Supplement for "Accurate prediction of DNA type specific binding residues in protein sequences"

Jian Zhang[1*], Sina Ghadermarzi[2], Akila Katuwawala[2] and Lukasz Kurgan[2*]

[1]School of Computer and Information Technology, Xinyang Normal University, Xinyang, China, 464000

[2]Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA, 23284

*corresponding authors: jianzhang@xynu.edu.cn (J.Z.); lkurgan@vcu.edu (L.K.)

Supplementary Tables

Supplementary Table S1. Features that are utilized in the first layer of DNAgenie. The features are computed in two ways: 1) for individual amino acids when using window size = 5 (i.e., values for each of the five residues in the window); and 2) based on values aggregated over the window size = 11 (typically average and standard deviations for the 11 values within the window).

Feature type	Description	Window Size	Number of features	Number of features per feature type
Relative Amino Acid Propensity (RAAP) for binding	RAAP for {ADNA; BDNA; ssDNA; protein+RNA+small molecules}-binding versus non-binding (per residue)	5	5×1=5	8
	Fraction of residues with high (above average) propensity for {ADNA; BDNA; ssDNA; protein+RNA+small molecules}- binding versus non-binding within the window	11	1	
	Average and standard deviation of RAAP values for {ADNA; BDNA; ssDNA; protein+RNA+small molecules}- binding versus non-binding within the window	11	2×1=2	
Putative Relative Solvent Accessibility (RSA) predicted by ASAquick	Putative RSA values (per residue)	5	5×1=5	9
	Fraction of putative exposed residues within the window (solvent exposed defined as RSA > 5% and as RSA > 20% in TRAINING dataset)	11	2	
	Average and standard deviation of putative RSA values within the window	11	2	
Evolutionary Conservation (ECO) produced by HHblits	ECO scores (per residue)	5	5×1=5	9
	Fraction of conserved residues within the window (conserved is defined as ECO > average in TRAINING dataset and as ECO > average + stdev in TRAINING dataset)	11	2	
	Average and standard deviation of ECO scores within the window	11	2	
Hydrophobicity, Polarity & Charge quantified with AAindex database	Hydrophobicity values using two amino acid hydrophobicity indices (per residue)	5	5×2=10	46
	Average and standard deviation of hydrophobicity values within the window using two amino acid hydrophobicity indices	11	2×2=4	
	Fraction of hydrophobic residues within the window	11	1	
	Polarity values using two polarity indices (per residue)	5	5×2=10	
	Average and standard deviation of polarity values within the window using two polarity indices	11	2×2=4	
	Fraction of polar residues within the window	11	1	
	Positive and negative charge values (per residue)	5	5×2=10	
	Average and standard deviation of positive and negative charge values within the window	11	2×2=4	
	Fraction of positively and negatively charged residues within the window	11	2	
Putative Disorder predicted by IUpred2A	Putative protein-binding disorder probability (per residue)	5	5×1=5	7
	Average and standard deviation of putative protein-binding disorder probabilities within the window	11	2	
Putative Secondary Structure predicted by PSIPRED	Putative secondary structure coded using three bits (for helix, beta-sheet and coil) (per residue)	5	5×3=15	65
	Fraction of residues in putative helix conformation, in putative coil conformation, and in putative beta-sheet conformation within the window	11	3	
	Fraction of residues in putative helix and beta sheet conformations within the window	11	1	
	Fraction of residues in the longest putative helix segment, the longest putative beta sheet segment, and the longest putative coil segment within the window	11	3	
	Residue position within current putative secondary segment (linear distance from the terminus of the current secondary structure segment)	N/A	1	
	Average length of putative secondary structure segments in the sequence	N/A	1	
	Fraction of residues in putative helix conformation, in putative coil conformation, and in putative beta-sheet conformation in the whole sequence	N/A	3	
	Presence of a secondary structure motif at the position of the predicted residue coded using fourteen bits per residue (X[H|E|C](the head motif), [H|E|C]X(the tail motif), CHC, CHE, EHC, EHE, HCH, ECH, HCE, ECE, CEC, HEC, CEH, and HEH)	N/A	14	
	Fraction of residues in a given motif type in the sequence (XHE,XHC,XEH,XEC, XCH, XCE, HEX, HCX, EHX, ECX, EHX, ECX, CHC, CHE, EHC, EHE, HCH, ECH, HCE, ECE, CEC, HEC, CEH, HEH)	N/A	24	
Physicochemical properties	Presence of aliphatic, sulphur containing, aromatic, hydrophobic, charged, polar, positive, acidic, small, tiny, and hydroxylic amino acids coded using 11 bits, one bit per property (per residue)	3	3×11=33	40
	Fraction of aliphatic, sulphur containing, aromatic, hydrophobic, charged, polar, positive, acidic, small, tiny, and hydroxylic amino acids within the window	11	7	

Sequence-based features	Linear distance (in sequence positions) to nearest putative helix, to the nearest putative coil, and to the nearest putative beta sheet	N/A	3	
	Linear distance (in sequence positions) to nearest conserved residue (conserved = ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset)	N/A	2	
	Linear distance (in sequence positions) to nearest putative solvent exposed residue (solvent exposed = RSA > 5% or RSA > 20%)	N/A	2	
	Linear distance (in sequence positions) to nearest sequence terminus	N/A	1	
	Linear distance (in sequence positions) to nearest residues with high (above average) propensity for {ADNA; BDNA; ssDNA; or protein+RNA+small molecules}-binding	N/A	1	34
	Linear distance (in sequence positions) to nearest residues that is hydrophobic, positively charged, negatively charged, polar, aliphatic, sulphur containing, aromatic, acidic, small, tiny, and hydroxylic	N/A	11	
	Linear distance (in sequence positions) to nearest secondary structure motifs (X[H|E|C](head), [H|E|C]X(tail), CHC, CHE, EHC, EHE, HCH, ECH, HCE, ECE, CEC, HEC, CEH, HEH)	N/A	14	
Features that combine multiple structural, physicochemical and evolutionary properties	Fraction of putative surface residues (two thresholds: RSA > 5% or RSA > 20%) in the set of residues with high propensity for {ADNA; BDNA; ssDNA; or protein+RNA+small molecules}-binding within the window	11	2×1=2	
	Fraction of conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset) in the set of residues with high propensity for {ADNA; BDNA; ssDNA; or protein+RNA+small molecules}-binding within the window	11	2×1=2	
	Fraction of aliphatic, sulphur containing, aromatic, hydrophobic, charged, polar, positive, acidic, small, tiny, and hydroxylic amino acids residues (11 residue types) in the set of residues with high propensity for {ADNA; BDNA; ssDNA; or protein+RNA+small molecules}-binding within the window	11	11×1=11	
	Fraction of residues with high propensity for {ADNA; BDNA; ssDNA; or protein+RNA+small molecules}-binding located in putative helix/coil/sheet segments within the window	11	3×1=3	
	Fraction of conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset) among putative surface residues (two thresholds: RSA > 5% or RSA > 20%) within the window	11	2×2=4	
	Fraction of aliphatic, sulphur containing, aromatic, hydrophobic, charged, polar, positive, acidic, small, tiny, and hydroxylic amino acids residues (11 residue types) among putative surface residues (two thresholds: RSA > 5% or RSA > 20%) within the window	11	2×11=22	
	Fraction of putative surface residues (two thresholds: RSA > 5% or RSA > 20%) located in putative helix/coil/sheet segments within the window	11	2×3=6	
	Fraction of aliphatic, sulphur containing, aromatic, hydrophobic, charged, polar, positive, acidic, small, tiny, and hydroxylic amino acids residues (11 residue types) among the conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset) within the window	11	2×11=22	
	Fraction of conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset) located in putative helix/coil/sheet segments within the window	11	2×3=6	
	Fraction of aliphatic, sulphur containing, aromatic, hydrophobic, charged, polar, positive, acidic, small, tiny, and hydroxylic amino acids residues (11 residue types) located in putative helix/coil/sheet segments within the window	11	11×3=33	
	Fraction of residues with high propensity for {ADNA; BDNA; ssDNA; or protein+RNA+small molecules}-binding on putative surface (two thresholds: RSA > 5% or RSA > 20%) in the entire protein sequence	N/A	2×1=2	
	Fraction of conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset) located on putative surface (two thresholds: RSA > 5% or RSA > 20%) in the entire protein sequence	N/A	2×2=4	205
	Fraction of putative helix, coil, and sheet residues on putative surface (two thresholds: RSA > 5% or RSA > 20%) in the entire protein sequence	N/A	2×3=6	
	Fraction of aliphatic, sulphur containing, aromatic, hydrophobic, charged, polar, positive, acidic, small, tiny, and hydroxylic amino acids residues (11 residue types) on putative surface (two thresholds: RSA > 5% or RSA > 20%) in the entire protein sequence	N/A	2×11=22	
	Fraction of residues in different secondary structure motifs (X[H|E|C](head), [H|E|C]X(tail), CHC, CHE, EHC, EHE, HCH, ECH, HCE, ECE, CEC, HEC, CEH, HEH) on putative surface (two thresholds: RSA > 5% or RSA > 20%) in the entire protein sequence	N/A	2×14=28	
	Fraction of residues with high (above average) propensity for {ADNA; BDNA; ssDNA; protein+RNA+small molecules}- binding versus non-binding within the segment of putative secondary structure that includes the predicted residue	N/A	1	
	Fraction of putative surface residues (two thresholds: RSA > 5% or RSA > 20%) within segment of putative secondary structure that includes the predicted residue	N/A	2	
	Fraction of conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset) within segment of putative secondary structure that includes the predicted residue	N/A	2	
	Fraction of aliphatic, sulphur containing, aromatic, hydrophobic, charged, polar, positive, acidic, small, tiny, and hydroxylic amino acids residues (11 residue types) within segment of putative secondary structure that includes the predicted residue	N/A	11	
	Fraction of residues with high propensity for {ADNA; BDNA, ssDNA; or protein+RNA+small molecules}-binding within motif of putative secondary structure that includes the predicted residue	N/A	2	
	Fraction of putative residues surface (two thresholds: RSA > 5% or RSA > 20%) within motif of putative secondary structure that includes the predicted residue	N/A	2	
	Fraction of conserved residues (two thresholds: ECO > average in TRAINING dataset or ECO > average + stdev in TRAINING dataset) residues within motif of putative secondary structure that includes the predicted residue	N/A	11	
	Fraction of aliphatic, sulphur containing, aromatic, hydrophobic, charged, polar, positive, acidic, small, tiny, and hydroxylic amino acids residues (11 residue types) within motif of putative secondary structure that includes the predicted residue	N/A	1	
		N/A	2	
		N/A	11	
TOTAL number of features				423

Supplementary Table S2. Features that are utilized in the second layer of DNAgenie. Features are computed based on values aggregated over the entire window for window size = 5 (typically average and standard deviations for the 5 values within the window).

Feature type	Description	Window Size	Number of features	Number of features per feature type
Local features computed from the A-DNA-binding prediction produced in the first layer	Putative ADNA-binding propensities generated in the first layer (per residue)	N/A	1	10
	Putative ADNA-binding propensities multiply by RSA and ECO scores	N/A	2	
	Average of putative ADNA-binding propensities generated in the first layer	11	1	
	Fraction of putative ADNA-binding residues within the window	11	1	
	Fraction of residues with high putative ADNA-binding propensities generated in the first layer (> average) among putative surface residues (RSA > 20%)	11	1	
	Fraction of residues with high putative ADNA-binding propensities generated in the first layer (> average) among putative conserved residues (ECO > average in TRAINING dataset)	11	1	
	Fraction of residues with high putative ADNA-binding propensities generated in the first layer (> average) among hydrophobic, polar and charged residues	11	1×3=3	
Local features computed from the B-DNA-binding prediction produced in the first layer	Putative BDNA-binding propensities generated in the first layer (per residue)	N/A	1	10
	Putative ADNA-binding propensities multiply by RSA and ECO scores	N/A	2	
	Average of putative BDNA-binding propensities generated in the first layer	11	1	
	Fraction of putative BDNA-binding residues within the window	11	1	
	Fraction of residues with high putative BDNA-binding propensities generated in the first layer (> average) among putative surface residues (RSA > 20%)	11	1	
	Fraction of residues with high putative BDNA-binding propensities generated in the first layer (> average) among putative conserved residues (ECO > average in TRAINING dataset)	11	1	
	Fraction of residues with high putative BDNA-binding propensities generated in the first layer (> average) among hydrophobic, polar and charged residues	11	1×3=3	
Local features computed from the ssDNA-binding prediction produced in the first layer	Putative ssDNA-binding propensities generated in the first layer (per residue)	N/A	1	10
	Putative ADNA-binding propensities multiply by RSA and ECO scores	N/A	2	
	Average of putative ssDNA-binding propensities generated in the first layer	11	1	
	Fraction of putative ssDNA-binding residues within the window	11	1	
	Fraction of residues with high putative ssDNA-binding propensities generated in the first layer (> average) among putative surface residues (RSA > 20%)	11	1	
	Fraction of residues with high putative ssDNA-binding propensities generated in the first layer (> average) among putative conserved residues (ECO > average in TRAINING dataset)	11	1	
	Fraction of residues with high putative ssDNA-binding propensities generated in the first layer (> average) among hydrophobic, polar and charged residues	11	1×3=3	
Local features computed from the protein+RNA+small molecules –binding (PSB) prediction produced in the first layer	Putative PSB propensities generated in the first layer (per residue)	N/A	1	10
	Putative PSB propensities multiply by RSA and ECO scores	N/A	2	
	Average of putative PSB propensities generated in the first layer	11	1	
	Fraction of putative PSB residues within the window	11	1	
	Fraction of residues with high putative PSB propensities generated in the first layer (> average) among putative surface residues (RSA > 20%)	11	1	
	Fraction of residues with high putative PSB propensities generated in the first layer (> average) among putative conserved residues (ECO > average in TRAINING dataset)	11	1	
	Fraction of residues with high putative PSB propensities generated in the first layer (> average) among hydrophobic, polar and charged residues	11	1×3=3	
Putative propensity difference (diff_A_max) = max{ADNA, BDNA, ssDNA, protein+RNA+small molecules} minus the A-DNA-binding propensity	Putative diff_A_max propensities generated in the first layer (per residue)	N/A	1	8
	Average of putative diff_A_max propensities generated in the first layer	11	1	
	Fraction of putative most likely A-DNA binding residues (diff_A_max>0) within the window	11	1	
	Fraction of residues with high putative diff_A_max propensities generated in the first layer (> average) among putative surface residues (RSA > 20%)	11	1	
	Fraction of residues with high putative diff_A_max propensities generated in the first layer (> average) among putative conserved residues (ECO > average in TRAINING dataset)	11	1	
	Fraction of residues with high putative diff_A_max propensities generated in the first layer (> average) among hydrophobic, polar and charged residues	11	1×3=3	
Putative propensity difference (diff_B_max) = max{ADNA, BDNA, ssDNA, protein+RNA+small molecules} minus the B-DNA-binding propensity	Putative diff_B_max propensities generated in the first layer (per residue)	N/A	1	8
	Average of putative diff_B_max propensities generated in the first layer	11	1	
	Fraction of putative most likely B-DNA binding residues (diff_B_max>0) within the window	11	1	
	Fraction of residues with high putative diff_B_max propensities generated in the first layer (> average) among putative surface residues (RSA > 20%)	11	1	
	Fraction of residues with high putative diff_B_max propensities generated in the first layer (> average) among putative conserved residues (ECO > average in TRAINING dataset)	11	1	
	Fraction of residues with high putative diff_B_max propensities generated in the first layer (> average) among hydrophobic, polar and charged residues	11	1×3=3	

Putative propensity difference (diff_ss_max) = max{ADNA, BDNA, ssDNA, protein+RNA+small molecules} minus the ssDNA-binding propensity	Putative diff_ss_max propensities generated in the first layer (per residue)	N/A	1	8
	Average of putative diff_ss_max propensities generated in the first layer	11	1	
	Fraction of putative most likely ss-DNA binding residues (diff_ss_max>0) within the window	11	1	
	Fraction of residues with high putative diff_ss_max propensities generated in the first layer (> average) among putative surface residues (RSA > 20%)	11	1	
	Fraction of residues with high putative diff_ss_max propensities generated in the first layer (> average) among putative conserved residues (ECO > average in TRAINING dataset)	11	1	
	Fraction of residues with high putative diff_ss_max propensities generated in the first layer (> average) among hydrophobic, polar and charged residues	11	1×3=3	
Sequence-based features	Fraction of putative A-DNA binding residues on exposed, conserved, hydrophobic, charged, and polar residues	N/A	5	22
	Fraction of putative B-DNA binding residues on exposed, conserved, hydrophobic, charged, and polar residues	N/A	5	
	Fraction of putative ss-DNA binding residues on exposed, conserved, hydrophobic, charged, and polar residues	N/A	5	
	Fraction of putative A-DNA binding residues on the sequence	N/A	1	
	Fraction of putative B-DNA binding residues on the sequence	N/A	1	
	Fraction of putative ss-DNA binding residues on the sequence	N/A	1	
	Fraction of putative PSB on the sequence	N/A	1	
	Fraction of putative diff_A_max on the sequence	N/A	1	
	Fraction of putative diff_B_max on the sequence	N/A	1	
	Fraction of putative diff_ss_max on the sequence	N/A	1	
TOTAL number of features				86

Supplementary Table S3. Comparison of different machine learning algorithms for the prediction of the A-DNA, B-DNA and ssDNA-binding residues. The results are computed by performing five-fold cross-validation on the training dataset. The metrics are defined in the Methods section in the main text. The binary assessments (sensitivity and accuracy) are normalized between different predictors to maintain the same specificity = 0.95. Bold font identifies the most accurate predictor for a given DNA type.

DNA type	Predictive model	Sensitivity	Accuracy	AUC	AULCratio	AUCPC-D	RatioCPR-D	AUCPC-L	RatioCPR-L	AUOPC	RatioOPR
A-DNA-binding	Random predictor	0.046	0.945	0.490	0.947	0.506	0.903	0.507	0.940	0.510	0.908
	Logistic regression	**0.593**	**0.948**	**0.907**	**16.672**	**0.246**	2.527	**0.099**	**11.717**	**0.085**	**12.696**
	Weighted *k*NN	0.362	0.947	0.807	9.615	0.328	**2.539**	0.228	5.340	0.190	7.659
	Naïve Bayes	0.174	0.945	0.804	3.518	0.384	1.312	0.218	2.588	0.191	3.502
B-DNA-binding	Random predictor	0.052	0.932	0.504	1.020	0.488	0.954	0.491	0.933	0.496	1.020
	Logistic regression	**0.560**	**0.942**	**0.904**	**15.281**	**0.239**	**2.865**	**0.136**	**6.822**	**0.093**	**11.693**
	Weighted *k*NN	0.316	0.937	0.788	8.158	0.350	1.658	0.245	4.003	0.209	4.911
	Naïve Bayes	0.135	0.933	0.734	2.742	0.410	1.306	0.262	2.429	0.265	2.592
ssDNA-binding	Random predictor	0.050	0.947	0.498	1.055	0.499	1.084	0.504	0.984	0.502	0.980
	Logistic regression	**0.573**	**0.949**	**0.912**	**16.253**	**0.124**	**7.100**	**0.113**	**7.039**	**0.084**	**11.694**
	Weighted *k*NN	0.307	0.949	0.782	8.515	0.284	3.263	0.269	3.608	0.218	5.384
	Naïve Bayes	0.156	0.948	0.766	3.388	0.283	2.158	0.299	2.018	0.231	3.167

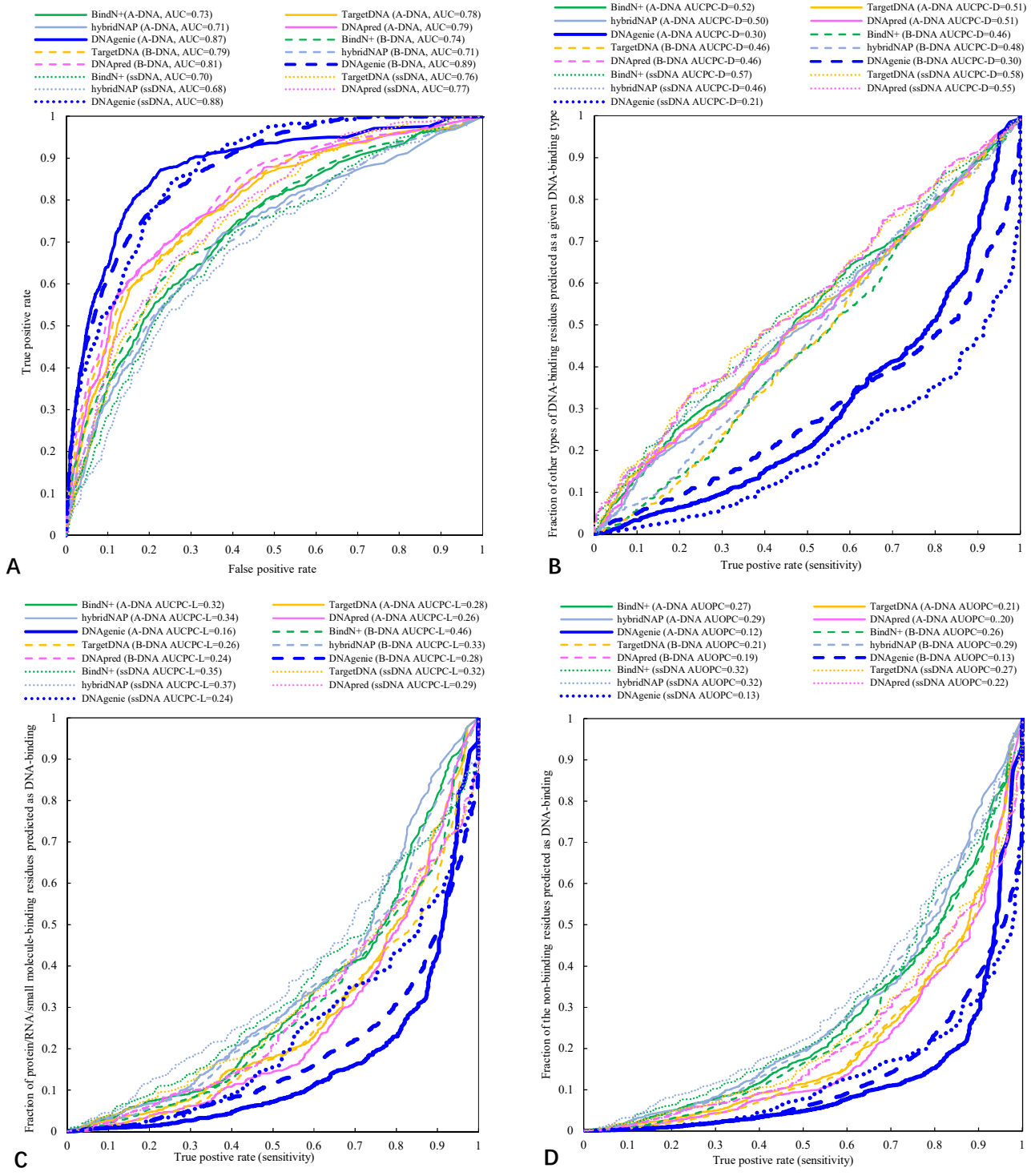
Supplementary Table S4. Relative amino acid propensities (RAAP) for binding A-DNA, B-DNA and ssDNA.

Amino Acid Type	Propensity for A-DNA binding	Propensity for B-DNA binding	Propensity for ssDNA binding
A	0.03	0.10	0.03
R	1.00	1.00	1.00
N	0.50	0.38	0.35
D	0.10	0.06	0.19
C	0.18	0.08	0.00
Q	0.35	0.36	0.14
E	0.09	0.03	0.06
G	0.22	0.22	0.34
H	0.56	0.56	0.50
I	0.14	0.09	0.22
L	0.06	0.00	0.10
K	0.76	0.76	0.56
M	0.15	0.18	0.15
F	0.19	0.19	0.38
P	0.00	0.15	0.08
S	0.34	0.34	0.18
T	0.36	0.50	0.36
W	0.62	0.35	0.62
Y	0.38	0.62	0.76
V	0.08	0.14	0.09

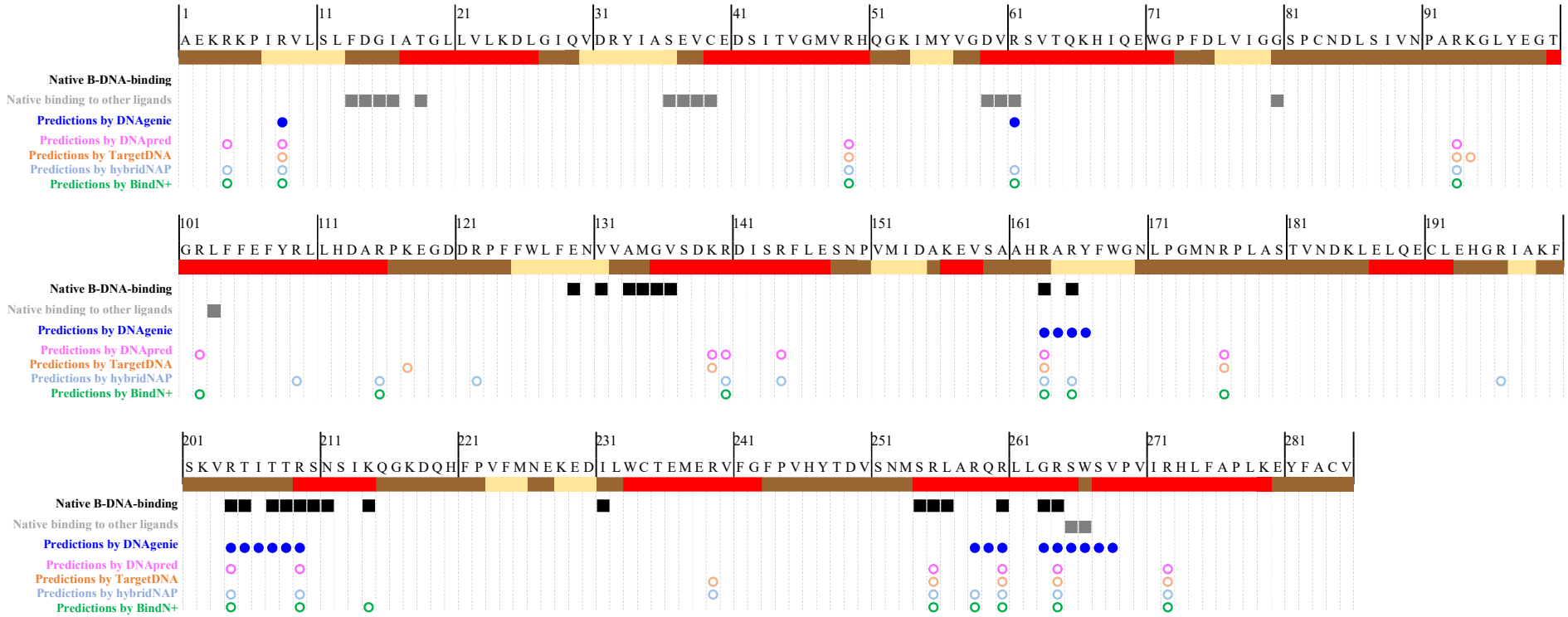
Supplementary Table S5. Fraction of the training proteins correctly predicted with the DNA type by using the relative amino acid propensities (RAAP) for binding A-DNA, B-DNA and ssDNA that are listed in Supplementary Table S4.

DNA-binding index type	ADNA-binding proteins	BDNA-binding proteins	ssDNA-binding proteins
A-DNA	**69.6%**	23.2%	7.1%
B-DNA	31.5%	**54.2%**	14.3%
ssDNA	18.2%	24.2%	**57.6%**

Supplementary Figures



Supplementary Figure S1. The ROC curves (panel A), the cross-prediction curves computed on the other types of DNA-binding residues (panel B), the cross-prediction curves computed on the residues that bind proteins, RNA and small molecules (panel C), and the over-prediction curves computed on the non-binding residues (panel D) that were assessed on the test dataset. Predictors are color-coded while the solid, dashed and dotted lines correspond to the predictions of the A-DNA, B-DNA and ssDNA binding residues, respectively.



Supplementary Figure S2. Predictions generated by DNAgenie and selected sequence-based predictors of DNA-binding residues: BindN+, TargetDNA, hybridNAP and DNApred for the human DNA methyltransferase 3A (PDB ID: 5YX2 chain D; UniProt ID: Q9Y6K1). The sequence continues over the three horizontal panels where the color-coded horizonal lines represents the secondary structure collected from the PDB structure; brown, yellow, and red regions represent coil, strand, and helix segments. The native binding residue that were extracted from the structure of the protein-D-DNA complex are shown with black (for B-DNA binding) and gray (for interactions with other ligands) square markers. The predictions are encoded with the round markers, where the solid dark blue circle identify the B-DNA binding predictions from DNAgenie and the hollow pink, orange, light blue and green circles correspond to the DNA type-agnostic predictions from DNApred, TargetDNA, hybridNAP, and BindN+, respectively.