SUPPLEMENTARY MATERIAL 1: Model Features

Bioinformatics: Application Note

Dragon PolyA Spotter: Predictor of poly(A) motifs within human genomic DNA sequences

Manal Kalkatawi[1,#], Farania Rangkuti[1,#], Michael Schramm[1,#], Boris R. Jankovic[1,#], Allan Kamau[1], Rajesh Chowdhary[2], John A.C. Archer[1], Vladimir B. Bajic[1,*]

[1] Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia

[2] Biomedical Informatics Research Center, MCRF, Marshfield Clinic, 1000 North Oak Avenue, Marshfield, WI 54449, USA

Our model uses selected features such as thermodynamic, structural, statistical and bioelectric properties of nucleotides and polynucleotide sequences. Various thermodynamic and structural properties of dinucleotides are available from a public database created by (Friedel et al., 2008). We selected 110 such properties associated with each of the 16 possible dinucleotide sequences in DNA. These were used as follows to derive model features: when parsing a nucleotide sequence, values of thermodynamic properties of encountered dinucleotides were averaged to generate a total score for the sequence. The range of values obtained in this way were then "stretched" by 10 percent in each direction and all values in between were normalized with respect to the new limits. This transformation was introduced to ensure that the corresponding thermodynamic values of test/validation data would be likely to fall within the stretched range and has been employed in other machine learning systems such as those implemented in WEKA (Hall et al 2009).

Another set of features used was derived from frequency analysis of nucleotide occurrences in the training sequences. To that effect, frequencies of individual nucleotides, dinucleotides and trinucleotides were calculated. Mono/Di/Tri Nucleotide Frequency: This group of features consisted of a/ four single nucleotide frequencies extracted from both upstream and downstream regions; b/ 16 dinucleotide frequencies extracted from both upstream and downstream regions; and c/ 64 trinucleotide frequencies extracted from both upstream and downstream regions. In addition to these, we also used the frequency of T nucleotide in the upstream and downstream regions, as well as G nucleotide frequency in the downstream region of poly(A) signal, even though these features are partially subsumed by previously utilized features. These last features were shown to have good predictive power (Liu et al., 2003). Electron-ion interaction potential (EIIP) of nucleotides (Veljkovic and Slavic, 1972; Veljkovic, 1972; Veljkovic and Lalovic, 1973) is used to characterize individual nucleotides.

Our models also use scores from position weight matrices (PWMs) that are calculated separately for upstream and downstream regions. The PWMs calculated for each window of 10 nucleotides with overlap of 5 nucleotides, which resulted in 38 features.

List of features used in our models

The following properties of nucleotide and polynucleotide sequences are used as model features:

Feature number	Feature Name
1	Twist Thermodynamics property
2	Stacking Thermodynamics property
3	Rise Thermodynamics property
4	Bend Thermodynamics property
5	Tip Thermodynamics property
6	Inclination Thermodynamics property
7	Major Groove Width Thermodynamics property
8	Major Groove Depth Thermodynamics property
9	Major Groove Size Thermodynamics property
10	Major Groove Distance Thermodynamics property
11	Minor Groove Width Thermodynamics property

12	Minor Groove Depth Thermodynamics property
13	Minor Groove Size Thermodynamics property
14	Minor Groove Distance Thermodynamics property
15	Presistance length Thermodynamics property
16	Melting Temperature Thermodynamics property
17	Probability contacting nucleosome core Thermodynamics property
18	Mobility to bend towards major groove Thermodynamics property
19	Mobility to bend towards minor groove Thermodynamics property
20	Propeller Twist Thermodynamics property
21	Clash Strength Thermodynamics property
22	Enthalpy Thermodynamics property
23	Entropy Thermodynamics property
24	Roll (DNA-protein complex) Thermodynamics property
25	Twist (DNA-protein complex) Thermodynamics property
26	Tilt (DNA-protein complex) Thermodynamics property
27	Slide (DNA-protein complex) Thermodynamics property
28	Shift (DNA-protein complex) Thermodynamics property
29	Rise (DNA-protein complex) Thermodynamics property
30	Stacking energy Thermodynamics property
31	Free energy Thermodynamics property
32	Free energy Thermodynamics property
33	Free energy Thermodynamics property
34	Twist (DNA-protein complex) Thermodynamics property
35	Free energy Thermodynamics property
36	Twist_twist Thermodynamics property
37	Tilt_tilt Thermodynamics property
38	Roll_roll Thermodynamics property
39	Twist_tilt Thermodynamics property
40	Twist_roll Thermodynamics property
41	Tilt_roll Thermodynamics property
42	Shift_shift Thermodynamics property
43	Slide_slide Thermodynamics property
44	Rise_rise Thermodynamics property
45	Shift_slide Thermodynamics property
46	Shift_rise Thermodynamics property
47	Slide_rise Thermodynamics property
48	Twist_shift Thermodynamics property
49	Twist_slide Thermodynamics property
50	Twist_rise Thermodynamics property
51	Tilt_shift Thermodynamics property
52	Tilt_slide Thermodynamics property
53	Tilt_rise Thermodynamics property
54	Roll_shift Thermodynamics property
55	Roll _slide Thermodynamics property
56	Roll _rise Thermodynamics property
57	Stacking energy Thermodynamics property
58	Twist Thermodynamics property
59	Tilt Thermodynamics property
60	Roll Thermodynamics property
61	Shift Thermodynamics property
62	Slide Thermodynamics property
63	Rise Thermodynamics property
64	Slide stiffness Thermodynamics property
65	Shift stiffness Thermodynamics property
66	Roll stiffness Thermodynamics property
67	Tilt stiffness Thermodynamics property
68	Twist stiffness Thermodynamics property
69	Free energy Thermodynamics property

70	Free energy Thermodynamics property
71	Free energy Thermodynamics property
72	Free energy Thermodynamics property
73	GC content Thermodynamics property
74	Purine (AG) content Thermodynamics property
75	Keto (GT) content Thermodynamics property
76	Adenine content Thermodynamics property
77	Guanine content Thermodynamics property
78	Cytosine content Thermodynamics property
79	Thymine content Thermodynamics property
80	Tilt (DNA-protein complex) Thermodynamics property
81	Roll (DNA-protein complex) Thermodynamics property
82	Shift (DNA-protein complex) Thermodynamics property
83	Slide (DNA-protein complex) Thermodynamics property
84	Rise (DNA-protein complex) Thermodynamics property
85	Twist Thermodynamics property
86	Tilt Thermodynamics property
87	Roll Thermodynamics property
88	Slide Thermodynamics property
89	Twist Thermodynamics property
90	Tilt Thermodynamics property
91	Roll Thermodynamics property
92	Shift Thermodynamics property
93	Slide Thermodynamics property
94	Rise Thermodynamics property
95	Twist Thermodynamics property
96	Wedge Thermodynamics property
97	Direction Thermodynamics property
98	Rise stiffness Thermodynamics property
99	Melting Temperature Thermodynamics property
100	Stacking energy Thermodynamics property
101	Roll Thermodynamics property
102	Tilt Thermodynamics property
103	Twist Thermodynamics property
104	Roll Thermodynamics property
105	Twist Thermodynamics property
106	Flexibility_slide Thermodynamics property
107	Flexibility_shift Thermodynamics property
108	Enthalpy Thermodynamics property
109	Entropy Thermodynamics property
110	Free energy Thermodynamics property
111	Frequency of A
112	Frequency of AA
113	Frequency of AAA
114	Frequency of AAC
115	Frequency of AAG
116	Frequency of AAT
117	Frequency of AC
118	Frequency of ACA
119	Frequency of ACC
120	Frequency of ACG
121	Frequency of ACT
122	Frequency of AG
123	Frequency of AGA
124	Frequency of AGC
125	Frequency of AGG
126	Frequency of AGT
127	Frequency of AT

128	Frequency of ATA
129	Frequency of ATC
130	Frequency of ATG
131	Frequency of ATT
132	Frequency of C
133	Frequency of CA
134	Frequency of CAA
135	Frequency of CAC
136	Frequency of CAG
137	Frequency of CAT
138	Frequency of CC
139	Frequency of CCA
140	Frequency of CCC
141	Frequency of CCG
142	Frequency of CCT
143	Frequency of CG
144	Frequency of CGA
145	Frequency of CGC
146	Frequency of CGG
147	Frequency of CGT
148	Frequency of CT
149	Frequency of CTA
150	Frequency of CTC
151	Frequency of CTG
152	Frequency of CTT
153	Frequency of G
154	Frequency of GA
155	Frequency of GAA
156	Frequency of GAC
157	Frequency of GAG
158	Frequency of GAT
159	Frequency of GC
160	Frequency of GCA
161	Frequency of GCC
162	Frequency of GCG
163	Frequency of GCT
164	Frequency of GG
165	Frequency of GGA
166	Frequency of GGC
167	Frequency of GGG
168	Frequency of GGT
169	Frequency of GT
170	Frequency of GTA
171	Frequency of GTC
172	Frequency of GTG
173	Frequency of GTT
174	Frequency of T
175	Frequency of TA
176	Frequency of TAA
177	Frequency of TAC
178	Frequency of TAG
179	Frequency of TAT
180	Frequency of TC
181	Frequency of TCA
182	Frequency of TCC
183	Frequency of TCG
184	Frequency of TCT
185	Frequency of TG

186	Frequency of TGA
187	Frequency of TGC
188	Frequency of TGG
189	Frequency of TGT
190	Frequency of TT
191	Frequency of TTA
192	Frequency of TTC
193	Frequency of TTG
194	Frequency of TTT
195	PWM in the upstream (-100:-90) extracted from positive samples
196	PWM in the upstream (-95:-85) extracted from positive samples
197	PWM in the upstream (-90:-80) extracted from positive samples
198	PWM in the upstream (-85:-75) extracted from positive samples
199	PWM in the upstream (-80:-70) extracted from positive samples
200	PWM in the upstream (-75:-65) extracted from positive samples
201	PWM in the upstream (-70:-60) extracted from positive samples
202	PWM in the upstream (-65:-55) extracted from positive samples
203	PWM in the upstream (-60:-50) extracted from positive samples
204	PWM in the upstream (-55:-45) extracted from positive samples
205	PWM in the upstream (-50:-40) extracted from positive samples
206	PWM in the upstream (-45:-35) extracted from positive samples
207	PWM in the upstream (-40:-30) extracted from positive samples
208	PWM in the upstream (-35:-25) extracted from positive samples
209	PWM in the upstream (-30:-20) extracted from positive samples
210	PWM in the upstream (-25:-15) extracted from positive samples
211	PWM in the upstream (-20:-10) extracted from positive samples
212	PWM in the upstream (-15:-5) extracted from positive samples
213	PWM in the upstream (-10:0) extracted from positive samples
214	PWM in the downstream (0:+10) extracted from positive samples
215	PWM in the downstream (+5:+15) extracted from positive samples
216	PWM in the downstream (+10:+20) extracted from positive samples
217	PWM in the downstream (+15:+25) extracted from positive samples
218	PWM in the downstream (+20:+30) extracted from positive samples
219	PWM in the downstream (+25:+35) extracted from positive samples
220	PWM in the downstream (+30:+40) extracted from positive samples
221	PWM in the downstream (+35:+45) extracted from positive samples
222	PWM in the downstream (+40:+50) extracted from positive samples
223	PWM in the downstream (+45:+55) extracted from positive samples
224	PWM in the downstream (+50:+60) extracted from positive samples
225	PWM in the downstream (+55:+65) extracted from positive samples
226	PWM in the downstream (+60:+70) extracted from positive samples
227	PWM in the downstream (+65:+75) extracted from positive samples
228	PWM in the downstream (+70:+80) extracted from positive samples
229	PWM in the downstream (+75:+85) extracted from positive samples
230	PWM in the downstream (+80:+90) extracted from positive samples
231	PWM in the downstream (+85:+95) extracted from positive samples
232	PWM in the downstream (+90:+100) extracted from positive samples
233	PWM in the upstream (-100:-90) extracted from negative samples
234	PWM in the upstream (-95:-85) extracted from negative samples
235	PWM in the upstream (-90:-80) extracted from negative samples
236	PWM in the upstream (-85:-75) extracted from negative samples
237	PWM in the upstream (-80:-70) extracted from negative samples
238	PWM in the upstream (-75:-65) extracted from negative samples
239	PWM in the upstream (-70:-60) extracted from negative samples
240	PWM in the upstream (-65:-55) extracted from negative samples
241	PWM in the upstream (-60:-50) extracted from negative samples
242	PWM in the upstream (-55:-45) extracted from negative samples
243	PWM in the upstream (-50:-40) extracted from negative samples

244	PWM in the upstream (-45:-35) extracted from negative samples
245	PWM in the upstream (-40:-30) extracted from negative samples
246	PWM in the upstream (-35:-25) extracted from negative samples
247	PWM in the upstream (-30:-20) extracted from negative samples
248	PWM in the upstream (-25:-15) extracted from negative samples
249	PWM in the upstream (-20:-10) extracted from negative samples
250	PWM in the upstream (-15:-5) extracted from negative samples
251	PWM in the upstream (-10:0) extracted from negative samples
252	PWM in the downstream (0:+10) extracted from negative samples
253	PWM in the downstream (+5:+15) extracted from negative samples
254	PWM in the downstream (+10:+20) extracted from negative samples
255	PWM in the downstream (+15:+25) extracted from negative samples
256	PWM in the downstream (+20:+30) extracted from negative samples
257	PWM in the downstream (+25:+35) extracted from negative samples
258	PWM in the downstream (+30:+40) extracted from negative samples
259	PWM in the downstream (+35:+45) extracted from negative samples
260	PWM in the downstream (+40:+50) extracted from negative samples
261	PWM in the downstream (+45:+55) extracted from negative samples
262	PWM in the downstream (+50:+60) extracted from negative samples
263	PWM in the downstream (+55:+65) extracted from negative samples
264	PWM in the downstream (+60:+70) extracted from negative samples
265	PWM in the downstream (+65:+75) extracted from negative samples
266	PWM in the downstream (+70:+80) extracted from negative samples
267	PWM in the downstream (+75:+85) extracted from negative samples
268	PWM in the downstream (+80:+90) extracted from negative samples
269	PWM in the downstream (+85:+95) extracted from negative samples
270	PWM in the downstream (+90:+100) extracted from negative samples
271	EIIP properties
272	Frequency of T in upstream
273	Frequency of T in downstream
274	Frequency of G in downstream

Supplementary References

Friedel M, Nikolajewa S, Suhnel J, Wilhelm T. (2008) DiProDB: a database for dinucleotide properties. Nucleic Acids Research, 1–4

Freund Y. and Mason L. The alternating decision tree learning algorithm. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 124--133, 1999

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Liu H, Han H, Li J, Wong L. An in-silico method for prediction of polyadenylation signals in human sequences. Genome Inform. 14, 84-93. 2003

Veljkovic V. and Slavic I. Simple general model pseudopotential, Physical Review Letters, 29(2):105, 1972.

Veljkovic V. Dependence of Fermi energy on atomic number, Physics Letters, 45A(1):41-42, 1973.

Veljkovic V and Lalovic DI. General model pseudopotential for positive-ions. Physics Letters, 45A(1):59-60, 1973.