the same procedure as in *part(b) of Question 01.* Then the respective column entropies were calculated and using an entropy threshold of *0.01* the redundant positions of 5, 6, 8, 9, 10 were eliminated (columns with blue outline) to derive the reduced PPM.

Table 8: PPM obtained from the first 1000 sequences for TTGACA box with Entropy measures

Base	**Position**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A	0.177	0.176	0.214	0.322	0.254	0.284	0.311	0.280	0.282	0.279
C	0.139	0.161	0.207	0.273	0.270	0.252	0.228	0.240	0.231	0.233
G	0.166	0.198	0.368	0.200	0.225	0.244	0.220	0.229	0.253	0.228
T	0.518	0.465	0.211	0.205	0.251	0.220	0.241	0.251	0.234	0.260
S	0.240	0.158	0.049	0.029	0.003	0.006	0.014	0.004	0.005	0.005

Discussion: Using the *initial PPM,* the consensus sequence identified was *TTGACAAAAA* with a consensus score of *-11.11.* As can be seen, the first 6 positions of the consensus directly represent the TTGACA Box. Using the *reduced PPM* the consensus sequence identified was *TTGAA* with a consensus score of *-4.72*. We can see that even position 5 and 6 corresponding to *C and A* have been eliminated due to the fact that they are more uncertain having all the bases equi-probable at those positions. Further the entropy values for Sigma site PPM is lower compared to Pribnow box PPM.

c) Statistical alignment of sequences in test set with the initial PPM

Results: Following the same procedure described in the *part (a) of Question 03* the statistical alignment to detect the TTGACA box was performed with Initial PPM and the following results were obtained.

Table 9: Results of Statistical Alignment for TTGACA Box search using Initial PPM

	Threshold				
	-1	**-2**	**-3**	**-4**	**-5**
# Genes *Positive* for TTGACA Box	333	2163	2884	2907	2907
% (out of 2907)	11.5%	74.4%	99.2%	100.0%	100.0%

d) Statistical alignment of sequences in test set with the reduced PPM

Results: Following the same procedure described in the *part (b) of Question 03* the statistical alignment to detect the TTGACA box was performed with Reduced PPM and the following results were obtained.

Table 10: Results of Statistical Alignment for TTGACA Box search using Reduced PPM

	Threshold				
	-1	**-2**	**-3**	**-4**	**-5**
# Genes *Positive* for TTGACA Box	1315	2797	2907	2907	2907
% (out of 2907)	45.2%	96.2%	100.0%	100.0%	100.0%

e) Comparison between the Pribnow Box search using Initial PPM and Reduced PPM

		-1	**-2**	**-3**	**-4**	**-5**
With Initial PPM	**# Genes *Negative* for *TTGACA Box***	2574	744	23	0	0
	% (out of 2907)	88.5%	25.6%	0.8%	0.0%	0.0%
With Reduced PPM	**# Genes *Negative* for *TTGACA Box***	1592	110	0	0	0
	% (out of 2907)	54.8%	3.8%	0.0%	0.0%	0.0%

Discussion: By comparing the success rates for statistical alignment methods, alignment with *reduced PPM* shows better results compared to the alignment with initial PPM. Clear distinction can be seen for the threshold vales from -1 to -2 where the alignment with reduced PPM misses only *54.8% and 3.8%* respectively as opposed to the *88.5% and 25.6%* for the alignment with initial PPM. When the threshold w.r.t. to consensus scores increases, *all the sequences* have been aligned for both the methods. Due to the reduction of columns having high uncertainty compared to others, alignment with reduced PPM has decreased the complexity of the calculations and improved the alignments by discarding uncommon mutations specifically at position 5 and 6 in initial PPM as all the bases have equal probability.

f) Detection of possible mutated TTGACA Box promoters using Non-Intact Query Local Search (NIQLS) for genes that returned negative of the presence of TTGACA Box from both alignment methods

Results: As similar to the Pribnow Box, it is likely to have mutation in the *TTGACA Box promoter* in the genes. Having all the sequences positive for the presence of *TTGACA Box* for thresholds -3 to -5, it is a fair doubt that the undetected genes for thresholds -1 and -2 should be having mutated *TTGACA boxes*. Hence as in *Question 04*, we look for the mutations (i.e. possible insertions and point mutations) using traditional NIQLS. In this we look at each position of the aligned sequences and check for possible point mutations after checking for the possible insertions. As in *Question 04*, for the convenience the results will be based on threshold -1. For NIQLS, scores were same as in *Question 04*.

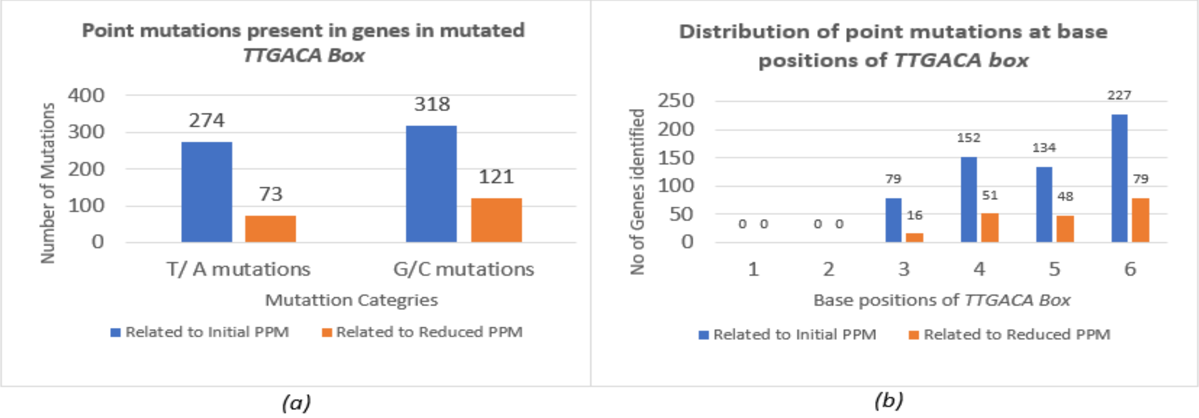

Figure 1: (a) Types of substitution mutations vs frequency of occurrence (b) distribution of mutations at each base position of TTGACA box (threshold -1)

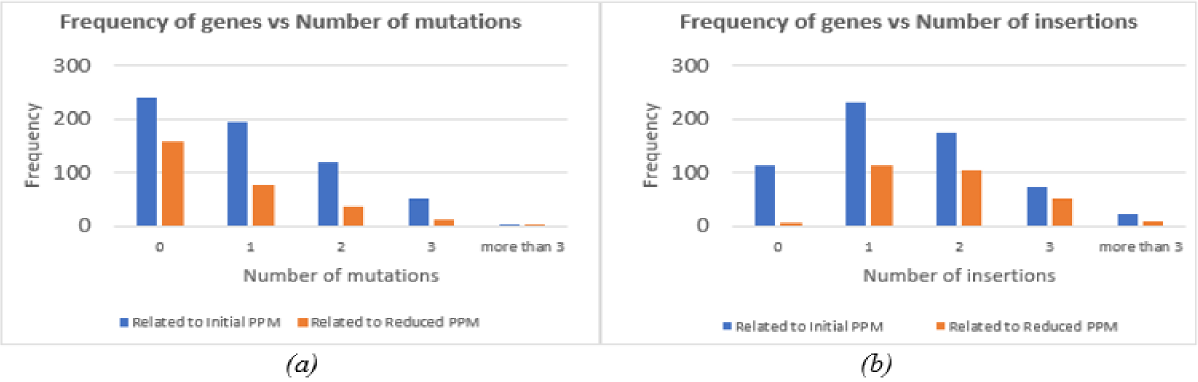

Figure 2: (a) The distribution of number of mutations (substitutions) detected (b) Distribution of number of insertions detected

Considering a maximum of 2 insertions and 2-point mutations (i.e. up to 4 maximum mutations), the detection rates for mutated TTGACA Box promoters were calculated as follows the thresholds -1 and -2.

Table 11: Summary of the mutated TTGACA Box detection for both the cases

		-1	**-2**
With Initial PPM	Total Number of Detections (% of negative genes from *part e*)	608 (23.6)	82 (11.0)
	# Genes with possible mutated TTGACA Boxes	468	65
	% mutations from Total Number of Detections	77.0	79.3
With Reduced PPM	Total Number of Detections (% of negative genes from *part e*)	267 (16.8)	2 (1.8)
	# Genes with possible mutated TTGACA Boxes	200	2
	% mutations from Total Number of Detections	75.0	100.0

Discussion: When searching for the mutated TTGACA boxes, we searched for mainly point mutations and insertions to the promoter. We used *traditional* NIQLS to detect possible TTGACA Boxes with insertions present as the sigma binding sites contains *G and C* that fails the effectiveness of the *W-search*. During the search, for both the statistical alignment scenarios, it was found that the returned aligned x-sequences had gaps when aligned with *TTGACA* box which suggests deletion of some bases from possible sigma binding sites of the genes. Since having gaps in x-sequences during *Local search* is not common, these sequences were considered as genes without *TTGACA boxes.* Due to the high number of deletions received (514 and 49 for -1 ,-2 of Initial PPM) and the partial alignments received from NIQLS, we only get a small portion of sequences that detects mutated TTGACA box promoters.

Based on Figure 1 (a), we can see that out of all the possible substitutions for threshold -1, bases tend mutate for *G or C* rather than *T or A* for both the scenarios. Further, based on Figure 1 (b), it is evident that the occurrence of substitutions at position 1 and 2 of the sigma binding sites (i.e. the bases of *T and* T) is 0 and the peaks of the distribution have occurred at position 4 and 6, both of which corresponds the base *Adenine(A)* for both the scenarios. Further based on Figure 2 (a), the same pattern for the number of point mutations present can be observed as in Question 04, making higher number of point mutations highly unlikely. Further we can see that based on the number of insertions present during the NIQLS alignment, most of the sequences appear to have 1 or 2 insertions.

As shown in Table 11, based on the criteria of selecting the success detection of mutated TTGACA box, even though the selected proportion of genes from *part e* is considerably low (23.6% and 16.8% for the two scenarios with threshold of -1), more than 70% of the sequences have been identified successfully for the presence of mutated TTGACA boxes for both the scenarios.

QUESTION 06

Results: Based on reduced PPM in *Question 2*, the following results were obtained for other genomes.

Accession Code	**Total test seq.s**	**Threshold**									
		-1		**-2**		**-3**		**-4**		**-5**	
		#	%	#	%	#	%	#	%	#	%
NZ_CP015020.1	3948	2088	**52.9**	3118	**79.0**	3623	**91.8**	3818	**96.7**	3898	**98.7**
NZ_CP015853.1	3939	2074	**52.7**	3101	**78.7**	3618	**91.9**	3808	**96.7**	3889	**98.7**
NZ_CP027338.1	3935	2066	**52.5**	3101	**78.8**	3618	**91.9**	3808	**96.7**	3889	**98.7**
NZ_CP027352.1	4067	2129	**52.3**	3190	**78.4**	3746	**92.1**	3937	**96.8**	4016	**98.7**
NZ_CP027387.1	4087	2140	**52.4**	3206	**78.4**	3758	**92.0**	3956	**96.8**	4038	**98.8**
NZ_CP027442.1	4081	2144	**52.5**	3216	**78.8**	3758	**92.1**	3950	**96.8**	4029	**98.7**
NZ_CP027472.1	4044	2143	**53.0**	3186	**78.8**	3721	**92.0**	3911	**96.7**	3993	**98.7**
NZ_CP027555.1	3980	2108	**53.0**	3148	**79.1**	3668	**92.2**	3856	**96.9**	3929	**98.7**
NZ_CP027577.1	4081	2154	**52.8**	3199	**78.4**	3749	**91.9**	3948	**96.7**	4026	**98.7**
NZ_CP028592.1	3973	2097	**52.8**	3129	**78.8**	3645	**91.7**	3834	**96.5**	3918	**98.6**
NZ_CP028607.1	4035	2126	**52.7**	3185	**78.9**	3702	**91.7**	3893	**96.5**	3979	**98.6**
NZ_CP032795.1	3973	2090	**52.6**	3131	**78.8**	3643	**91.7**	3839	**96.6**	3921	**98.7**
NZ_CP032803.1	3964	2092	**52.8**	3133	**79.0**	3640	**91.8**	3827	**96.5**	3908	**98.6**
NZ_CP032808.1	3971	2102	**52.9**	3144	**79.2**	3652	**92.0**	3840	**96.7**	3922	**98.8**
NZ_CP034806.1	3941	2082	**52.8**	3101	**78.7**	3616	**91.8**	3803	**96.5**	3888	**98.7**
NZ_CP037945.1	4033	2098	**52.0**	3161	**78.4**	3714	**92.1**	3903	**96.8**	3983	**98.8**
NZ_CP040305.1	3914	2061	**52.7**	3077	**78.6**	3592	**91.8**	3781	**96.6**	3863	**98.7**
NZ_CP045827.1	3991	2104	**52.7**	3145	**78.8**	3674	**92.1**	3858	**96.7**	3939	**98.7**
NZ_CP045975.1	3902	2067	**53.0**	3087	**79.1**	3584	**91.9**	3768	**96.6**	3849	**98.6**
NZ_CP047378.1	4045	2142	**53.0**	3201	**79.1**	3715	**91.8**	3912	**96.7**	3988	**98.6**
NZ_AP018808.1	3154	1664	**52.8**	2465	**78.2**	2903	**92.0**	3047	**96.6**	3115	**98.8**

As shown above, using the reduced PPM in Question 02, the statistical alignment has resulted in similar hit rates for all the threshold values. This means that although the genomes are differentiated due to its own mutations at replications, the number of Pribnow box detections using reduced PPM results almost the same percentiles compared to the genes available for each genome.