1. Dataset

Supplementary Table 1 Protein chain IDs of datasets (PDB ID_chain)

DBP-374

1A0A_A	1D2I_A	1GTW_A	1IU3_C	1LB2_B	1OZJ_A	1RZ9_A	1XSD_A	2C62_A	2HZV_A
1A35_A	1D5Y_A	1GXP_A	1IV6_A	1LLM_C	1P4E_A	1RZR_L	1Y1W_D	2C6Y_A	2I0Q_B
1A36_A	1D66_A	1H0M_A	1IXY_A	1LQ1_A	1P7D_A	1S40_A	1Y1W_G	2C9L_Y	2I13_A
1A6Y_A	1D8Y_A	1H6F_A	1J1V_A	1LRR_A	1P8K_Z	1SA3_A	1Y6F_A	2CCZ_A	2IIE_A
1A73_A	1DC1_A	1H88_C	1J4W_A	1LWS_A	1PER_L	1SFU_A	1YNW_B	2CGP_A	2ISZ_A
1AKH_A	1DCT_A	1H89_A	1J5N_A	1M06_F	1PH1_B	1SKN_P	1YO5_C	2D45_A	2IX1_A
1AKH_B	1DEW_A	1H8A_C	1J5O_A	1M06_G	1PP8_F	1SKR_B	1YZ8_P	2D5V_A	2JEA_A
1AM9_A	1DIZ_A	1H9D_A	1J75_A	1M07_A	1PUE_E	1SVC_P	1Z1B_A	2D7D_A	2JEA_B
1AN4_A	1DMU_A	1H9T_A	1JB7_A	1M5X_A	1PUF_A	1T2K_D	1Z63_A	2DDG_A	2JEA_I
1AOI_A	1DNK_A	1HAO_H	1JB7_B	1MDM_	1PUF_B	1T38_A	1Z9C_A	2DGC_A	2NLL_A
1AOI_B	1DP7_P	1HBX_A	1JE8_A	1MDY_A	1PV4_A	1TAU_A	1ZGW_A	2DPI_A	2NOB_A
1AOI_D	1DSZ_A	1HBX_G	1JEY_A	1MJE_A	1PYI_A	1TEZ_A	1ZLK_A	2DRP_A	2NP2_A
1AV6_A	1DUX_C	1HHT_P	1JEY_B	1MJE_B	1Q0T_A	1TF3_A	1ZME_C	2DWL_A	2NRA_C
1AWC_A	1ECR_A	1HJB_A	1JFI_A	1MM8_A	1Q9Y_A	1TF6_A	1ZQ3_P	2ER8_A	2NTC_A
1AWC_B	1EMH_A	1HLO_A	1JFI_B	1MNM_	1QAI_A	1TQE_P	1ZS4_A	2ES2_A	2O4I_A
1AZP_A	1EO3_A	1HLV_A	1JMC_A	1MNM_	1QBJ_A	1TRO_A	1ZTG_A	2ETW_A	2OAA_A
1B2M_A	1EQZ_C	1HWT_C	1JNM_A	1MOW_	1QN3_A	1TSR_A	1ZX4_A	2EX5_A	2ODI_A
1B3T_A	1EQZ_D	1I3J_A	1JT0_A	1MSE_C	1QPI_A	1TTU_A	1ZZI_A	2EZV_A	2OFI_A
1B72_A	1EWN_A	1I6H_A	1K61_A	1MTL_A	1QRV_A	1U1K_A	2A07_F	2F8N_K	2OH2_A
1B72_B	1EXI_A	1I6H_B	1K6O_B	1MVM_	1QUM_A	1U3E_M	2A0I_A	2F8X_K	2OST_A
1B8I_A	1EYG_A	1I6H_C	1K78_A	1MW8_X	1QZG_A	1U78_A	2A1R_A	2F8X_M	2OWO_A
1B8I_B	1EYU_A	1I6H_E	1K8G_A	1NFK_A	1R0N_B	1U8B_A	2A3V_A	2FD8_A	2PJR_A
1BDH_A	1F0V_A	1I6H_F	1KB2_A	1NG9_A	1R0O_A	1U8R_A	2A66_A	2FIO_A	2PJR_B
1BF5_A	1F2I_G	1I6H_H	1KBU_A	1NGM_B	1R4I_A	1UBD_C	2A6O_A	2FKC_A	2STT_A
1BG1_A	1F4K_A	1I6H_I	1KC6_A	1NH2_C	1R4O_A	1V14_A	2ACJ_A	2FO1_D	3CRO_L
1BHM_A	1F4S_P	1I6H_J	1KDH_A	1NK2_P	1R71_A	1VAS_A	2AJQ_A	2FO1_E	3HTS_B
1BRN_L	1F5T_A	1I6H_K	1KQQ_A	1NKP_A	1R7M_A	1VFC_A	2AOQ_A	2FQZ_A	3KTQ_A
1BVO_A	1FIU_A	1I6H_L	1KSY_A	1NLW_A	1R8D_A	1VRR_A	2AQ4_A	2G1P_A	3ORC_A
1C7Y_A	1FJL_A	1I7D_A	1KU7_A	1NOP_A	1RC7_A	1W36_B	2ASD_A	2GAT_A	3PJR_A
1C9B_A	1FOK_A	1IAW_A	1L1M_A	1NOY_A	1RC8_A	1W36_C	2AYB_A	2GLI_A	4GAT_A
1CEZ_A	1FOS_E	1IC8_A	1L1T_A	1NWQ_A	1REP_C	1W36_D	2B9S_A	2GXA_A	6PAX_A
1CF7_A	1FZP_B	1ID3_C	1L3S_A	1O4X_A	1RIO_A	1WVL_A	2B9S_B	2GZK_A	10MH_A
1CF7_B	1G38_A	1ID3_D	1L9Z_A	1O4X_B	1RM1_A	1X9N_A	2BGW_A	2H1O_E	
1CIT_A	1G4D_A	1IF1_A	1L9Z_C	1ODG_A	1RM1_B	1XF2_B	2BOP_A	2H27_A	
1CKQ_A	1GCC_A	1IG4_A	1L9Z_D	1ODH_A	1RM1_C	1XHZ_A	2BSQ_A	2H7G_X	
1CMA_A	1GD2_E	1IGN_A	1L9Z_E	1ORN_A	1RRQ_A	1XJV_A	2BSQ_E	2H8C_A	
1CW0_A	1GDT_A	1IHF_A	1L9Z_H	1OSB_A	1RTD_B	1XPX_A	2BZF_A	2HDC_A	
1D02_A	1GM5_A	1IO4_D	1LAU_E	1OUP_A	1RXV_A	1XS9_A	2C5R_A	2HVR_A	

PDNA-62

1A02_F	1BF5_A	1CMA_A	1GCC_A	1HWT_D	1MDY_A	1PAR_B	1REP_C	1UBD_C	2DRP_D
1A02_J	1BHM_A	1D02_A	1GDT_A	1IF1_A	1MEY_F	1PDN_C	1SRS_A	1XBR_A	2GLI_A
1A02_N	1BL0_A	1D66_A	1HCQ_A	1IGN_A	1MHD_A	1PER_L	1SVC_P	1YRN_A	2HDC_A
1A74_A	1C0W_B	1DP7_P	1HCR_A	1IHF_A	1MNM_	1PNR_A	1TC3_C	1YRN_B	3CRO_L
1AAY_A	1CDW_A	1ECR_A	1HDD_C	1IHF_B	1MNM_	1PUE_E	1TF3_A	1YSA_C	
1AZQ_A	1CF7_A	1FJL_A	1HLO_A	1J59_A	1MSE_C	1PVI_B	1TRO_A	1YUI_A	

1B3T_A	1CJG_A	1GAT_A	1HRY_A	1LMB_4	1OCT_C	1PYI_A	1TSR_A	2BOP_A	
TS75									
1A6Y_A	1CIT_A	1EMH_A	1HHT_P	1L9Z_H	1O4X_B	1RXV_A	1YO5_C	2C6Y_A	2PJR_B
1AKH_A	1CKQ_A	1EO3_A	1I6H_E	1LAU_E	1QN3_A	1RZ9_A	1ZGW_A	2EZV_A	3PJR_A
1AKH_B	1CW0_A	1F0V_A	1IAW_A	1M06_G	1QZG_A	1RZR_L	2A1R_A	2FO1_E	4GAT_A
1AOI_B	1D2I_A	1FIU_A	1J4W_A	1MJE_B	1R0N_B	1SKN_P	2A66_A	2GZK_A	
1AOI_D	1DC1_A	1FOK_A	1J5N_A	1NGM_B	1RC8_A	1TQE_P	2ASD_A	2H7G_X	
1AWC_A	1DNK_A	1GD2_E	1JNM_A	1NK2_P	1RM1_B	1U3E_M	2B9S_A	2I0Q_B	
1B2M_A	1DSZ_A	1HAO_H	1JT0_A	1NKP_A	1RM1_C	1Y1W_D	2B9S_B	2JEA_I	
1C9B_A	1DUX_C	1HBX_A	1K61_A	1NLW_A	1RRQ_A	1Y1W_G	2BSQ_E	2OFI_A	

2. Performance comparisons with DISIS.

DISIS (http://cubic.bioc.columbia.edu/services/disis)(Ofran, et al., 2007) predicts DNA-binding sites (6.0 Å is designated as the cutoff distance in the definition of them) in a DNA-binding protein from its amino acid sequence through neural networks (NN) and SVM relying on sequence environment, evolutionary profiles and predicted structural features (secondary structure, solvent accessibility, globularity). Putative DNA-binding residues in the test dataset TS75 were predicted by DISIS with the default values for the optional parameters. We also trained RF models with the exactly same strategy as that for DP-Bind, except that the different cutoff distance (6.0 Å) in the definition of a binding residue. Then, the RF models were used to predict putative DNA-binding residues in the test dataset TS75. The overall accuracy is 81.59% with a low SE 7.65% for the DISIS predictor, and the low SE value may be due to DISIS's more focus on the accuracy of predicted positive samples (Supplementary Table 2). The total accuracy is 78.24% for the RF predictor (Supplementary Table 2).

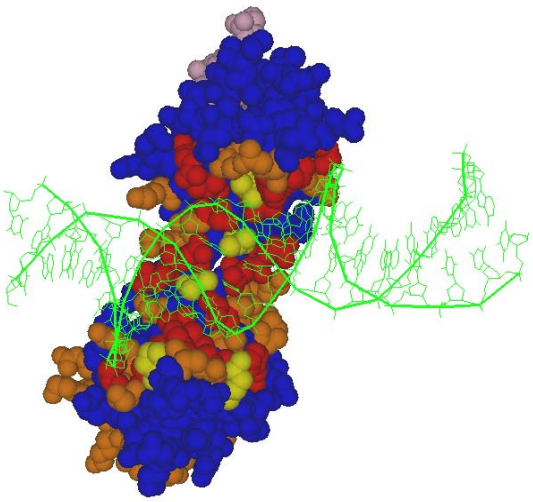
Supplementary Table 2 . Performance comparisons with DISIS.

Classifiers	ACC(%)	SE (%)	PR (%)	SP (%)	MCC
DISIS	81.59	7.65	70.37	99.23	0.190
RF	78.24	51.38	44.25	84.62	0.341

3. Detailed presentation of the prediction of a representative protein-DNA complexe.

To demonstrate that the RF-based model is a useful tool for understanding protein-nucleic acid interactions, we have applied it to predict DNA-binding residues in the archaeal TATA-box-binding protein (TBP) in the structure of TBP/TFB/promoter complex (PDB ID: 1D3U) by visualizing them in the format of three-dimensional structures. The TBP contains 181 residues and was not used for training the RF classifier. Its only homologue in the DBP-374 dataset is the wild-type TATA box-binding protein (TBP) in the structure of TBP–TATA box complex (PDB ID: 1QN3) with 38% sequence identity. The actual DNA-contacting residues were verified by Otis Littlefield et al. (Littlefield, et al., 1999) and these are E12,N13,V15,K37,F43,P44,I47,H49,L58,F60,S62,V66,T68,Q103,N104,V106,F134,P135,R140,V147,L149,F151,S153,V157 and S159. As shown in Supplementary Fig.1., twenty out of twenty-five DNA-binding residues (80.00%) are correctly identified and highlighted in red. The five residues in orange are false negatives (DNA-binding residues but predicted as negatives). For the non-binding residues, 104 of 146 (71.23%) are predicted correctly (residues in blue) and 42

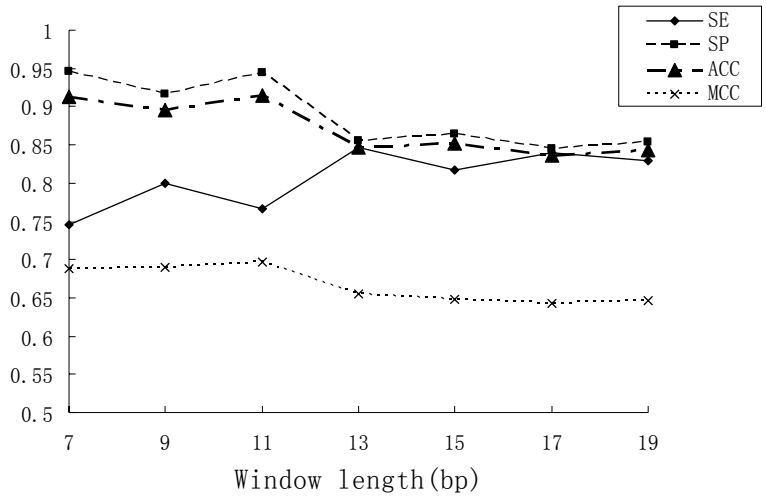
are wrongly predicted (residues in yellow). The total 10 residues located at the N-terminal and C-terminal of the TBP were not used for reporting prediction performance by the RF model and shown in light pink. The results show that DNA-binding residues were predicted by the RF model at 72.51% overall accuracy with Matthew's correlation coefficient (MCC) of 0.377, and with a sensitivity of 80.00% and a specificity of 71.23%. The overall accuracy is 79.56% and 65.14% with the low sensitivity 24.00% and 36.00% for the BindN and Ho et al predictors, and 83.42%, 84.53% and 76.24% for the SVM, KLR and PLR predictors in DP-Bind, respectively.



Supplementary Fig.1. Prediction performance of residues within the archaeal TATA-box-binding protein (TBP) in the structure of TBP/TFB/promoter complex (PDB ID: 1D3U) and its presentation in the format of three-dimensional structures. The correctly identified binding residues (true positives,TPs) are in red space fill; the correctly identified non-binding residues (true negatives,TNs) are in blue space fill; the binding residues with negative predictions (false negatives,FNs) are in orange space fill; the non-binding residues but wrongly predicted as positives (false positives,FPs) are in yellow space fill; the total 10 residues located in the N-terminal and C-terminal of the TBP protein were not used in reporting prediction performance by our model and shown in light pink space fill. The DNA molecule is indicated in green wire frame. The picture is generated with PyMOL (http://www.pymol.org). Prediction performance: ACC 72.51%, SE 80.00%, SP 71.23%, MCC 0.377.

4. Comparison of prediction performances of the different lengths of a data instance.

The length of the structural motifs ranges from 7 to ~20 residues in the protein-DNA complexs. Comparison of prediction performances is studied on considering different lengths of a data instance from 7 to 19 residues (Supplementary Fig.2). Compared with other window sizes, the RF classifiers constructed with $\beta = 11$ presented the best performance.



Supplementary Fig.2. Comparison of prediction performances when considering different lengths of a data instance.