Supporting Information

DNAPred: accurate identification of DNA-binding sites from protein sequence by ensembled hyperplane-distance-based support vector machines

Yi-Heng Zhu [†], *Jun Hu* [‡], *Xiao-Ning Song* [¶] *and Dong-Jun Yu* [†,*]

[†] School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing, 210094, P. R. China, [‡] College of Information Engineering, Zhejiang University of Technology, Hangzhou, 310023, P. R. China, and [¶] School of Internet of Things, Jiangnan University, 1800 Lihu Road, Wuxi, 214122, P. R. China

Supporting Texts

Text S1. The details of CNN-A and CNN-B.

In this work, we design two types of CNN models, i.e., CNN-A and CNN-B, which use the same architecture but different training strategies. As illustrated in Figure S1, the architecture of designed CNN models consists of one input layer, five convolution layers, one fully connected layer, and one output layer.

Input layer. The input layer is a 9×27 feature matrix, which is generated using the following three steps: (1) for each residue in a protein, we extract its corresponding PSSM feature vector (20-D), PSS feature vector (3-D), PRSA feature vector (3-D) and AAFD-BN feature vector (1-D); (2) a 27-D feature vector for each residue is obtained by serially combining its PSSM, PSS, PRSA, and AAFD-BN feature vectors; (3) for a target residue, its corresponding 9×27 feature matrix is generated by using the sliding window technique with size of 9 centered at the residue. In other words, the 27-D feature vector of each residue in the sliding window is a row of the 9×27 feature matrix.

Convolution layers. There are five convolution layers. Each convolution layer (taking the *i*-th layer as an example) works as follows: (1) we use a filter with size of $S_i \times S_i \times N_i$ to execute convolution operation; (2) we use batch normalization technology to normalize the output of convolution; (3) the normalized result is fed to ReLU activation function; (4) we select a window size of $P_i \times P_i$ to execute the max pooling operation on the result of last procedure; (5) the dropout technology with the rate of P_d is adopted to avoid over-fitting.

Fully connected layer. The fully connected layer with N_f neurons is connected with the output of the last convolution layer. Noted that the dropout technology is adopted again after the fully

connected layer.

Output layer. The output layer consists of two nodes and a softmax function is used to compute the final output. For each target residue, the output of the CNN is a 2-D vector, which consists of the predicted probabilities of a sample (residue) separately belonging to the positive class (DNA-binding site) and the negative class (non-binding site). All the hyper-parameters of the CNN models are optimized by ten-fold cross-validation on benchmark datasets.

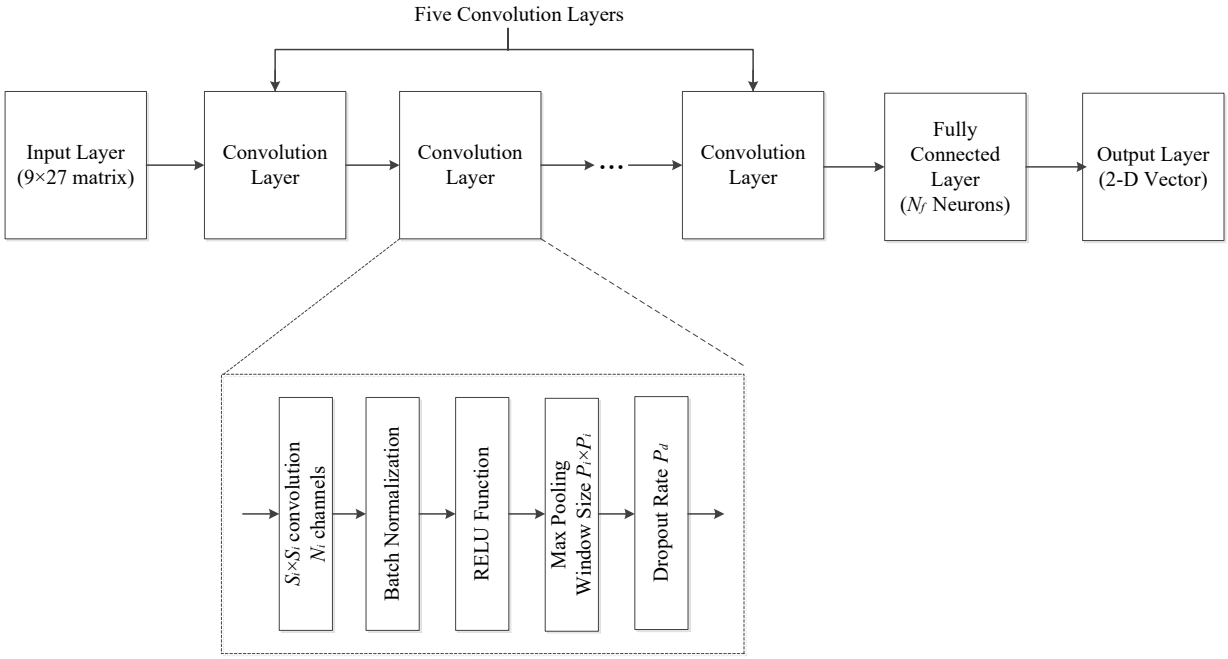

Figure S1. The architecture of the designed CNN models

Training strategies. The batch dataset used for training CNN in each iteration is randomly selected from the original training dataset, including a positive training dataset and a negative training dataset. Due to the severe imbalance of the original training dataset, the batch dataset is also an imbalanced dataset, which may do harm to the performance improvement of CNN. To further investigate the impact of data imbalance to the performance of CNN model, we use two training strategies, denoted as TS-A and TS-B. In TS-A, we randomly select samples from

the original training dataset to construct an imbalanced batch dataset used for training CNN, and the trained CNN model is represented as CNN-A. In TS-B, we separately randomly select same number of samples from the positive training dataset and the negative training dataset to form a balanced batch dataset, and the corresponding trained CNN model is called CNN-B. In this study, we optimize the hyper-parameters of the CNN models by ten-fold cross-validation. The hyper-parameters of two CNN models on four training datasets, i.e., PDNA-543, PDNA-335, PDNA-316, and PDNA-1151, are listed in Table S1.

Table S1. The hyper-parameters of CNN-A and CNN-B on PDNA-543, PDNA-335, PDNA-316, and PDNA-1151 over ten-fold cross-validation.

Dataset	Model	S_1	N_1	P_1	S_2	N_2	P_2	S_3	N_3	P_3	S_4	N_4	P_4	S_5	N_5	P_5	N_f	P_d
PDNA-543	CNN-A	5	512	2	5	256	3	3	512	3	3	512	3	3	256	2	256	0.5
	CNN-B	7	256	2	5	512	4	7	512	2	7	256	4	3	128	2	256	0.5
PDNA-335	CNN-A	5	128	4	5	512	2	5	512	2	3	256	2	3	64	2	256	0.5
	CNN-B	7	256	2	3	256	4	5	256	2	5	64	2	3	128	4	128	0.6
PDNA-316	CNN-A	7	128	2	3	256	2	7	128	2	5	512	4	7	256	2	512	0.6
	CNN-B	5	512	2	5	128	4	3	512	4	3	128	4	3	64	4	256	0.5
PDNA-1151	CNN-A	5	256	4	3	256	2	7	32	4	3	64	2	5	32	4	256	0.6
	CNN-B	5	256	4	3	256	2	7	32	4	3	64	2	5	32	4	256	0.6

Text S2. Analysis of the contributions of different types of features on the independent test datasets.

Table S2 illustrates the performances comparisons of these four combination features, i.e., PSSM (P), PSSM+PSS (PP), PSSM+PSS+PRSA (PPP), and PSSM+PSS+PRSA+AAFD-BN (PPPA), on two independent test datasets, i.e., PDNA-41 and PDNA-52. From Table S2, the same phenomena with Table 4 in the manuscript can be found: (1) the PSSM feature is very useful to predict the protein-DNA binding sites; (2) PSS, PRSA, and AAFD-BN are also beneficial for predicting the

protein-DNA binding sites. Taking PDNA-52 as an example, the *MCC* and *AUC* values of E-HDSVM with the PSSM feature reach 0.371 and 0.858, respectively, which are not far below those of the E-HDSVM with the PPPA feature; moreover, the *MCC* value of E-HDSVM is improved from 0.371 to 0.384, 0.393, and 0.405 after gradually adding PSS, PRSA, and AAFD-BN to the PSSM feature. In addition, it has not escaped our notice that the *MCC* value of E-HDSVM using the PPPA is slightly lower than that of the E-HDSVM using the PPP on PDNA-41. However, the model with the PPPA approximately still obtains 0.8% improvement of *AUC* value compared with the model with the PPP.

Table S2. Performances comparisons of four combination features on two independent test datasets.

Dataset	Feature	*Sen* (%)	*Spe* (%)	*Acc* (%)	*MCC*	*AUC*
PDNA-41 [a]	P	42.4	94.1	91.5	0.297	0.826
	PP	**46.7**	94.0	91.6	0.325	0.838
	PPP	43.1	**95.6**	**93.0**	**0.345**	0.847
	PPPA	44.0	95.2	92.7	0.340	**0.851**
PDNA-52 [b]	P	53.0	93.4	91.1	0.371	0.858
	PP	**55.4**	93.3	91.2	0.384	0.868
	PPP	52.4	94.4	92.0	0.393	0.872
	PPPA	51.8	**94.9**	**92.5**	**0.405**	**0.876**

[a] The corresponding training dataset is PDNA-543.

[b] The corresponding training dataset is PDNA-335.