

University of Moratuwa
Faculty of Engineering



Department of Electronic and Telecommunication Engineering

BM4321 : Genomic Signal Processing
Assignment 1

De Silva K.G.G.L.A

150103P

This is submitted as a partial fulfillment for the module
BM4321 – Genomic Signal Processing
Department of Electronic and Telecommunication Engineering
University of Moratuwa
29th of May 2019

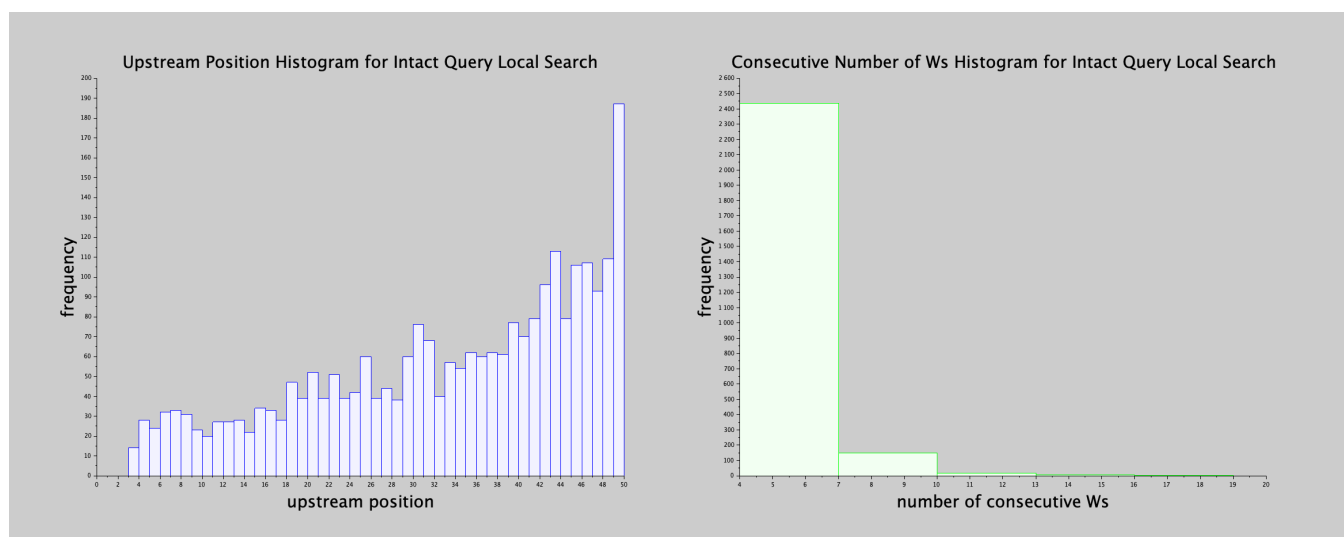
Summary of the Genome of the Organism

Organism	:	<i>Acetobacter tropicalis</i>
Definition	:	Acetobacter tropicalis strain BDGP1 chromosome, complete genome
Accession	:	NZ_CP022699
Version	:	NZ_CP022699.1
Number of Base Pairs	:	3988649 (Chromosome) / 4,139,662 (All)
Number of Proteins	:	3444
Number of Genes in the Sense Strand	:	1741
Number of Genes in the Anti-Sense Strand	:	1703

Results

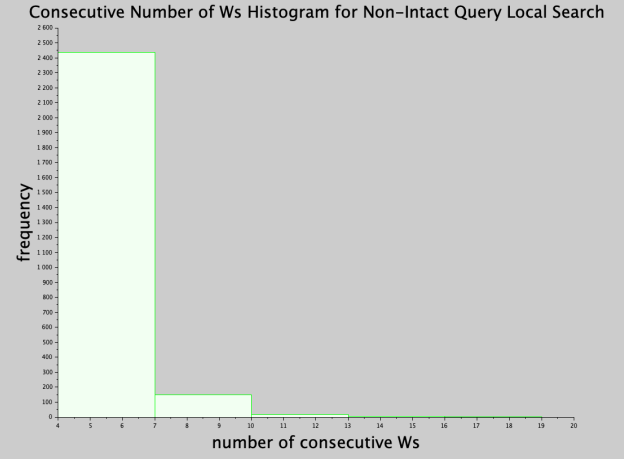
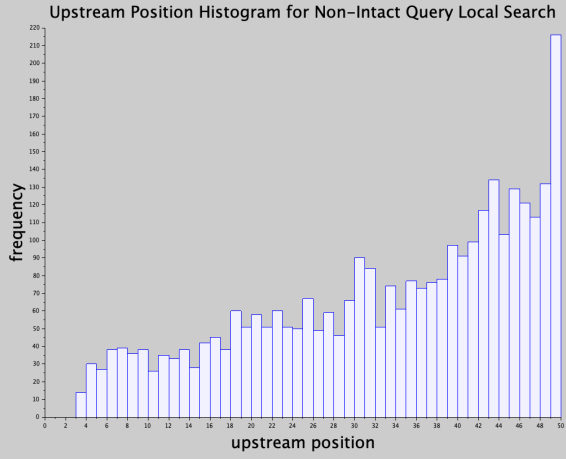
Standard Local Search Alignment (Intact Query)

Number of Genes with Potentials Promoters	:	2610
Percentage of Genes with Potential Promoters	:	75.78 %
Upstream Position Distribution	:	33.69 +/- 12.97
Number of Consecutive Ws Distribution	:	5.05 +/- 1.47
Maximum Number of Consecutive Ws	:	19



Standard Local Search Alignment (Non-Intact Query)

Number of Genes with Potentials Promoters	:	3191
Percentage of Genes with Potential Promoters	:	92.65 %
Upstream Position Distribution	:	33.56 +/- 12.93
Number of Consecutive Ws Distribution	:	5.05 +/- 1.47
Maximum Number of Consecutive Ws	:	19



Statistical Alignment (via Intact Query)

Consensus Sequence

: **ATTTCCGCCAAAATAAGA**

Number of Genes with Potential Promoters

: 3379

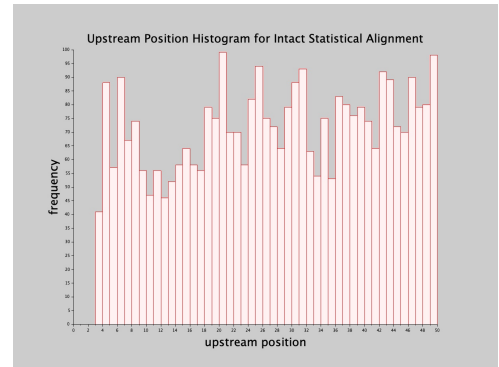
Percentage of Genes with Potential Promoters

: 98.11 %

Upstream Position Distribution

: 28.24 +/- 13.53

	A	C	G	T	
1.	0.5406661	0.0001133	0.0001133	0.4591074	0.6943497
2.	0.4373584	0.0001133	0.0001133	0.562415	0.6988612
3.	0.4758722	0.0001133	0.0001133	0.5239012	0.6921734
4.	0.4427957	0.0001133	0.0001133	0.5569778	0.6975538
5.	0.2311962	0.2547576	0.2189624	0.2950838	0.006623
6.	0.2284776	0.2751473	0.2352741	0.261101	0.002864
7.	0.2275714	0.2688038	0.2529452	0.2506797	0.0017459
8.	0.2407114	0.2456955	0.2737879	0.2398052	0.0015202
9.	0.2339148	0.2683507	0.2547576	0.2429769	0.0013311
10.	0.2574762	0.2601948	0.2552107	0.2271183	0.0014504
11.	0.2579293	0.24343	0.2547576	0.2438831	0.000332
12.	0.2633666	0.2411645	0.247961	0.2475079	0.0005299
13.	0.2683507	0.2352741	0.2488672	0.2475079	0.0011151
14.	0.2529452	0.2443362	0.2511328	0.2515859	0.0000895
15.	0.2443362	0.2398052	0.2511328	0.2647259	0.0007034
16.	0.2615541	0.2398052	0.2452424	0.2533983	0.0005423
17.	0.2574762	0.2524921	0.2470548	0.2429769	0.0002401
18.	0.2511328	0.2497734	0.2543045	0.2447893	0.0000942
19.	0.2606479	0.2497734	0.2552107	0.2343679	0.0007769



Statistical Alignment (via Non-Intact Query)

Consensus Sequence

: **ATTTCCGCGGGAGGGGCC**

Number of Genes with Potential Promoters

: 3355

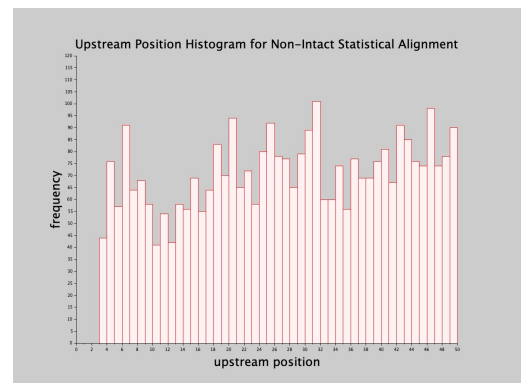
Percentage of Genes with Potential Promoters

: 97.42 %

Upstream Position Distribution

: 28.31 +/- 13.45

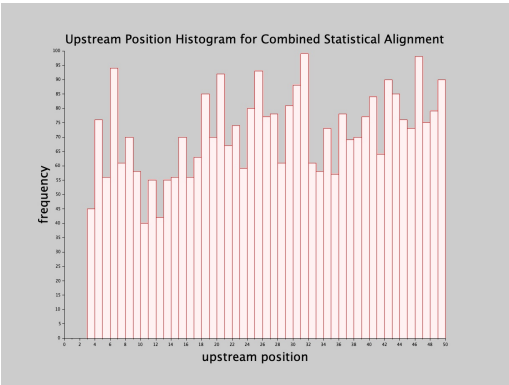
	A	C	G	T	
1.	0.5386406	0.0000931	0.0000931	0.4611732	0.6943658
2.	0.4038175	0.0187151	0.0339851	0.5434823	0.4993345
3.	0.4418063	0.0436685	0.0332402	0.4812849	0.4235471
4.	0.4332402	0.0324953	0.0235568	0.5107076	0.4810815
5.	0.2682495	0.210149	0.182216	0.3393855	0.0285216
6.	0.2120112	0.2864991	0.2600559	0.2414339	0.0059362
7.	0.231378	0.2853818	0.2529795	0.2302607	0.0039231
8.	0.2298883	0.2518622	0.2864991	0.2317505	0.0040648
9.	0.222067	0.2749534	0.2712291	0.2317505	0.0043877
10.	0.2432961	0.2637803	0.2734637	0.21946	0.0034782
11.	0.2485102	0.2514898	0.2634078	0.2365922	0.0007283
12.	0.2507449	0.2481378	0.2671322	0.2339851	0.0011064
13.	0.2626629	0.2451583	0.2548417	0.2373371	0.0007354
14.	0.2447858	0.2496276	0.2626629	0.2429236	0.0004716
15.	0.2380819	0.2507449	0.2596834	0.2514898	0.0004794
16.	0.2570764	0.2391993	0.2611732	0.2425512	0.0006941
17.	0.2470205	0.2552142	0.2578212	0.2399441	0.0003979
18.	0.2436685	0.2682495	0.2619181	0.2261639	0.0021852
19.	0.2470205	0.2630354	0.2615456	0.2283985	0.0015759



Statistical Alignment (via Combined PPM)

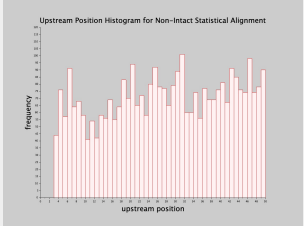
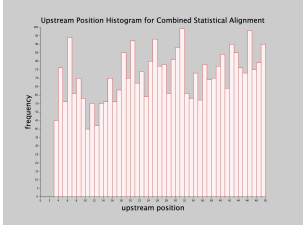
Consensus Sequence : **ATT**TCCGCGGGAGTACCG
Number of Genes with Potential Promoters : 3358
Percentage of Genes with Potential Promoters : 97.50 %
Upstream Position Distribution : 28.30 +/- 13.46

	A	C	G	T	
1.	0.5396533	0.0001032	0.0001032	0.4601403	0.6943546
2.	0.420588	0.0094142	0.0170492	0.5529487	0.5810652
3.	0.4588393	0.0218909	0.0166767	0.5025931	0.5311316
4.	0.4380179	0.0163043	0.011835	0.5338427	0.5700218
5.	0.2497229	0.2324533	0.2005892	0.3172346	0.0141961
6.	0.2202444	0.2808232	0.247665	0.2512675	0.0036862
7.	0.2294747	0.2770928	0.2529623	0.2404702	0.0024858
8.	0.2352998	0.2487789	0.2801435	0.2357778	0.0026047
9.	0.2279909	0.2716521	0.2629933	0.2373637	0.002567
10.	0.2503862	0.2619875	0.2643372	0.2232891	0.0021673
11.	0.2532198	0.2474599	0.2590827	0.2402376	0.0003898
12.	0.2570557	0.2446511	0.2575466	0.2407465	0.0004425
13.	0.2655068	0.2402162	0.2518545	0.2424225	0.0007881
14.	0.2488655	0.2469819	0.2568979	0.2472548	0.0001303
15.	0.2412091	0.245275	0.2554081	0.2581078	0.0003895
16.	0.2593152	0.2395022	0.2532078	0.2479747	0.0004237
17.	0.2522483	0.2538531	0.252438	0.2414605	0.000199
18.	0.2474006	0.2590115	0.2581113	0.2354766	0.0007345
19.	0.2538342	0.2564044	0.2583782	0.2313832	0.0009605



Comparison

Search	Promoter Percentage	Position Distribution (mean \pm stdev)	Distribution
Intact Query Local Search	75.78%	33.69 +/- 12.97	
Non Intact Query Local Search	92.65%	33.56 +/- 12.93	
Statistical Alignment (Intact Query)	98.11%	28.24 +/- 13.53	

Search	Promoter Percentage	Position Distribution (mean ± stdev)	Distribution
Statistical Alignment (Non-Intact Query)	97.42%	28.31 +/- 13.45	
Combined Statistical Alignment	97.50%	28.30 +/- 13.46	

Discussion

In the Intact Query Local Search (IQLS), an alignment is found only when 4 consecutive Ws are present without gaps. In the Non-Intact Query Local Search (NIQLS), an alignment is found not only when WWWW is present, but also when W-WWW, WW-WW, W-WW-W etc are present in the sequence. Therefore, it is intuitive that the NIQLS would yield more alignments than the IQLS. In IQLS, we assume that the promoter is a region where there is at least strictly 4 consecutive Ws and in NIQLS it is assumed that promoter regions where you can find at least 4 Ws in very close proximity. Therefore, it is apparent that promoter percentage from NIQLS (92.65%) is higher than that from the IQLS (75.78 %). It is also important to note that IQLS alignments act as a strict definition of the promoter region while NIQLS alignments act as a more relaxed definition.

After that, the maximum possible number (N) of consecutive Ws that follow after the WWWW alignment was found. In this case, N was equal to 19. The motifs for the statistical alignments were extracted by taking 19 bases towards the downstream direction starting from the detected upstream position of each sequence which potentially contained a promoter region as per the IQLS and NIQLS respectively. The Position Probability Matrices (PPMs) for Intact, Non-Intact and Combined cases were calculated using these motifs. This was followed by computing the Net Information Content of each base position according to each PPM.

$$I = \sum_{j=1}^L \sum_N I_{j,N} \text{ where } I_{j,N} = p_{j,N} \ln \left(\frac{p_{j,N}}{p_{0,N}} \right) ; p_{0,N} = 0.25$$

It was found that only the first four rows of the PPMs contained a considerable information content (as marked in **Red** in the results earlier). Therefore, the statistical alignment was performed by only considering the first four rows of the PPMs. The maximum of probability scores of CCCC and GGGG was used as the lower threshold since a promoter regions is not likely to have CCCC or GGGG.

After performing the statistical alignment, it was noticed that around 98% (98.11%, 97.42%, 97.50%) of sequences contained a promoter region in all three cases (When the alignment was coincided with at least one of the 3 downstream bases, — most commonly the start codon, ATG —that sequence was considered to be a sequence without a promoter region). All the statistical alignments resulted in a higher promoter percentage than the IQLS and NIQLS because the statistical alignment always yield a hit (the alignment with the highest

probability score, $P(S|X) = \sum_j \ln p_{S_j,j}$) except for the instances where the alignment coincides with at least one of the 3 downstream bases.

The upstream position distributions for the IQLS and NIQLS took an exponential shape (increasing towards the downstream direction) while that of the statistical distributions took a shape that loosely resembles a uniform distribution. Therefore, it could be concluded that the promoters as defined by IQLS or NIQLS are more likely to be found closer to the start of the gene (between -1 and -20). And it could also be concluded that according to the statistical alignments, the promoter regions are likely to be found between -1 and -30. The mean upstream position for IQLS and NIQLS was around -33 while for the statistical alignments it was around -28. But it should also be noted that these mean values were accompanied by significant variances. In reality, the promoter regions are usually found in -10 and -35 positions.

We cannot draw a direct connection from the above results and the reality, but however, the above methods introduce different ways of defining the promoter regions and analyzing them.