

Supplementary Material

Title: Prediction of polyadenylation signals in human DNA sequences using nucleotide frequencies

Firoz Ahmed, Manish Kumar and G. P. S. Raghava*

Institute of Microbial Technology, Sector-39A, Chandigarh, India

Table S1: The dimension of vectors required to present nucleotide sequence around PAS for various types of nucleotide frequencies.

Type of frequency	UP100 +DW100	Two part of UP100+DW100 (UP1_50+UP51_100 + DW1_50+DW51_100)	Three parts of UP100 and DW100	Four parts of UP100 and DW100
(A) Mononucleotide	$(4) \times 2 = 8$	$(4) \times 4 = 16$	$(4) \times 6 = 24$	$(4) \times 8 = 32$
(B) Dinucleotide	$(16) \times 2 = 32$	$(16) \times 4 = 64$	$(16) \times 6 = 96$	$(16) \times 8 = 128$
(C) Trinucleotide	$(64) \times 2 = 128$	$(64) \times 4 = 256$	$(64) \times 6 = 384$	$(64) \times 8 = 512$
(D) Tetranucleotide	$(256) \times 2 = 512$	$(256) \times 4 = 1024$	$(256) \times 6 = 1536$	$(256) \times 8 = 2048$

Table S2: Comparison of nucleotide composition of PAS and pseudo-PAS (pPAS) sequences.

Nuct	-51/-100			-1/-50			+1/+50			+51/+100.		
	Mean of PAS	Mean of pPAS	p-value	Mean of PAS	Mean of pPAS	p-value	Mean of PAS	Mean of pPAS	p-value	Mean of PAS	Mean of pPAS	p-value
A	26.38	29.47	0.00	27.72	32.29	0.00	24.02	34.05	0.00	26.47	29.19	0.00
C	20.84	20.27	3.5E-2	18.75	19.50	3.3E-3	19.06	19.24	0.46	20.59	19.97	0.01
G	20.88	23.14	0.00	19.04	20.31	0.00	19.83	18.86	2.4E-5	22.95	23.27	0.29
T	31.89	26.88	0.00	34.47	27.81	0.00	37.07	27.67	0.00	29.99	27.07	0.00

Table S3: Comparison of dinucleotide composition of PAS and pseudo-PAS (pPAS) sequences.

Dinuct	-51/-100			-1/-50			+1/+50			+51/+100.		
	Mean of PAS	Mean of pPAS	p-value	Mean of PAS	Mean of pPAS	p-value	Mean of PAS	Mean of pPAS	p-value	Mean of PAS	Mean of pPAS	p-value
AA	8.46	9.37	1.3E-5	9.42	12.33	0.00	7.57	13.56	0.00	8.18	9.15	5.0E-6
AC	4.64	5.86	0.00	4.68	5.98	0.00	4.20	6.18	0.00	4.62	6.20	0.00
AG	5.95	7.82	0.00	5.11	6.17	0.00	4.73	6.45	0.00	6.86	7.55	0.00
AT	7.31	6.41	0.00	8.27	7.60	0.00	7.48	7.93	5.4E-4	6.80	6.30	2.0E-4
CA	6.22	6.23	0.99	5.95	6.18	2.9E-2	5.20	6.54	0.00	6.58	6.08	3.0E-6
CC	5.96	5.22	4.0E-6	4.75	4.81	0.69	4.49	4.82	1.1E-2	5.66	4.82	0.00
CG	1.19	2.90	0.00	0.91	2.80	0.00	1.18	2.04	0.00	1.16	2.84	0.00
CT	7.44	5.95	0.00	7.04	5.82	0.00	8.16	5.81	0.00	7.18	6.17	0.00
GA	5.30	7.52	0.00	5.00	6.48	0.00	4.69	6.09	0.00	5.82	7.54	0.00
GC	4.37	3.51	0.00	3.75	2.85	0.00	3.77	2.69	0.00	4.44	3.19	0.00
GG	5.36	6.12	2.0E-6	4.19	5.13	0.00	4.26	4.61	4.9E-3	6.98	6.67	0.08
GT	5.87	5.91	0.64	6.24	5.83	1.5E-4	7.07	5.43	0.00	5.70	5.93	2.8E-2
TA	6.38	6.37	0.95	7.37	7.34	0.83	6.28	7.58	0.00	5.92	6.44	1.0E-4
TC	5.84	5.67	0.12	5.53	5.84	2.2E-3	6.69	5.66	0.00	5.89	5.75	0.17
TG	8.34	6.33	0.00	8.81	6.10	0.00	9.54	5.83	0.00	7.94	6.26	0.00
TT	11.32	8.53	0.00	12.92	8.61	0.00	14.64	8.58	0.00	10.24	8.65	0.00

Table S4: The performance of SVM model based on simple mononucleotide frequency.
 SVM_light parameters: **g:0.01, c:1, j:1.**

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	97.12	35.11	66.07	0.41
-0.9	96.30	40.98	68.61	0.45
-0.8	95.10	46.04	70.54	0.47
-0.7	93.51	48.99	71.22	0.47
-0.6	92.14	52.76	72.42	0.49
-0.5	90.63	55.98	73.28	0.50
-0.4	89.51	58.94	74.21	0.51
-0.3	88.23	61.12	74.66	0.51
-0.2	86.29	63.99	75.13	0.52
-0.1	84.31	67.00	75.64	0.52
0	82.29	68.88	75.58	0.52
0.1	79.93	70.90	75.41	0.51
0.2	77.40	72.10	74.74	0.50
0.3	74.56	73.85	74.21	0.48
0.4	71.21	75.48	73.35	0.47
0.5	66.78	77.58	72.19	0.45
0.6	62.05	80.37	71.22	0.43
0.7	56.64	82.85	69.76	0.41
0.8	50.32	85.34	67.85	0.38
0.9	42.46	88.00	65.26	0.34
1	32.10	91.26	61.72	0.29

Table S5: The performance of SVM model based on simple dinucleotide frequency.
 SVM_light parameters: **g:0.001, c:9, j:1.**

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	97.16	56.45	76.78	0.59
-0.9	96.73	58.21	77.45	0.60
-0.8	96.30	60.27	78.26	0.61
-0.7	95.66	61.77	78.69	0.61
-0.6	95.02	63.78	79.38	0.62
-0.5	94.28	66.31	80.28	0.63
-0.4	93.51	67.77	80.62	0.63
-0.3	92.39	69.10	80.73	0.63
-0.2	91.41	70.77	81.07	0.64
-0.1	89.47	72.87	81.16	0.63
0	87.88	74.37	81.12	0.63
0.1	86.03	76.17	81.09	0.63
0.2	84.01	78.18	81.09	0.62
0.3	81.22	79.94	80.58	0.61
0.4	78.08	81.65	79.87	0.60
0.5	74.26	83.71	78.99	0.58
0.6	70.82	85.60	78.22	0.57
0.7	67.13	87.23	77.19	0.55
0.8	63.00	88.68	75.86	0.53
0.9	58.57	90.27	74.44	0.52
1	52.69	91.51	72.12	0.48

Table S6: The performance of SVM model based on simple trinucleotide frequency.
 SVM_light parameters: g:0.01, c:2, j:1.

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	99.83	17.75	58.73	0.31
-0.9	99.66	24.09	61.82	0.36
-0.8	99.48	30.35	64.87	0.41
-0.7	99.31	37.55	68.39	0.47
-0.6	99.01	44.79	71.87	0.52
-0.5	98.71	50.58	74.61	0.56
-0.4	97.85	57.01	77.40	0.60
-0.3	96.48	63.87	80.15	0.64
-0.2	94.37	69.61	81.97	0.66
-0.1	91.41	75.05	83.22	0.67
0	86.42	80.50	83.45	0.67
0.1	80.62	84.61	82.62	0.65
0.2	73.06	87.91	80.49	0.62
0.3	62.74	90.91	76.85	0.56
0.4	52.34	93.27	72.83	0.50
0.5	41.99	94.68	68.37	0.43
0.6	32.75	96.14	64.48	0.37
0.7	24.84	97.21	61.07	0.32
0.8	18.44	98.41	58.48	0.28
0.9	13.28	98.71	56.05	0.23
1	8.59	99.36	54.03	0.19

Table S7: The performance of SVM model based on simple tetranucleotide frequency.

SVM_light parameters: **g:0.01, c:1, j:2.**

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	100.00	15.77	57.83	0.29
-0.9	100.00	22.67	61.29	0.36
-0.8	99.87	29.53	64.66	0.41
-0.7	99.79	36.73	68.22	0.47
-0.6	99.61	42.95	71.24	0.52
-0.5	99.14	49.29	74.18	0.56
-0.4	98.58	55.12	76.82	0.60
-0.3	97.64	61.29	79.44	0.63
-0.2	96.35	67.00	81.65	0.66
-0.1	93.68	72.82	83.24	0.68
0	90.07	77.58	83.82	0.68
0.1	85.30	82.30	83.80	0.68
0.2	77.87	86.37	82.12	0.64
0.3	69.36	89.46	79.42	0.60
0.4	59.95	92.03	76.01	0.55
0.5	48.04	94.17	71.14	0.48
0.6	38.63	95.97	67.34	0.42
0.7	29.57	97.43	63.54	0.37
0.8	21.44	98.46	60.00	0.31
0.9	13.75	98.89	56.37	0.24
1	8.25	99.49	53.93	0.19

Table S8: The performance of SVM model based on split mononucleotide frequency, where upstream/downstream sequence is divided in two equal parts and frequency of each part is calculated separately. In this case input vector of SVM is of dimension 16 (8 for upstream and 8 for downstream). SVM_light parameters: **g:0.001, c:5, j:1.**

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	96.78	50.19	73.45	0.53
-0.9	96.26	52.38	74.29	0.54
-0.8	95.53	54.31	74.89	0.55
-0.7	95.06	56.97	75.99	0.56
-0.6	94.07	59.15	76.59	0.57
-0.5	93.64	61.38	77.49	0.58
-0.4	92.48	63.35	77.90	0.58
-0.3	91.45	65.71	78.56	0.59
-0.2	89.94	67.17	78.54	0.59
-0.1	88.23	69.44	78.82	0.59
0	86.16	71.15	78.65	0.58
0.1	84.27	73.30	78.78	0.58
0.2	81.82	75.65	78.73	0.58
0.3	78.86	77.67	78.26	0.57
0.4	75.59	79.85	77.73	0.55
0.5	72.24	81.78	77.02	0.54
0.6	67.64	83.75	75.71	0.52
0.7	63.39	85.73	74.57	0.50
0.8	58.96	87.66	73.33	0.49
0.9	54.32	89.63	72.00	0.47
1	48.69	91.38	70.06	0.44

Table S9: The performance of SVM models based on split dinucleotide frequency, where upstream/downstream sequence is divided in two equal parts and frequency of each part is calculated separately. In this case input vector of SVM is of dimension 64 (32 for upstream and 32 for downstream). SVM_light parameters: **g:0.001, c:4, j:1.**

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	98.15	57.91	78.00	0.61
-0.9	97.85	60.14	78.97	0.63
-0.8	97.34	62.37	79.83	0.64
-0.7	96.73	64.51	80.60	0.65
-0.6	96.43	67.00	81.70	0.66
-0.5	95.66	69.27	82.45	0.67
-0.4	94.59	71.20	82.88	0.68
-0.3	93.34	73.38	83.35	0.68
-0.2	92.14	75.23	83.67	0.68
-0.1	90.80	77.11	83.95	0.69
0	88.96	79.00	83.97	0.68
0.1	87.02	81.57	84.29	0.69
0.2	84.44	83.28	83.86	0.68
0.3	81.39	84.35	82.88	0.66
0.4	78.43	85.94	82.19	0.65
0.5	75.08	87.18	81.14	0.63
0.6	71.38	88.98	80.19	0.61
0.7	67.51	90.31	78.93	0.59
0.8	63.94	91.56	77.77	0.58
0.9	59.43	93.06	76.27	0.56
1	53.67	93.96	73.84	0.52

Table S10: The performance of SVM model based on split trinucleotide frequency, where upstream/downstream sequence is divided in two equal parts and frequency of each part is calculated separately. In this case input vector of SVM is of dimension 256 (128 for upstream and 128 for downstream). SVM_light parameters: **g:0.001, c:6, j:1.**

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	98.80	57.65	78.20	0.62
-0.9	98.37	60.05	79.18	0.63
-0.8	97.85	62.75	80.28	0.65
-0.7	97.68	65.67	81.65	0.67
-0.6	96.91	67.90	82.38	0.68
-0.5	95.87	70.55	83.20	0.69
-0.4	94.07	72.82	83.43	0.68
-0.3	92.74	74.71	83.71	0.69
-0.2	91.19	76.85	84.01	0.69
-0.1	89.60	79.08	84.33	0.69
0	87.88	81.53	84.70	0.70
0.1	85.47	83.07	84.27	0.69
0.2	82.81	85.04	83.93	0.68
0.3	79.50	86.67	83.09	0.66
0.4	76.71	88.21	82.47	0.65
0.5	72.84	89.46	81.16	0.63
0.6	70.05	91.21	80.64	0.63
0.7	66.57	92.50	79.55	0.61
0.8	61.28	93.31	77.32	0.58
0.9	56.90	94.43	75.69	0.55
1	52.43	95.37	73.93	0.53

Table S11: The performance of SVM model based on split tetranucleotide frequency, where upstream/downstream sequence is divided in two equal parts and frequency of each part is calculated separately. In this case input vector of SVM is of dimension 1024 (512 for upstream and 512 for downstream). SVM_light parameters: **g:0.001, c:1, j:1**.

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	99.48	47.49	73.45	0.55
-0.9	99.31	50.96	75.11	0.57
-0.8	99.05	54.52	76.76	0.60
-0.7	98.50	58.08	78.26	0.62
-0.6	98.07	61.17	79.59	0.64
-0.5	97.38	64.21	80.77	0.65
-0.4	96.13	66.91	81.50	0.66
-0.3	95.02	69.65	82.32	0.67
-0.2	93.51	73.21	83.35	0.68
-0.1	91.32	76.43	83.86	0.68
0	89.56	79.55	84.55	0.69
0.1	87.32	82.08	84.70	0.69
0.2	84.01	84.10	84.06	0.68
0.3	80.28	86.20	83.24	0.67
0.4	76.28	88.56	82.42	0.65
0.5	72.02	90.06	81.05	0.63
0.6	67.17	91.90	79.55	0.61
0.7	61.67	93.48	77.60	0.58
0.8	55.35	94.51	74.96	0.54
0.9	49.59	95.29	72.47	0.50
1	44.78	96.10	70.47	0.48

Table S12: The performance of SVM model based on split mononucleotide frequency, where upstream/downstream sequence is divided into three nearly equal parts and frequency of each part is calculated separately. In this case input vector of SVM is of dimension 24 (12 for upstream and 12 for downstream). SVM_light parameters: **g:0.01, c:1, j:1.**

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	98.71	28.03	63.33	0.38
-0.9	98.37	33.13	65.71	0.42
-0.8	97.64	37.68	67.62	0.44
-0.7	97.03	42.31	69.64	0.47
-0.6	96.48	47.24	71.82	0.50
-0.5	95.53	51.61	73.54	0.52
-0.4	94.33	55.51	74.89	0.54
-0.3	92.87	59.67	76.24	0.56
-0.2	90.55	63.74	77.12	0.56
-0.1	87.92	68.07	77.98	0.57
0	84.53	72.95	78.73	0.58
0.1	79.63	76.94	78.28	0.57
0.2	74.09	80.20	77.15	0.54
0.3	68.24	83.11	75.69	0.52
0.4	61.19	86.41	73.82	0.49
0.5	54.19	88.38	71.31	0.45
0.6	46.71	90.10	68.43	0.41
0.7	40.52	91.94	66.27	0.38
0.8	33.09	93.48	63.33	0.33
0.9	26.60	94.73	60.71	0.29
1	20.71	96.01	58.41	0.25

Table S13: The performance of SVM model based on split dinucleotide frequency, where upstream/downstream sequence is divided into three nearly equal parts and frequency of each part is calculated separately. In this case input vector of SVM is of dimension 96 (48 for upstream and 48 for downstream). SVM_light parameters: **g:0.001, c:3, j:1.**

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	98.71	55.55	77.10	0.60
-0.9	98.20	58.08	78.11	0.61
-0.8	97.81	60.74	79.25	0.63
-0.7	97.21	63.05	80.11	0.64
-0.6	96.52	65.71	81.09	0.65
-0.5	95.75	67.81	81.76	0.66
-0.4	94.76	70.55	82.64	0.67
-0.3	93.47	73.00	83.22	0.68
-0.2	92.14	75.01	83.56	0.68
-0.1	90.07	76.90	83.48	0.68
0	87.97	78.91	83.43	0.67
0.1	85.43	81.35	83.39	0.67
0.2	83.24	83.07	83.15	0.66
0.3	80.02	84.78	82.40	0.65
0.4	77.18	86.67	81.93	0.64
0.5	73.57	87.96	80.77	0.62
0.6	69.83	89.54	79.70	0.61
0.7	65.41	90.87	78.15	0.58
0.8	61.75	91.94	76.87	0.56
0.9	56.90	93.14	75.04	0.54
1	51.35	94.13	72.77	0.50

Table S14: The performance of SVM model based on split trinucleotide frequency, where upstream/downstream sequence is divided into three nearly equal parts and frequency of each part is calculated separately. In this case input vector of SVM is of dimension 384 (192 for upstream and 192 for downstream). SVM_light parameters: **g:0.001, c:2, j:1.**

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	98.84	53.79	76.29	0.59
-0.9	98.58	56.62	77.58	0.61
-0.8	98.11	59.58	78.82	0.62
-0.7	97.64	62.92	80.26	0.65
-0.6	97.21	65.50	81.33	0.66
-0.5	96.09	68.71	82.38	0.67
-0.4	95.06	71.75	83.39	0.69
-0.3	93.30	73.64	83.45	0.68
-0.2	91.32	75.65	83.48	0.68
-0.1	89.60	77.97	83.78	0.68
0	87.71	79.94	83.82	0.68
0.1	85.73	81.78	83.76	0.68
0.2	83.11	83.67	83.39	0.67
0.3	79.42	85.51	82.47	0.65
0.4	76.45	86.97	81.72	0.64
0.5	73.06	88.43	80.75	0.62
0.6	68.50	89.84	79.18	0.60
0.7	64.20	90.96	77.60	0.57
0.8	59.73	93.06	76.42	0.56
0.9	54.62	94.00	74.33	0.53
1	49.94	94.81	72.40	0.50

Table S15: The performance of SVM model based on split tetranucleotide frequency, where upstream/downstream sequence is divided into three nearly equal parts and frequency of each part is calculated separately. In this case input vector of SVM is of dimension 1536 (768 for upstream and 768 for downstream). SVM_light parameters: **g:0.001, c:1, j:1.**

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	99.53	41.58	70.52	0.50
-0.9	99.44	45.65	72.51	0.53
-0.8	99.27	49.29	74.25	0.56
-0.7	98.80	53.02	75.88	0.58
-0.6	98.45	57.27	77.83	0.61
-0.5	97.72	60.57	79.12	0.63
-0.4	96.99	64.08	80.52	0.65
-0.3	95.32	67.30	81.29	0.65
-0.2	93.94	70.90	82.40	0.67
-0.1	91.96	74.54	83.24	0.68
0	89.82	78.14	83.97	0.68
0.1	86.46	81.18	83.82	0.68
0.2	83.41	84.05	83.73	0.67
0.3	79.63	85.94	82.79	0.66
0.4	74.13	88.08	81.12	0.63
0.5	67.98	90.14	79.08	0.60
0.6	61.75	91.94	76.87	0.56
0.7	56.34	93.36	74.87	0.54
0.8	50.54	94.64	72.62	0.50
0.9	43.96	95.89	69.96	0.47
1	37.99	96.61	67.34	0.43

Table S16: The performance of SVM model based on split mononucleotide frequency, where upstream/downstream sequence is divided into four equal parts and frequency of each part is calculated separately. In this case input vector of SVM is of dimension 32 (16 for upstream and 16 for downstream). SVM_light parameters: **g:0.01, c:1, j:1.**

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	98.84	27.95	63.35	0.38
-0.9	98.45	31.80	65.09	0.41
-0.8	97.94	36.99	67.42	0.44
-0.7	97.51	42.05	69.74	0.48
-0.6	96.56	46.76	71.63	0.50
-0.5	95.66	50.88	73.24	0.52
-0.4	94.37	56.02	75.17	0.55
-0.3	92.61	60.18	76.37	0.56
-0.2	89.99	64.94	77.45	0.57
-0.1	87.58	69.22	78.39	0.58
0	83.93	73.60	78.76	0.58
0.1	79.50	78.74	79.12	0.58
0.2	73.70	81.95	77.83	0.56
0.3	67.47	84.35	75.92	0.53
0.4	60.64	86.50	73.58	0.49
0.5	53.67	89.11	71.42	0.46
0.6	46.33	91.21	68.80	0.42
0.7	39.36	93.14	66.29	0.39
0.8	32.36	94.34	63.39	0.34
0.9	25.35	95.71	60.58	0.30
1	19.04	96.91	58.03	0.25

Table S17: The performance of SVM model based on split dinucleotide frequency, where upstream/downstream sequence is divided into four equal parts and frequency of each part is calculated separately. In this case input vector of SVM is of dimension 128 (64 for upstream and 64 for downstream). SVM_light parameters: **g:0.01, c:1, j:2.**

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	99.91	18.69	59.25	0.32
-0.9	99.74	24.26	61.95	0.37
-0.8	99.70	30.13	64.87	0.42
-0.7	99.57	35.83	67.66	0.46
-0.6	99.44	40.93	70.15	0.50
-0.5	98.97	47.06	72.98	0.54
-0.4	98.41	53.49	75.92	0.58
-0.3	97.51	60.18	78.82	0.62
-0.2	95.96	65.80	80.86	0.65
-0.1	93.77	72.01	82.88	0.67
0	89.90	78.05	83.97	0.68
0.1	84.40	83.24	83.82	0.68
0.2	77.09	86.71	81.91	0.64
0.3	69.79	89.71	79.76	0.61
0.4	59.17	92.07	75.64	0.54
0.5	50.32	94.00	72.19	0.49
0.6	41.81	95.29	68.58	0.44
0.7	33.78	96.66	65.26	0.39
0.8	26.21	98.03	62.17	0.35
0.9	19.34	98.59	59.01	0.29
1	13.62	98.97	56.35	0.24

Table S18: The performance of SVM model based on split trinucleotide frequency, where upstream/downstream sequence is divided into four equal parts and frequency of each part is calculated separately. In this case input vector of SVM is of dimension 512 (256 for upstream and 256 for downstream). SVM_light parameters: **g:0.01, c:3, j:1.**

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	99.96	16.29	58.07	0.30
-0.9	99.83	21.30	60.52	0.34
-0.8	99.74	27.73	63.69	0.40
-0.7	99.44	34.25	66.80	0.44
-0.6	99.40	40.72	70.02	0.50
-0.5	99.05	47.45	73.22	0.54
-0.4	98.45	53.84	76.12	0.58
-0.3	97.51	61.34	79.40	0.63
-0.2	95.79	68.11	81.93	0.66
-0.1	92.57	74.62	83.58	0.68
0	87.80	81.40	84.59	0.69
0.1	81.22	85.86	83.54	0.67
0.2	73.36	89.67	81.52	0.64
0.3	63.00	92.41	77.73	0.58
0.4	52.43	94.86	73.67	0.52
0.5	41.86	96.57	69.25	0.46
0.6	32.32	97.90	65.15	0.40
0.7	23.55	98.59	61.12	0.34
0.8	16.63	99.10	57.92	0.28
0.9	10.83	99.44	55.19	0.22
1	5.89	99.74	52.88	0.16

Table S19: The performance of SVM model based on tetranucleotide frequency, where upstream/downstream sequence is divided into four equal parts and frequency of each part is calculated separately. In this case input vector of SVM is of dimension 2048 (1024 for upstream and 1024 for downstream). SVM_light parameters: **g:0.001, c:2, j:1.**

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	99.18	45.82	72.47	0.53
-0.9	99.10	49.12	74.08	0.56
-0.8	98.84	52.04	75.41	0.58
-0.7	98.11	55.42	76.74	0.59
-0.6	97.64	58.77	78.18	0.61
-0.5	96.99	62.41	79.68	0.63
-0.4	96.22	66.14	81.16	0.65
-0.3	94.80	69.57	82.17	0.67
-0.2	93.51	73.04	83.26	0.68
-0.1	91.36	76.30	83.82	0.68
0	88.44	79.55	83.99	0.68
0.1	85.60	82.17	83.88	0.68
0.2	82.17	84.57	83.37	0.67
0.3	77.91	86.41	82.17	0.65
0.4	73.23	88.51	80.88	0.62
0.5	68.33	90.74	79.55	0.61
0.6	63.47	92.28	77.90	0.58
0.7	57.67	93.48	75.60	0.55
0.8	51.10	94.86	73.00	0.51
0.9	46.11	95.67	70.92	0.48
1	41.21	96.31	68.80	0.45

Table S20: The performance of composition based SVM modules using various features of sequence around PAS (Nuct: Mononucleotides; Dinuct: Dinucleotides; Trinuct: Trinucleotides; Tetnuct: Tetranucleotides; SN: Sensitivity; SP: Specificity; MCC: Matthews correlation coefficient).

Type	100 nt. around PAS			(50+50) nt. around PAS			(33+33+34) nt. around PAS			(25nt x4) nt. around PAS		
	SN	SP	MCC	SN	SP	MCC	SN	SP	MCC	SN	SP	MCC
Nuct	76.06	74.80	0.51	81.99	74.75	0.57	79.9	76.4	0.56	76.8	78	0.55
Dinuct	80.40	80.45	0.61	85.47	82.13	0.68	86.68	80.15	0.67	86.89	77.28	0.64
Trinuct	83.50	82.08	0.66	86.33	83.20	0.70	87.11	80.80	0.68	89.69	73.42	0.64
Tetnuct	85.99	81.01	0.67	87.11	82.85	0.70	83.84	82.73	0.67	92.35	66.10	0.61

Table S21: The performance of SVM based hybrid model (PolyApred) based on split nucleotide frequency, where upstream/downstream sequence is divided in two equal parts and frequency of each part is calculated separately. In this case, frequency of dinucleotide, 2nd order dinucleotide and tetranucleotide are combined. SVM_light parameters: **g:0.001, c:2, j:1.**

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	99.40	49.55	74.44	0.56
-0.9	99.18	53.75	76.44	0.59
-0.8	98.75	57.44	78.07	0.62
-0.7	98.37	60.91	79.61	0.64
-0.6	97.64	64.51	81.05	0.66
-0.5	97.08	67.72	82.38	0.68
-0.4	96.26	71.07	83.65	0.70
-0.3	95.06	73.81	84.42	0.70
-0.2	93.12	76.85	84.98	0.71
-0.1	91.32	79.85	85.58	0.72
0	89.34	82.17	85.75	0.72
0.1	86.38	84.44	85.41	0.71
0.2	82.94	86.88	84.91	0.70
0.3	78.81	88.90	83.86	0.68
0.4	75.20	90.36	82.79	0.66
0.5	70.31	92.11	81.22	0.64
0.6	65.58	93.57	79.59	0.62
0.7	58.36	94.38	76.39	0.57
0.8	52.21	95.37	73.82	0.53
0.9	45.90	96.27	71.12	0.49
1	40.18	97.09	68.67	0.45

Table S22: The performance of SVM based hybrid model of 650 sequences of PAS as positive and 650 sequences of CDS as negative datasets. SVM_light parameters: **g:0.001, c:2, i:3.**

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	99.85	45.54	72.69	0.54
-0.9	99.69	52.15	75.92	0.59
-0.8	99.54	57.38	78.46	0.63
-0.7	99.38	62.46	80.92	0.67
-0.6	98.92	67.54	83.23	0.70
-0.5	98.31	70.92	84.62	0.72
-0.4	96.92	76.00	86.46	0.75
-0.3	95.69	80.00	87.85	0.77
-0.2	94.31	84.00	89.15	0.79
-0.1	92.92	86.46	89.69	0.80
0	90.46	88.31	89.38	0.79
0.1	89.08	90.31	89.69	0.79
0.2	86.31	92.31	89.31	0.79
0.3	82.92	94.15	88.54	0.78
0.4	79.23	95.38	87.31	0.76
0.5	74.62	96.46	85.54	0.73
0.6	69.54	97.54	83.54	0.70
0.7	65.08	98.00	81.54	0.67
0.8	59.08	98.46	78.77	0.63
0.9	53.08	99.23	76.15	0.59
1	46.00	99.54	72.77	0.54

Table S23: The performance of SVM based hybrid model of 400 sequences of PAS as positive and 400 sequences of CDS as negative datasets. SVM_light parameters: **g:0.001, c:2, i:1**.

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	99.75	41.00	70.38	0.50
-0.9	99.50	48.25	73.88	0.56
-0.8	99.50	58.00	78.75	0.63
-0.7	99.50	62.00	80.75	0.66
-0.6	99.25	65.75	82.50	0.69
-0.5	98.50	69.75	84.12	0.71
-0.4	97.25	74.50	85.88	0.74
-0.3	96.00	78.50	87.25	0.76
-0.2	94.00	82.50	88.25	0.77
-0.1	92.75	86.00	89.38	0.79
0	89.75	90.75	90.25	0.81
0.1	86.50	93.25	89.88	0.80
0.2	83.75	94.00	88.88	0.78
0.3	79.75	95.25	87.50	0.76
0.4	76.00	95.50	85.75	0.73
0.5	71.00	96.25	83.62	0.70
0.6	65.25	97.25	81.25	0.66
0.7	60.00	98.25	79.12	0.63
0.8	54.50	98.50	76.50	0.59
0.9	48.25	99.00	73.62	0.55
1	40.75	99.25	70.00	0.49

Table S24: The performance of independent datasets of 250 PAS and 250 CDS on SVM based hybrid model (Table S23) trained on 400 sequences of PAS as positive and 400 sequences of CDS as negative datasets.

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	99.6	48.0	73.8	0.56
-0.9	99.6	56.4	78.0	0.62
-0.8	99.2	62.4	80.8	0.66
-0.7	98.0	67.6	82.8	0.69
-0.6	97.2	72.0	84.6	0.72
-0.5	96.0	74.0	85.0	0.72
-0.4	93.6	78.4	86.0	0.73
-0.3	92.0	81.2	86.6	0.74
-0.2	88.8	86.8	87.8	0.76
-0.1	87.6	88.4	88.0	0.76
0	84.8	90.8	87.8	0.76
0.1	82.4	92.4	87.4	0.75
0.2	79.6	93.6	86.6	0.74
0.3	76.4	94.4	85.4	0.72
0.4	70.4	96.0	83.2	0.69
0.5	66.0	97.6	81.8	0.67
0.6	61.6	97.6	79.6	0.63
0.7	56.4	99.2	77.8	0.62
0.8	50.8	99.6	75.2	0.58
0.9	44.4	99.6	72.0	0.53
1	40.0	99.6	69.8	0.49

Table S25: The performance of independent datasets of 2327 of positive and 2333 as negative (used for polyApred model development) on SVM based hybrid model (Table S22) trained on 650 sequences of PAS as positive and 650 sequences of CDS as negative datasets.

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	99.79	3.39	51.52	0.12
-0.9	99.66	4.67	52.10	0.14
-0.8	99.53	6.13	52.77	0.16
-0.7	99.27	8.14	53.65	0.18
-0.6	98.93	10.63	54.72	0.20
-0.5	98.62	13.42	55.97	0.23
-0.4	98.07	15.73	56.85	0.24
-0.3	97.08	18.65	57.81	0.25
-0.2	96.39	23.02	59.66	0.29
-0.1	95.02	27.43	61.18	0.30
0	93.55	30.65	62.06	0.31
0.1	91.49	35.23	63.33	0.32
0.2	88.87	39.56	64.18	0.33
0.3	86.42	43.89	65.13	0.33
0.4	83.28	49.16	66.20	0.35
0.5	80.10	54.31	67.19	0.36
0.6	75.63	59.79	67.70	0.36
0.7	70.56	64.77	67.66	0.35
0.8	65.41	70.08	67.75	0.36
0.9	58.92	75.01	66.97	0.34
1	51.83	79.64	65.75	0.33

Table S26: The performance of SVM based hybrid model of combined datasets of PolyApred (2327,2333) and independent (650,650) based on split nucleotide frequency. SVM_light parameters: **g:0.001, c:4, j:1.**

Threshold	Sensitivity	Specificity	Accuracy	MCC
-1	98.96	45.12	72.01	0.52
-0.9	98.42	49.41	73.89	0.55
-0.8	98.05	54.48	76.24	0.58
-0.7	97.58	58.57	78.05	0.61
-0.6	96.84	62.15	79.48	0.63
-0.5	96.14	65.81	80.96	0.65
-0.4	95.20	69.53	82.35	0.67
-0.3	93.85	72.68	83.26	0.68
-0.2	91.70	76.40	84.04	0.69
-0.1	89.49	79.79	84.63	0.70
0	86.87	82.67	84.77	0.70
0.1	83.10	85.05	84.08	0.68
0.2	79.44	87.66	83.56	0.67
0.3	74.44	89.81	82.13	0.65
0.4	68.83	91.45	80.15	0.62
0.5	63.76	92.99	78.39	0.59
0.6	58.05	94.40	76.24	0.56
0.7	52.74	95.37	74.08	0.53
0.8	46.86	96.38	71.64	0.50
0.9	40.75	97.25	69.03	0.46
1	33.99	97.92	65.99	0.42

Figure S1: Schematic diagram of 3'-end of pre-mRNA. A number of protein complex recognizes upstream polyadenylation signal (PAS) and downstream G/U rich regions. Premature transcript gets cleaved at polyA site and polyA polymerase carries out polymerization at the newly formed end to form polyA tail at 3'-end of mRNA.

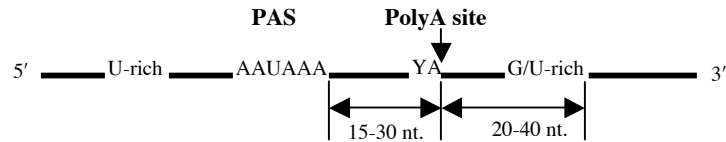


Figure S2: Schematic representation of four framed sequence, having PAS in middle, used to generate the models for SVM. See material & methods for details.

