Supplementary Information

LBi-DBP, an accurate DNA-binding protein prediction method based lightweight interpretable BiLSTM network

Wenwu Zeng[1,†], Xuan Yu [2,†], Wenjuan Liu[1,*], Jun Hu[3,*], and Shaoliang Peng[1,*]

[1] College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China;

[2] Glasgow College, University of Electronic Science and Technology of China, Chengdu, 611731, China;

[3] College of Information Engineering, Zhejiang University of Technology, Hangzhou, 310023, China

[*] Address correspondence to W.J. Liu at liuwenjuan89@hnu.edu.cn or J. Hu at hujunum@zjut.edu.cn or S.L. Peng at slpeng@hnu.edu.cn

[†] The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint Authors.

Supporting Texts

Text S1. Feature Source

Position-Specific Scoring Matrix

For each protein contained L residues, the PSSM profile is first generated by feeding protein sequence into PSI-BLAST [1] after three rounds of iterative search with 0.001 as E-value cutoff for multiple sequence alignment against the non-redundant protein sequence database [2]; then, each element (x) in PSSM is normalized by the logistic function, i.e., $f(x)=1/(1+e^{-x})$; finally, a feature matrix of size $L \times 20$ can be obtained.

Hidden Markov Model Profile

It has been demonstrated that HMM generated by HHblits [3] is complementary to PSSM for representing the evolutionary information [4]. In this study, for each protein contained L residues, its sequence is fed into HHblits to generate the raw profile; then, each element in the raw profile is normalized in turn by two normalization functions, i.e., $g(x)$=$2^{-x/1000}$ and $f(x)$=$1/(1+e^{-x})$; finally, the HMM profile of size $L \times 30$ is generated.

Predicted Secondary Structure Probability Matrix

For one protein with L residues, its PSSPM profile, which contains L rows and three columns, is predicted by PSIPRED [5]. Each row in PSSPM contains three probability values of belonging to three SS classes, i.e., coil, helix, and strand, of the corresponding residue.

Predicted Solvent Accessibility Probability Matrix

For each protein with L residues, the PSAPM predicted by the SANN [6] program contains L rows. Each row includes three elements, which represent the probabilities of belonging to the classes of buried, intermediate, and exposed of the corresponding residue.

Predicted Probabilities of DNA-Binding Sites

Theoretically, if all DNA-binding sites (DBSs) in proteins can be accurately identified, the DBP prediction will degenerate into an easy task. Unluckily, the DBS prediction stays a big room for improvements. However, the PPDBS results could be employed to act as a feature view and extract the discriminative feature representation. To fairly use this feature view, a new DBS prediction model is trained on the training data set of TargetDBP [7] after removing all non-DBPs and those DBPs that have a sequence identity larger than 25% with at least one protein in *UniSwiss-Tst*. This model employs PSSM and PSAPM as input features. For each protein, its PPDBS profile of size L×2 can be generated using the above model.

Text S2. Pseudo feature extraction

For each vector $\boldsymbol{u} = (u_1, u_2, \dots, u_L)^T$, e.g., the l-th column vector of PSSM ($1 \leq l \leq 20$), the corresponding SO feature vector $\boldsymbol{o} = (o_1, o_2, \dots, o_G)^T$ could be obtained by calculating the correlation factors of $\boldsymbol{u}$. The g-tier correlation factor (o_g) of $\boldsymbol{u}$ is calculated by:

$$o_g = \frac{1}{L-g} \sum_{i=1}^{L-g} \left(u_i - u_{i+g}\right)^2 \tag{1}$$

where $1 \leq g \leq G$ and $G < L$, G is a hyperparameter that needs to be adjusted. In this study, G is set to 18 according to our previous study [8]. For a feature source of size $L \times D$, where L is the protein length and D is the feature dimension of each residue, a pseudo SO feature of size $18 \times D$ could be easily generated according to the above steps. Based on the above steps, the SO information embedding in PSSM, HMM, PSSPM, PSAPM and PPDBS are extracted. The corresponding features are named as PsePSSM, PseHMM, PsePSSPM, PsePSAPM and PsePPDBS with sizes of 18×20, 18×30, 18×3, 18×3, and 18×2, respectively.

Text S3. Evaluation indices

To evaluate the effectiveness of the proposed LBi-DBP, six evaluation indexes, i.e., sensitivity (Sen), specificity (Spe), accuracy (Acc), precision (Pre), Matthew's correlation coefficient (MCC), and F_1-score (F_1), are utilized. The formulas of these six evaluation indexes are as follows:

$$Sen = \frac{TP}{TP+FN} \times 100 \tag{2}$$

$$Spe = \frac{TN}{TN+FP} \times 100 \tag{3}$$

$$Acc = \frac{TN+TP}{TN+TP+FN+FP} \times 100 \tag{4}$$

$$Pre = \frac{TP}{TP+FP} \times 100 \tag{5}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FN) \cdot (TP+FP) \cdot (TN+FN) \cdot (TN+FP)}} \tag{6}$$

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP+FN+FP} \tag{7}$$

where true positive (TP) and false positive (FP) are the numbers of proteins that are correctly and mistakenly predicted as DBP, and true negative (TN) and false negative (FN) are the numbers of proteins that are correctly and mistakenly predicted as non-DBP, respectively. Besides the above six indexes, the receiver operating characteristic curve (ROC) and the area under ROC (AUC) are also utilized to further evaluate the overall predictive performance of LBi-DBP.

Text S4. Performance Comparison of Sequence-based Feature Sources

The performance of six sequence-based feature sources, i.e., PSSM, HMM, PSSPM, PSAPM, PPDBS and the combination of them, are investigated on *UniSwiss-Tr* over a

Table S1. Performance comparison of PSSM, HMM, PSSPM, PSAPM, PPDBS and their combination on *UniSwiss-Tr* over 10-fold cross-validation test

Feature	Sen	Spe	Acc	Pre	F_1	MCC	AUC	*p*-value
PSSM	84.47	75.93	80.20	77.82	0.810	0.606	0.881	4.51e-06
HMM	**85.33**	59.62	72.47	67.88	0.756	0.465	0.790	1.91e-15
PSSPM	70.84	64.22	67.53	66.44	0.686	0.351	0.737	5.05e-18
PSAPM	82.15	61.89	72.02	68.31	0.746	0.450	0.778	2.44e-05
PPDBS	81.89	52.60	67.24	63.34	0.714	0.361	0.744	4.96e-06
Combination	80.82	**85.42**	**83.12**	**84.72**	**0.827**	**0.663**	**0.903**	**-**

"Combination" means the combination of HMM, PSSPM, PSAPM, PSSPM and PPDBS.
The *p*-values in Student's t-test are calculated for the differences between combination feature and other features.
The bolded font indicates the highest result.

10-fold cross-validation test. It is noted that only the module in Figure 1 is employed to train the prediction model in this section. Table S1 demonstrates the performance comparison of the above six feature sources.

From Table S1, it is easy to find that the MCC values of all six sequence-based feature sources are larger than 0.35, which means they can give a positive impact on the prediction of DBPs. Among the six feature sources, the combination one gains the highest Spe, ACC, Pre, F_1, MCC and AUC values of 85.42, 83.12, 84.72, 0.827, 0.663 and 0.903, which are 12.50%, 3.64%, 8.86%, 2.10%, 9.40% and 2.61% higher than the second-best PSSM, respectively. The difference in the predicted probability values between the combination feature and PSSM is statistically significant which has a *p*-value $<10^{-5}$ in the Student's t-test. The above data demonstrates that there are complementary information embedding in the five sequence-based feature sources, i.e., PSSM, HMM, PSSPM, PSAPM and PPDBS.

Figure S1 demonstrates the ROC curves of the six feature sources on *UniSwiss-Tr* over a 10-fold cross-validation test. It is obviously found that the true positive rate (TPR) of the combination feature is consistently higher than that of the other five feature sources, i.e., PSSM, HMM, PSSPM, PSAPM and PPDBS, when the false positive rate (FPR) is less than 0.37. When FPR>0.37, the TPR of the combination feature source is still higher than that of HMM, PSSPM, PSAPM and PPDBS and comparable to that of PSSM.

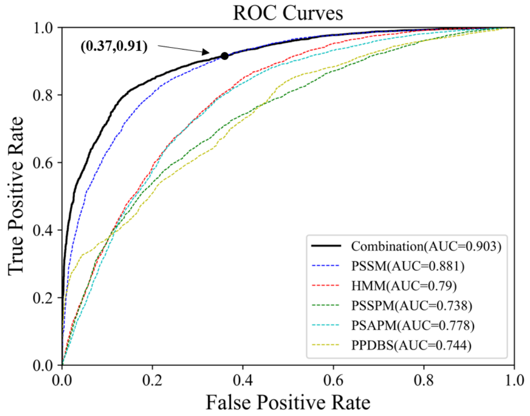

Figure S1. ROC curves of different sequence features on BiLSTM-based on *UniSwiss-Tr* over 10-fold cross-validation.

Text S5. Performance Comparison of Pseudo Sequence Order Features

To evaluate the performance of five pseudo SO features, i.e., PsePSSM, PseHMM, PsePSSPM, PsePSAPM and PsePPDBS, and their combination, the MLP module in Figure 2 is employed to train the prediction model in this section. Table S2 lists the performance of the above six pseudo SO features on *UniSwiss-Tr* over a 10-fold cross-validation test.

From Table S2, we can find that PsePSSM achieves the best performance over five pseudo features in this study. The MCC value of PsePSSM is 0.616, which is 24.19, 54.00, 20.55 and 28.06 percent higher than PseHMM, PsePSSPM, PsePSAPM and PsePPDBS, respectively. The values of Spe, ACC, Pre, F_1, MCC and AUC of the combination feature are 87.40, 84.71, 86.68, 0.843, 0.695 and 0.925, which are 10.27, 4.82, 8.51, 3.94, 12.82 and 3.93 percent higher than that of PsePSSM. Figure S2 also demonstrates the ROC curves of five pseudo SO features, i.e., PsePSSM, PseHMM, PsePSSPM, PsePSAPM and PsePPDBS, and their combination on *UniSwiss-Tr* over 10-fold cross-validation test.

Table S2. Performance comparison of PsePSSM, PseHMM, PsePSSPM, PsePSAPM and PsePPDBS, and their combination on *UniSwiss-Tr* over 10-fold cross-validation test

Feature	Sen	Spe	Acc	Pre	F_1	MCC	AUC	*p*-value
PsePSSM	82.35	79.26	80.81	79.88	0.811	0.616	0.890	1.50e-01
PseHMM	**83.62**	64.53	74.07	70.22	0.763	0.496	0.817	5.18e-20
PsePSSPM	70.60	69.31	69.95	69.70	0.701	0.400	0.749	9.51e-16
PsePSAPM	75.91	75.15	75.53	75.34	0.756	0.511	0.833	5.15e-01
PsePPDBS	62.22	84.69	73.45	80.25	0.701	0.481	0.795	3.72e-12
Combination	82.02	**87.40**	**84.71**	**86.68**	**0.843**	**0.695**	**0.925**	-

"Combination" means the combination of PsePSSM, PseHMM, PsePSSPM, PsePSAPM and PsePPDBS.
The *p*-values in Student's t-test are calculated for the differences between combination feature and other features.
The bolded font indicates the highest result.

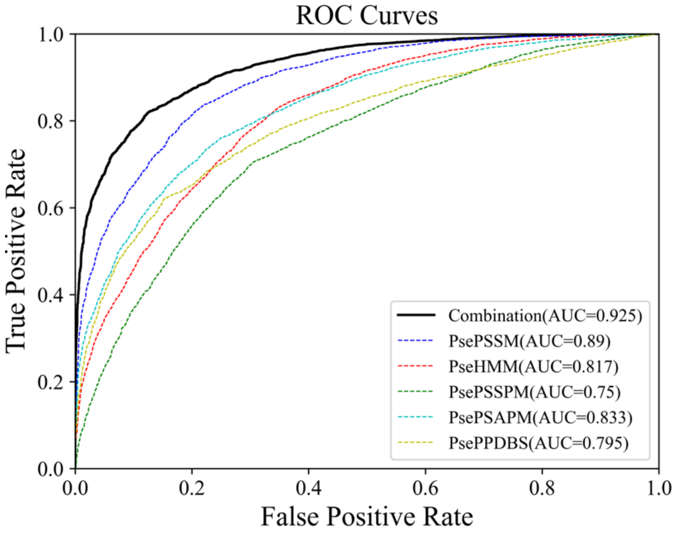

Figure S2. ROC curves of five pseudo SO features, i.e., PsePSSM, PseHMM, PsePSSPM, PsePSAPM and PsePPDBS, and their combination on *UniSwiss-Tr* over 10-fold cross-validation test.

Text S6. Architecture of LBi-DBP-single and LBi-DBP-together

Figures S3 and S4 show the structures of LBi-DBP-single and LBi-DBP-together, respectively. Unlike LBi-DBP, these two models do not use an independent module to extract SCI feature from the sequence feature source, but feed the sequence feature and SO feature into the model together for training. In LBi-DBP-single, the output of BiLSTM layer is directly combined with SO feature and then train the DBP identification model based on MLP. In LBi-DBP-together, the output of BiLSTM layer and SO feature are first fed into two MLP modules respectively; then, the output of these two MLP modules are combined and input to final MLP module.

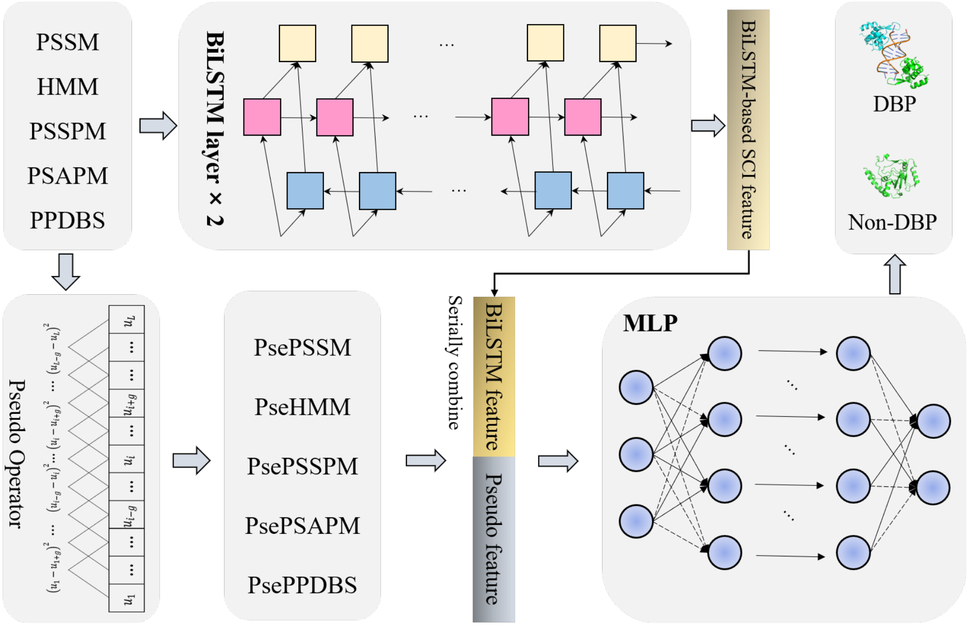

Figure S3. Structure of LBi-DBP-single.

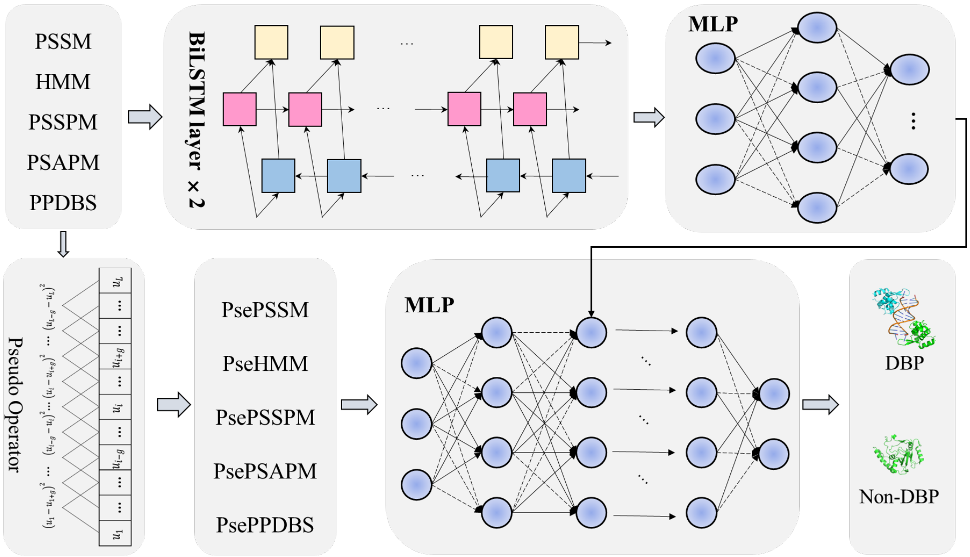

Figure S4. Structure of LBi-DBP-together.

Text S7. Performance comparison with other method on *UniSwiss-Tst* using PDB148, PDB424, and PDB1075 as training set, respectively

Table S3 shows the comparison results of LBi-DBP and existing DBP prediction methods on *UniSwiss-Tst* using PDB148, PDB424, and PDB1075 as training data set, respectively. The prediction results of these control methods are reported in previous studies [9] [8]. From Table S3, we can know that, LBi-DBP achieves the highest MCC value when using PDB424 as the training set and the second highest MCC when using PDB1075 and PDB148 as the training set.

Table S3. Performance Comparison of LBi-DBP and State-of-the-art DBP Prediction Methods on *UniSwiss-Tst* using PDB148, PDB424, and PDB1075 as training set, respectively

PDB148 [a]	IKP-DBPPred	52.91	56.58	54.75	54.79	0.538	0.095
	TargetDBP+	34.38	74.80	54.59	57.71	0.431	0.100
	TPSO-DBP	**62.99** [e]	62.20	**62.60**	**62.50**	0.627	**0.252**
	LBi-DBP	35.43	**77.16**	56.30	60.81	**0.631**	0.138
PDB424 [b]	iDNA-Prot	48.56	51.97	50.26	50.27	0.494	0.005
	TargetDBP+	84.25	34.91	59.58	56.41	0.676	0.220
	TPSO-DBP	52.75	**75.06**	63.91	**67.91**	0.594	0.285
	LBi-DBP	**84.51**	43.83	**64.17**	60.07	**0.691**	**0.310**
PDB1075 [c]	PSFM-DBT	87.30	48.35	67.28	61.52	**0.722**	0.385
	PseDNA-Pro	74.28	41.21	57.74	55.82	0.637	0.164
	Local-DPP	13.53	**92.89**	53.37	65.38	0.224	0.106
	iDNAPro-PseAAC	64.55	32.63	48.55	48.80	0.556	-0.030
	HMMBinder	**99.74**	2.36	51.05	50.53	0.671	0.092
	DPP-PseAAC	54.59	55.91	55.25	55.32	0.620	0.120
	iDNA-Prot|dis	72.44	38.85	55.64	54.22	0.620	0.120
	TargetDBP+	66.40	73.23	69.82	**71.27**	0.688	0.397
	TPSO-DBP	70.08	70.34	**70.21**	70.26	0.702	**0.404**
	LBi-DBP	81.10	49.08	65.09	61.43	0.715	0.318

a. PDB148 contains 74 DBPs and 74 non-DBPs, which are randomly selected from PDB186, as the training data set of IKP-DBPPred is not given.

b. PDB424 is the data set used to train the prediction model of iDNA-Prot.

c. PDB1075 is the data set used to train the prediction models of PSFM-DBT, PseDNA-Pro, Local-DPP, iDNAPro-PseAAC, HMMBinder, DPP-PseAAC, and iDNA-Prot|dis.

Text S8. Performance comparison with other method on PDB186

Table S4 shows the performance comparison of LBi-DBP trained on PDB1075 and state-of-the-art DBP prediction methods on PDB186. The prediction results of these control methods are reported in previous study [9]. By visiting Table S4, we can find that out of these 20 methods, the performance of LBi-DBP is relatively good. The MCC value of LBi-DBP is 0.735 and ranks fourth. It is worth noting that the MCC value of the best method MLapSVA-LBS is 0.760, which is 3.40% higher than LBi-DBP. Since the performance of neural networks relies heavily on large datasets, the potential reason of this phenomenon is that the amount of protein in PDB1075 is too low, leading model to be prone to overfitting. On the contrary, as a machine learning-based method, MLapSVM-LBS is relatively not so easy to overfit.

Table S4. Performance Comparison among LBi-DBP trained on PDB1075 and State-of-the-art DBP Prediction Methods on PDB186

Method	Acc	Sen	Spe	MCC	AUC
iDNA-Prot|dis [10]	72.00	79.50	64.50	0.450	0.786
iDNAPro-PseAAC [11]	71.50	82.80	60.20	0.440	0.778
HMMBinder [4]	69.00	61.50	76.30	0.394	0.632
DBPPred [12]	76.90	79.60	74.20	0.538	0.791
iDNAProt-ES [13]	80.64	81.31	80.00	0.613	- [a]
Local-DPP [14]	79.00	92.50	65.60	0.625	-
MKSVM-HKA [15]	81.20	94.60	67.70	0.650	0.887
FKRR-MVSF [16]	81.70	**98.90** [b]	64.50	0.680	0.901
DPP-PseACC [17]	77.40	83.90	71.00	0.550	0.799
PseDNA-Pro [18]	71.50	82.80	60.20	0.240	-
MSFBinder [19]	81.70	89.30	74.20	0.640	-
MKL-HSIC with H-LapSVM [20]	87.10	91.70	82.80	0.750	0.931
MKSVM with MKL-CKA [21]	83.70	93.60	74.20	0.690	0.899
MsDBP [22]	80.10	86.00	74.20	0.610	0.875
KK-DBP [23]	81.20	97.80	64.50	0.660	-
StackPDB [24]	84.41	83.87	84.95	0.690	-
MLapSVM-LBS [25]	**88.70**	90.30	**87.00**	**0.760**	**0.957**
TPSO-DBP [9]	87.16	94.59	79.73	0.752	0.907
LBi-DBP	83.33	91.40	81.72	0.735	0.882

a. "-" indicates that the value is not available.

b. The bolded font indicates the highest result.

Table S5. Max sequence identities of disDNA-TEST140 against to the training data set, i.e., *UniSwiss-Tr*, calculated by NWAlign tool

DBP	Sequence identity	Non-DBP	Sequence identity
O00470	0.352	A0A1B0GTW7	0.228
O09185	0.259	A1A4Y4	0.243
O14770	0.371	A1L167	0.261
O96028	0.22	A2PYH4	0.288
P02340	0.256	A2RU14	0.263
P02751	0.208	A2VDN5	0.335
P04150	0.364	A4D126	0.231
P04637	0.254	A4D1T9	0.243
P11308	0.954	A5D8V7	0.23
P11473	0.251	A5D8W1	0.217
P13481	0.257	A5YKK6	0.265
P17096	0.303	A6H8Y1	0.215
P21675	0.451	A6NFY7	0.28
P27694	0.272	A6NGG8	0.226
P28482	0.244	A6NI61	0.244
P35398	0.486	A6NK58	0.283
P35680	0.53	A8K2U0	0.211
P35711	0.581	A8MTZ0	0.261
P37275	0.452	A8MXD5	0.255
P51608	0.233	A8TX70	0.196
P52926	0.294	B0YJ81	0.237
P56423	0.256	B1AK53	0.232
P56424	0.256	B2RPY5	0.229
P61260	0.256	B2RTY4	0.246
P98177	0.376	B2RXF5	0.451
Q00366	0.254	B2RY04	0.206
Q02556	0.317	B5SY89	0.227
Q14653	0.299	B7U540	0.226
Q60641	0.273	C9JE40	0.232
Q92731	0.458	C9JR72	0.258
Q95330	0.251	D3ZQF4	0.254
Q9NR48	0.203	D6RGH6	0.252
Q9UMN6	0.295	E2RDM9	0.253
Q9UN79	0.531	E2RDP2	0.25
Q9Y5R6	0.272	E9PY46	0.212
Q9Y6Y1	0.416	Q8IV33	0.209
O14529	0.466	Q8NDG6	0.252
O15353	0.243	Q9C091	0.521
O35160	0.331	Q86YR7	0.21
O43316	0.31	Q5RHB5	0.218
O43364	0.465	E2R1I5	0.218
O75364	0.334	Q8IZD2	0.206
O95718	0.279	A0A1W2PR82	0.255
P06798	0.361	Q969U7	0.254
P20264	0.348	O95267	0.217
P26367	0.354	Q9P021	0.257
P35716	0.443	Q3ZAQ7	0.324

P41235	0.302	Q6NUK1	0.258
P51448	0.489	Q8IYB7	0.313
P63015	0.354	Q9BXB7	0.215
P63016	0.336	Q6UWS5	0.284
P78337	0.337	Q86XA0	0.268
Q01851	0.362	Q2M385	0.219
Q04743	0.341	Q6ZUV0	0.234
Q13285	0.274	Q9BVQ7	0.24
Q15306	0.322	O14654	0.223
Q1HGE8	0.389	B0FYL5	0.23
Q60954	0.352	Q6PJT7	0.228
Q61575	0.247	F1PZQ5	0.213
Q86UP3	0.55	Q96RN1	0.207
Q8N635	0.251	O95996	0.35
Q92753	0.473	Q8NB90	0.253
Q96RI1	0.289	Q8NDX2	0.218
Q99JB6	0.242	Q5HYJ1	0.229
Q9BZS1	0.253	Q7LDG7	0.225
Q9C0A1	0.285	Q6V0I7	0.193
Q9ULV5	0.395	Q6NXT6	0.336
Q9UMR3	0.346	Q96QF7	0.217
Q9Y5X4	0.285	Q6ZMP0	0.219
B5RHS5	0.479	P0DPH8	0.301

REFERENCES

[1] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. J. N. a. r. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," vol. 25, no. 17, pp. 3389-3402, 1997.

[2] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic acids research,* vol. 35, no. suppl_1, pp. D61-D65, 2007.

[3] M. Remmert, A. Biegert, A. Hauser, and J. Söding, "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment," *Nature methods,* vol. 9, no. 2, pp. 173-175, 2012.

[4] R. Zaman, S. Y. Chowdhury, M. A. Rashid, A. Sharma, A. Dehzangi, and S. Shatabda, "Hmmbinder: Dna-binding protein prediction using hmm profile based features," *BioMed research international,* vol. 2017, pp. 4590609, 2017.

[5] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *Journal of Molecular Biology,* vol. 292, no. 2, pp. 195-202, 1999.

[6] K. Joo, S. J. Lee, and J. Lee, "Sann: solvent accessibility prediction of proteins by nearest neighbor method," *Proteins: Structure, Function, and Bioinformatics,* vol. 80, no. 7, pp. 1791-1797, 2012.

[7] J. Hu, X.-G. Zhou, Y.-H. Zhu, D.-J. Yu, and G.-J. Zhang, "TargetDBP: accurate DNA-binding protein prediction via sequence-based multi-view feature learning," *IEEE/ACM transactions on computational biology and bioinformatics,* vol. 17, no. 4, pp. 1419-1429, 2019.

[8] J. Hu, L. Rao, Y. H. Zhu, G. J. Zhang, and D. J. Yu, "TargetDBP+ : Enhancing the Performance of Identifying DNA-Binding Proteins via Weighted Convolutional Features," *Journal of Chemical Information and Modeling,* vol. 61, no. 1, pp. 505-515, Jan, 2021.

[9] J. Hu, W. W. Zeng, N. X. Jia, M. Arif, D. J. Yu, and G. J. Zhang, "Improving DNA-Binding Protein Prediction Using Three-Part Sequence-Order Feature Extraction and a Deep Neural Network Algorithm," *Journal of Chemical Information and Modeling,* vol. 63, no. 3, pp. 1044-1057, Feb, 2023.

[10] B. Liu, J. Xu, X. Lan, R. Xu, J. Zhou, X. Wang, and K.-C. Chou, "iDNA-Prot| dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition," *PloS one,* vol. 9, no. 9, pp. e106691, 2014.

[11] B. Liu, S. Wang, and X. Wang, "DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation," *Scientific reports,* vol. 5, pp. 15479, 2015.

[12] W. Lou, X. Wang, F. Chen, Y. Chen, B. Jiang, and H. Zhang, "Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes," *PloS one,* vol. 9, no. 1, pp. e86703, 2014.

[13] S. Y. Chowdhury, S. Shatabda, and A. Dehzangi, "iDNAProt-ES: Identification of DNA-binding proteins using evolutionary and structural features," *Scientific reports,* vol. 7, pp. 14938, 2017.

[14] L. Wei, J. Tang, and Q. Zou, "Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information," *Information Sciences,* vol. 384, pp. 135-144, 2017.

[15] Y. Ding, F. Chen, X. Guo, J. Tang, and H. Wu, "Identification of DNA-binding proteins by multiple kernel support vector machine and sequence information," *Current Proteomics,* vol. 17, no. 4, pp. 302-310, 2020.

[16] Y. Zou, Y. Ding, J. Tang, F. Guo, and L. Peng, "FKRR-MVSF: a fuzzy kernel ridge regression model for identifying DNA-binding proteins by multi-view sequence features via Chou's five-step rule," *International Journal of Molecular Sciences,* vol. 20, no. 17, pp. 4175, 2019.

[17] M. S. Rahman, S. Shatabda, S. Saha, M. Kaykobad, and M. S. Rahman, "Dpp-pseaac: A dna-binding protein prediction model using chou's general pseaac," *Journal of theoretical biology,* vol. 452, pp. 22-34, 2018.

[18] B. Liu, J. Xu, S. Fan, R. Xu, J. Zhou, and X. Wang, "PseDNA-Pro: DNA-binding protein identification by combining Chou's PseAAC and physicochemical distance transformation," *Molecular Informatics,* vol. 34, no. 1, pp. 8-17, 2015.

[19] X.-J. Liu, X.-J. Gong, H. Yu, and J.-H. Xu, "A model stacking framework for identifying DNA binding proteins by orchestrating multi-view features and classifiers," *Genes,* vol. 9, no. 8, pp. 394, 2018.

[20] Y. Qian, H. Meng, W. Lu, Z. Liao, Y. Ding, and H. Wu, "Identification of DNA-binding proteins via hypergraph based laplacian support vector machine," *Current Bioinformatics,* vol. 17, no. 1, pp. 108-117, 2022.

[21] Y. Qian, L. Jiang, Y. Ding, J. Tang, and F. Guo, "A sequence-based multiple kernel model for identifying DNA-binding proteins," *BMC bioinformatics,* vol. 22, no. 3, pp. 1-18, 2021.

[22] X. Du, Y. Diao, H. Liu, and S. Li, "MsDBP: exploring DNA-binding proteins by integrating multiscale sequence information via Chou's five-step rule," *Journal of Proteome Research,* vol. 18, no. 8, pp. 3119-3132, 2019.

[23] Y. Jia, S. Huang, and T. Zhang, "KK-DBP: A multi-feature fusion method for DNA-binding protein identification based on random forest," *Frontiers in Genetics,* vol. 12, pp. 811158, 2021.

[24] Q. Zhang, P. Liu, X. Wang, Y. Zhang, Y. Han, and B. Yu, "StackPDB: predicting DNA-binding proteins based on XGB-RFE feature optimization and stacked ensemble classifier," *Applied Soft Computing,* vol. 99, pp. 106921, 2021.

[25] M. W. Sun, P. Tiwari, Y. Q. Qian, Y. J. Ding, and Q. Zou, "MLapSVM-LBS: Predicting DNA-binding proteins via a multiple Laplacian regularized support vector machine with local behavior similarity," *Knowledge-Based Systems,* vol. 250, pp. 8, Aug, 2022.