A Statistical Model for Prokaryotic Promoter Prediction

Joseph Oppon **Winston Hide**
ekow@sanbi.ac.za winhide@sanbi.ac.za
South African National Bioinformatics Institute, University of the Western Cape
Bellville 7535, South Africa

1 Introduction

A simple, fast and sensitive model for identifying and predicting sequences which have non-random statistical properties and therefore biologically active, such as promoters has been developed. This statistical model, Penalized Triplet Frequency Distribution (PTFD) utilizes the information content of promoters (in triplets) and those of other set of sequences of different category e.g. coding sequences (also in triplets) to generate a hash table of scores for each of the 64 possible triplets. The hash table is unique for each set of promoter and non- promoter sequences but generally similar in composition. Cumulative score and therefore the performance of each sequence is assessed by (a) opening a 3bp window and moving along the sequence one bp at a time to extract all the triplets ; (b) obtaining each triplet's corresponding hash table value and (c) summing up all the hash table values of the triplets found in the sequence. A cut-off value obtained by implementing the model on test promoter sequences is used to predict promoters from non promoters. Our prediction results using Penalized Triplet Frequency Distribution (PTFD) method are consistently around 93% True Positives (TP) and 10-14% False Positives (FP). These results are comparable to the kind of results obtained with Neural Network and Hidden Markov Model predictions of promoters from non-promoters (results not shown). Our method is in addition, able to identify promoters sandwiched between other sequences whether these are coding sequences, or non coding sequences between coding sequences with no known promoter activity.

2 Method

Fifteen (15) pairs of promoter and non-promoter sequences were analyzed for their triplet frequency distribution. Each set in the pair consisted of 40 sequences and each sequence length was 50bp long. *E. coli* promoter sequences were selected 50bp upstream of their respective Transcription Start Point [1] 1993). Coding sequences used as non- promoter data were obtained from the current Genbank release. To obtain the triplet frequency distribution of each set of sequence, all the sequences in the set (40) were concatenated. Actual triplet frequency distribution in each set was calculated by using the formula:

$$\text{Frequency of each triplet} = \frac{(\text{No of triplets found inset})(4^3)}{\text{Total number of Nucleotides in Set}}$$

A system of rewarding triplets more common in promoter sequences and penalizing triplets prevalent in non- promoters was implemented by subtracting each triplet's frequency in the promoter set from that of the corresponding frequency in the non-promoter set. Fifteen separate hash tables were created for all fifteen promoter/non promoter pair. An average hash table was also generated from all the fifteen hash tables (Table 1). Two types of tests were conducted with the newly created hash tables. First, 1000 coding sequences were tested with each hash table after a threshold figure has been obtained by testing them on another set of promoter sequences (86) not used in the hash table generation Each set's threshold was selected to obtain the desired percentage of true positive (TP) and used to determine

AAA	=	0.77828	AAC	=	-0.18764	AAG	=	-0.18124	AAT	=	1.08746
ACA	=	0.21749	ACC	=	-0.51175	ACG	=	-0.60343	ACT	=	0.37741
AGA	=	-0.03838	AGC	=	-0.27293	AGG	=	-0.04051	AGT	=	0.23242
ATA	=	0.79747	ATC	=	0.17058	ATG	=	0.02985	ATT	=	0.96805
CAA	=	-0.02345	CAC	=	-0.02345	CAG	=	-0.39021	CAT	=	0.28359
CCA	=	-0.33263	CCC	=	-0.17058	CCG	=	-0.73776	CCT	=	-0.03412
CGA	=	-0.45204	CGC	=	-0.81666	CGG	=	-0.84225	CGT	=	-0.44564
CTA	=	0.30705	CTC	=	-0.00640	CTG	=	-0.85930	CTT	=	0.66527
GAA	=	-0.15352	GAC	=	-0.33903	GAG	=	-0.09382	GAT	=	-0.13007
GCA	=	-0.25800	GCC	=	-0.66740	GCG	=	-1.12370	GCT	=	-0.48189
GGA	=	-0.14499	GGC	=	-1.15782	GGG	=	-0.36888	GGT	=	-0.30278
GTA	=	0.31344	GTC	=	- 0.10661	GTG	=	-0.26866	GTT	=	-0.30918
TAA	=	0.79960	TAC	=	0.04904	TAG	=	0.49895	TAT	=	0.70578
TCA	=	0.21536	TCC	=	0.10875	TCG	=	-0.07676	TCT	=	0.27506
TGA	=	-0.10448	TGC	=	-0.27506	TGG	=	-0.75909	TGT	=	0.79960
TTA	=	0.67593	TTC	=	0.49895	TTG	=	0.72924	TTT	=	1.88919

Table 1: Triplets and their corresponding hash values generated by subtracting the actual frequency of the triplet in coding sequences from those of actual promoter sequences.

Sets	Cut-off	TP		FP		Set	Cut-off	FP	
	---	(/86)	%	(/1000)	%		---	(/1000)	%
1	2.33640	80	93.0	105	10.5	9	1.92010	126	12.6
2	0.99220	"	"	130	13.0	10	1.98400	115	11.5
3	0.32000	"	"	125	12.5	11	-0.54330	199	19.9
4	0.22400	"	"	141	14.1	12	0.25670	149	14.9
5	-1.95200	"	"	168	16.8	13	1.02420	131	13.1
6	1.31220	"	"	150	15.0	14	-0.80000	168	16.8
7	1.95200	"	"	141	14.1	15	-0.73580	139	13.9
8	2.72010	"	"	128	12.8	Av.	1.09819	146	14.6

Figure 1: Results from the fifteen hash tables used to test the same set of promoter and non-promoter sequences. Each threshold value was selected to obtain the specified percentage (93%) of true positives (TP). Also included is the results obtained on the average hash table.

the prediction from the coding sequences Fig. 1.The second test was done by selecting forty nine (49) sequences ranging from 120-900bp which contained promoter sequences. These sequences hereby referred to as inter-orfs were obtained by selecting regions in *E. coli* genome between TAG/TGA/TAA and ATG A 75bp window was opened and performance scores were obtained for each window. For each inter-orf sequence, the best predicted score for the 75bp window was retained together with its score. Thirty-eight (37) of the 49 predictions had either all of the promoter sequences or part of it (Fig. 2).

References

[1] Lisser, L. and Margalit, H., Compilation of *E. coli* mRNA promoter sequences, *Nucleic Acid Res.*, 21(7):1507–1516, 1993.

PREDICTED SEQUENCE	BEST SC.	NAME
TTTTCAATACTTCAATGACCGGTTATCAAGAAATCCTCACTGATCCTTCCTATTCTCGTCAAATCGTTACTCTTA	18.00706	
TGTATTGAGGTTATTAGCGAATAGACAAATCGGTTGCCGTTTGTTGTTTAAAAATTGTTAACAATTTTGTAAAAT	34.80297	lpd
AAATAAAAATACGGCTTGAAACGAGCCAAATAGGGTTCTCGTAGGGGGAATAAGATGAATATTTTAGGTTTTTTT	25.21843	
ATTATCAATTTTAAAAAACTAACAGTTGTCAGCCTGTCCCGCTTATAAGATCATACGCCGTTATACGTTGTTTAC	20.41013	glnS
GTAACAAAGAAATGCAGGAAATCTTTAAAAACTGCCCCTGACACTAAGAGAGTTTTTAAAGGTTCCTTCGCGAGC	15.92599	suc AB
AAACAGGTTCGGAAAACGTTTGCGTTTTTTTTGCCGCAGGTCAATTCCCTTTTGGTCCGAACTCGCACATAATAC	18.30775	pyrd
AAATATTGATAGCCTGAATCAGTATTGATCTGCTGGCAAGAACAGACTACTGTATATAAAAACAGTATAACTTCA	14.05808	umu
TGTTAATTATCCTAAAGGGGTATCTTAGGAATTTACTTTATTTTTCATCCCCATCACTCTTGATCGTTATCAATT	32.28686	narG
GGGAGAAATCGCAACTGTTAATTTTTTATTTCCACGGGTAGAATGCTCGCCGTTTACCTGTTTCGCGCCACTTCC	15.00058	pyrF
TTTTGTCTCACCTTTTAATTTGCTACCCTATCCATACGCACAATAAGGCTATTGTACGTATGCAAATTAATAATA	28.19079	sodB
TTTTTTTATTTAATCGATAACCAGAAGCAATAAAAAATCAAATCGGATTTCACTATATAATCTCACTTTATCTAA	38.71991	
TCGATATCATGGGCCTTAGTCGCCGAATGTACTAGAGAACTAGTGCATTAGCTTATTTTTTTGTTATCATGCTAA	19.58493	aroH
CATATTAAAAATCAGAAAAACTGTAGTTTAGCCGATTTAGCCCCTGTACGTCCCGCTTTGCGTGTATTTCATAAC	20.09882	katE
AAATTTCTGCTAATCGAAAGTTAAATTACGGATCTTCATCACATAAAATAATTTTTTCGATATCTAAAATAAATC	37.61114	manX
CGTTGATATTTTCGCCTAACGTCAGAGGTAGCACCGTAATCCGCGTCTTTTCCCCGCTTTGTTGCGCTCAAGACG	6.83401	flaA
CTAAATAAGTCGCGGGCATAAGGCATATTTTCATCAACAAGGATTTTCACGTTTGTGTTACCTGTATGAGACGAG	15.91532	div
TTTGTATATCTTGGTTGAGTTTATTGGCAACCCTATCACTGCCATGTTTATCGCCGTGTTTGTCGCCTATTATGT	18.45270	
GTACTGTACTAAAGTCACTTAAGGAAACAAACATGAAACACATACCGTTTTTCTTCGCATTCTTTTTTACCTTCC	25.98812	phe
AAAAATGTTATCCACATCACAATTTCGTTTTGCAAATTGGGAATGTTTGCAATTATTTGCCACAGGTAACAAAAA	30.19303	nupG
TGTTTCTCTAACGACTTCCCTTTTAGCCTTAAAGATAAAATCCATTTTAATTTCAGTCATTTAATAAAGAATTTT	41.06753	
AAAAGTTAACCCTTCGACCCACTTCACTCGCGCTTGCATTTTTGCTACTCCACTGCGTCAATTTTCCTGACAGAG	11.32669	
GTAGTATTTTGCTTTTTTCAGAAAATAATCAAAAAAAGTTAGCGTGGTGAATCGATACTTTACCGGTTGAATTTG	30.71754	
CCTTATAACCATTAATTACGAAGCGCAAAAAAAATAATATTTCCTCATTTTCCACAGTGAAGTGATTAACTATGC	25.54891	malT
TTATTCCTCAACCCTTTTTTTAAACATTAAAATTCTTACGTAATTTATAATCTTTAAAAAAAGCATTTAATATTG	47.73726	
GAAACGTTTCGCTGATGGAGAAAAAAATGAAAAAAGGCACCGTTCTTAATTCTGATATTTCATCGGTGATCTCCC	12.10496	rbs
GAGTAAACCTCTCCTTAGTAAACTCTGAAAAAGTAATAACACAACGTTACGACCCGATATTTTCTAAGTCTAATG	18.29702	
TGTTGACTTCGTATTAAACATACCTTATTAAGTTTGAATCTTGTAATTTCCAACGCTTCCCGTTTTATCTTAAAT	31.26764	rho
TTTCTTTACGGTCAATCAGCAAGGTGTTAAATTGATCACGTTTTAGACCATTTTTTCGTCGTGAAACTAAAAAAA	26.87733	cya
TTTCTCGCGACCGGGTTTTTTATTTGTCACGATTTTGCGTTACCCTTGCATCTCTTTGAGGTACAGGGAAAAAAA	21.54236	cdh
AATGCATATAATTTTAACGGCTATTTGGGATTTGCTCAATCTATACGCAAAGAAGTTTAGATGTCCAGATGTATT	22.32491	metBL
GTTTCTGTGAGCAATTATCAGTCAGAATGCTTGATAGGGATAATCGTTCATTGCTATTCTACCTATCGCCATGAA	13.70628	oxyR
TGTTTCTTCATCGTGTCGCATAAAATGTGACCAATAAAACAAATTATGCAATTTTTTAGTTGCATGAACTCGCAT	26.96261	tufB
ATCATTTGATGCCCTTTTTGCACGCTTTCGTACCAGAACCTGGCTCATCAGTGATTTTCTTTGTCATAATCATTG	20.24805	secE
AATAAATTTTATTCATATTGTTATCAACAAGTTATCAAGTATTTTTAATTAAAATGGAAATTGTTTTTGATTTTG	51.44744	aceB
ATTTTGGATAACCCTTCCAGAATTCGATAAATCTCTGGTTTATTGTGCAGTTTATGGTTCCAAAATCGCCTTTTG	24.26313	exA
AAAACTCTGCTTTTCAGGTAATTTATTCCCATAAACTCAGATTTACTGCTGCTTCACGCAGGATCTGAGTTTATG	17.76397	melA
AAATTGCGATGAAATGTGAGGTGAATCAGGGTTTTCACCCGATTTTTGTGCTGATCAGAATTTTTTTTCTTTTTCC	30.92437	groE
CTTTTGTAAAGACGAACAATAAATTTTTACCTTTTGCAGAAACTTTAGTTCGGAACTTCAGGCTATAAAACGAAT	29.40191	htrA
CTTATTGAATATGATTGCTATTTGCATTTAAAATCGAGACCTGGTTTTTCTACTGAAATGATTATGACTTCAATG	27.65771	tonB
CCTCCAGTGCGGTGTTTAAATCTTTGTGGGATCAGGGCATTATCTTACGTGATCAGAATAAACAACCCTCTTTAA	13.44826	hisB
GTTACAGGAAAAGCCAAAGCTGAATCGATTTTATGATTTGGTTCAATTCTTCCTTTAGCGGCATAATGTTTAATG	22.10529	ptsH
CCATAATGTTATACATATCACTCTAAAATGTTTTTTCAATGTTACCTAAAGCGCGATTCTTTGCTAATATGTTCG	31.86892	glpD
TGTCAATGATTGTTGACAGAAACCTTCCTGCTATCCAAATAGTGTCATATCATCATATTAATTGTTCTTTTTTCA	29.80917	tyrR
GCATTTTTACACACTGTGATGAAAAAATCTCCCGTCATTTATAATGATAAGTGTTTTTACCACTTCCCCTTTTCG	32.02034	purR
CAAAAAGGTTGTGTAAAGCAGTCTCGCAAACGTTTGCTTTCCCTGTTAGAATTGCGCCGAATTTTATTTTTCTAC	23.60644	purMN
CTTTTATTCAAACTTTCAAATTAAAATATTTATCTTTCATTTTGCGATCAAAATAACACTTTTAAATCTTTCAAT	49.43671	
TTTTAATAAATGCTCACGTTCTACGGTAAATTTCATAGGTTTACGATGACAATGTTCTGATTAAATTTGAAAAAT	30.17592	
GTAGGGATTGCTCATCAGATGTCCAGATCTTGATGAATTCCTATTTGTGAGCTACGTCTGGACAGTAACTTGTTA	11.42687	btuB
GATTTTTGCAAGCAACATCACGAAATTCCTTACATGACCTCGGTTTAGTTCACAGAAGCCGTGTTCTCATCCTCC	12.91093	

Figure 2: The best predicted sequences (75bp) from each of the 49 inter-orfs using penalized triplet frequency distribution together with their corresponding scores. Underlined are nucleotide sequences found in original promoter sequence (Lisser and Margalit, 1993). Also the rightmost column shows the names of the promoters, which were partially or fully predicted.