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;

[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. Architecture of bidirectional long short-term memory

BiLSTM is the bidirectional Long Short-Term Memory (LSTM) (1), which has a great ability to capture the short-range and long-range information of sequence data. As shown in Figure S1(A), the formulas of each LSTM unit can be expressed as follow:

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + B_i) \tag{S1}$$

$$f_t = \sigma\left(W_f \times [h_{t-1}, x_t] + B_f\right) \tag{S2}$$

$$\widetilde{c_t} = tanh(W_c \times [h_{t-1}, x_t] + B_c) \tag{S3}$$

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c_t} \tag{S4}$$

$$o_t = \sigma(W_O \times [h_{t-1}, x_t] + B_o) \tag{S5}$$

$$h_t = o_t * tanh(c_t) \tag{S6}$$

where input gate, forget gate and output gate are denoted as i_t, f_t and o_t, respectively. c_t and $\widetilde{c_t}$ represent the new memory cell and the final memory cell, respectively. h_{t-1} and h_t mean the hidden state vector at the position $t-1$ and position t. W_i, W_f, W_c and W_o are weight matrices that wait to be updated in the train process. B_i, B_f, B_c and B_o are bias vectors. $\sigma(\cdot)$ is the sigmoid function. As shown in Figure S1(B), BiLSTM consists of two layers of unidirectional LSTM, i.e., backward and forward LSTM. Compared with LSTM, BiLSTM can better capture the sequence dependence in both directions.

Text S2. Architecture of multi-layer perceptron

Multi-layer perceptron (MLP) is a classical deep learning network with strong nonlinear fitting ability. As shown in Figure S2, the formula can be expressed as follow:

$$X_{i+1} = \delta(W_i * X_i + B_i) \tag{S7}$$

where X_i is the input vector of i-th MLP layer. W_i is the weight matrix to be learned. B_i is the bias vector. $\delta(\cdot)$ is the activation function. X_{i+1} is the input of $(i+1)$-th MLP layer.

Text S3. Architecture of LBi-DBP-single and LBi-DBP-together

Figure 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.

Table S1. 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

Figure S1. Architecture of LSTM unit and BiLSTM. (A) LSTM unit; (B) BiLSTM.

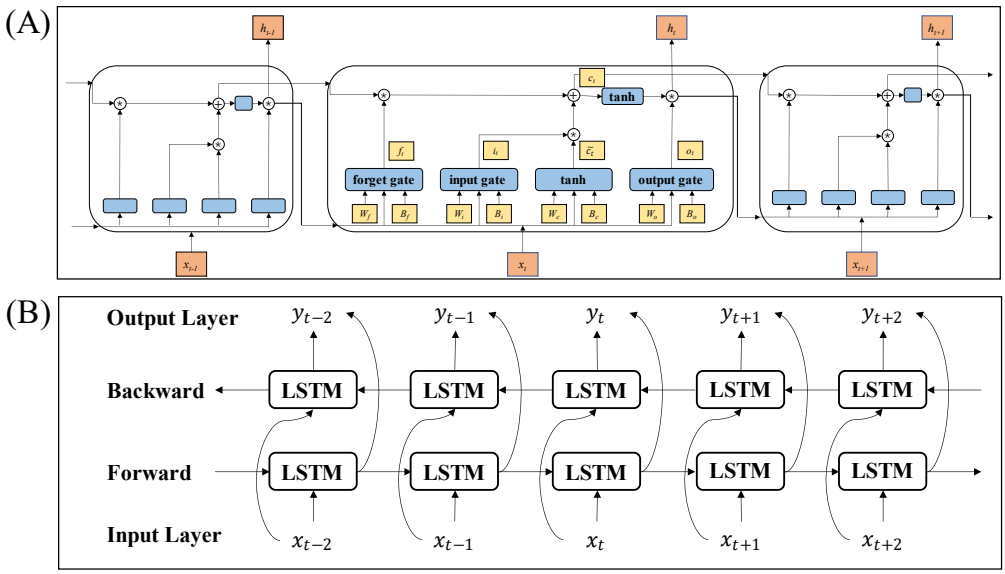


Figure S2. Architecture of multi-layer perceptron

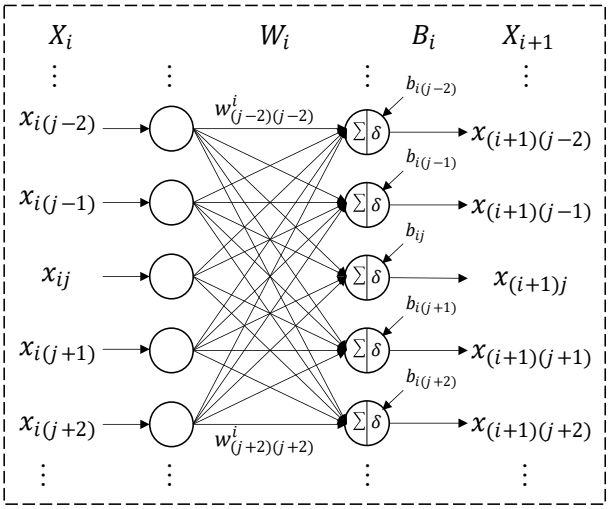


Figure S3. Structure of LBi-DBP-single.

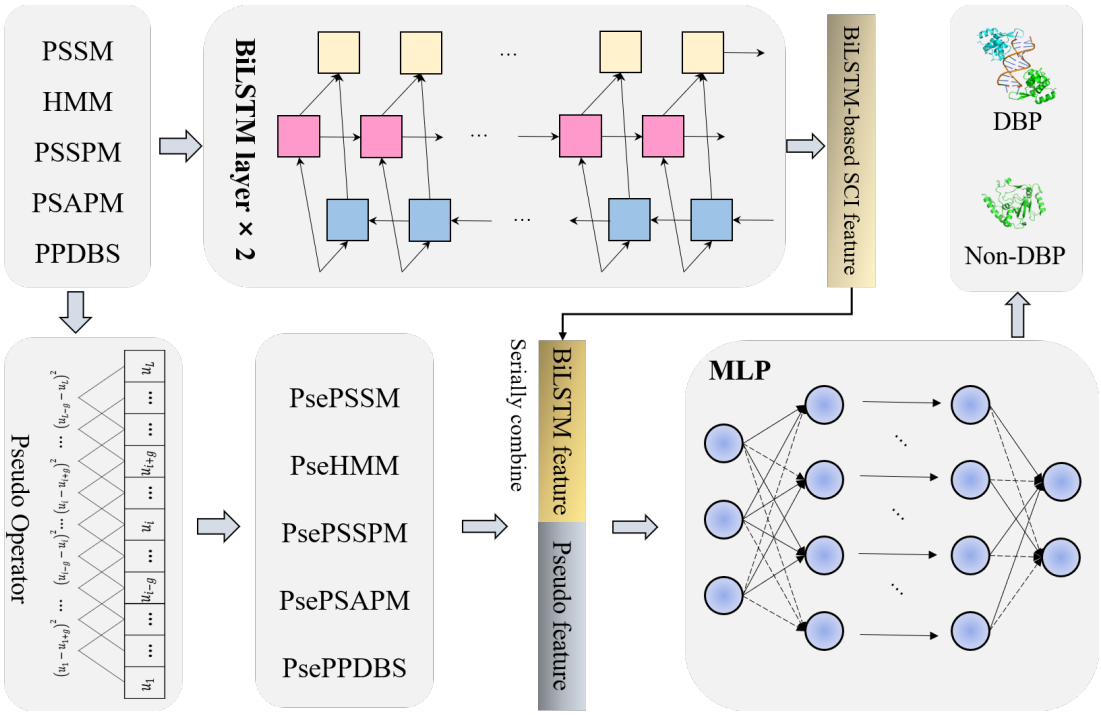
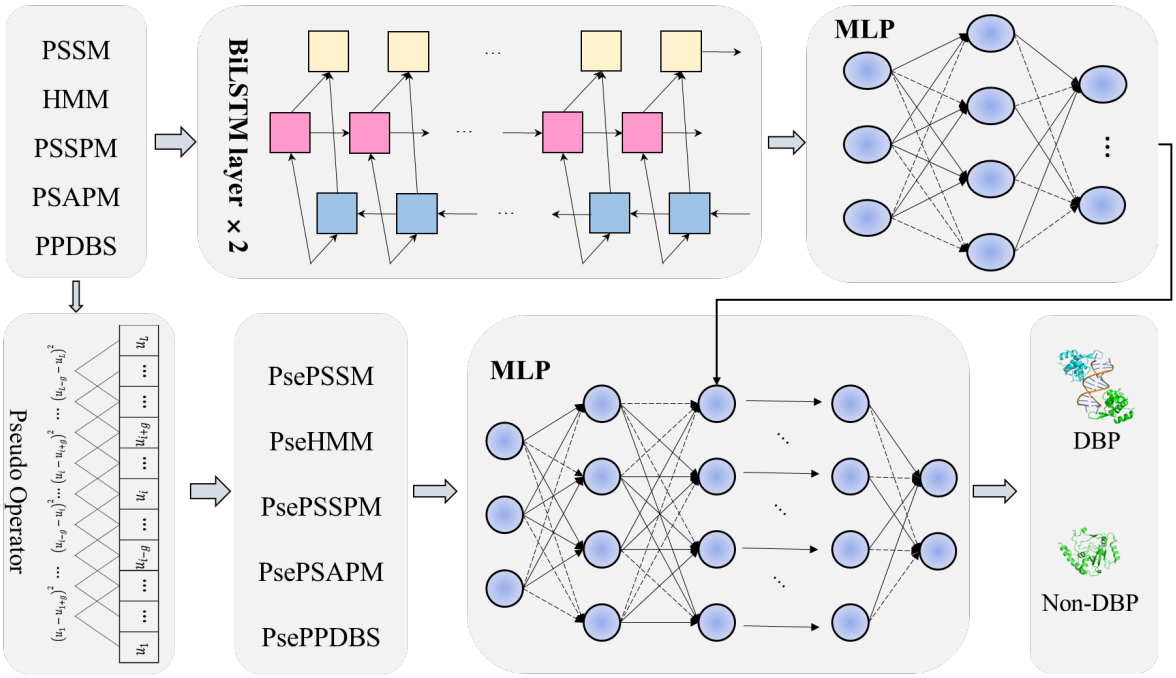


Figure S4. Structure of LBi-DBP-together.



REFERENCES

1. Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780.