Supplementary Information

Improved Protein Secondary Structure Prediction Using Bidirectional Long Short-Term Memory Neural Network and Bootstrap Aggregating

Wen-Wu Zeng [1], Ning-Xin Jia [1], Jun Hu [1,*]

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

[*] Address correspondence to J. Hu at hujunum@zjut.edu.cn

Supporting Texts

Text S1. Feature Representation of BiBagPSS

A. Position Specific Scoring Matrix

The PSSM of a protein sequence which obtained by the PSI-BLAST [1] program presents important evolutionary information. It is an $M \times 20$ dimensional matrix, where M represents the number of residues in a protein sequence, and 20 represents the evolution score of each residue against the 20 residues found in nature which based on BLOSUM62 [2]. The PSSM is often used by a variety of methods in the field of protein structure prediction, such as identification of DNA-binding protein prediction [3, 4], protein-nucleotide binding sites prediction [5, 6] and solvent accessibility prediction [7]. A lot of studies have proved that PSSM is a very effective feature representation which can observably improve the predict performance. In this study, we obtained the PSSM profile after three rounds of iterative search with 0.001 as the E-value cutoff for multiple sequence alignment against the query sequence. The PSSM matrix was changed by logistic function (Eq. S1) to normalize each element in PSSM.

$$f(x) = \frac{1}{1+e^{-x}} \tag{S1}$$

where x is the original element derived from PSSM profile.

B. Hidden Markov Model sequence profile

Besides PSSM, Hidden Markov Model sequence profile (HMM) is also employed to excavate the protein evolutionary information for reflecting the residue conservation and further improving the accuracy of protein secondary structure prediction. HMM is a $M \times 30$ dimensional matrix generated by HHBlits program [8] where the M represents the total number of residues in the protein sequences. HHBlits emerged in 2011 with higher speed, sensitivity and accuracy than PSI-BLAST. Every element in HMM is generated by the following equation (Eq. S2):

$$f(x) = -1000 \times log_2(x) \tag{S2}$$

where x are positive integers representing the match state amino acid emission frequencies. In this study, we first calculate the value of x from the HMM sequence file; then, process it with the logistic function (Eq. S1); finally, it is spliced with the PSSM and PSAPM as the input feature of the model.

C. Predicted Solvent Accessibility Probability Matrix

Protein solvent accessibility is an important local structural characteristic which closely related to the spatial arrangement and packing of amino acid residues. In this study, the predicted solvent accessibility probability matrix which obtained by feeding the protein sequence to the standalone SANN [9] program is utilized to represent the protein solvent accessibility information. For each protein sequence with L residues, the output profile of SANN program includes the probabilities of three solvent accessibility classes (i.e., buried (B), intermediate (I), and exposed (E)) of each residue. We extracted a predicted solvent accessibility probability matrix (PSAPM) of size $L \times 3$ from this file. In this study, PSAPM was combined with PSSM and HMM to represent based protein features and train the model.

Text S2. Architecture of BiLSTM

BiLSTM is the bidirectional Long Short-Term Memory (LSTM) [10], 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:

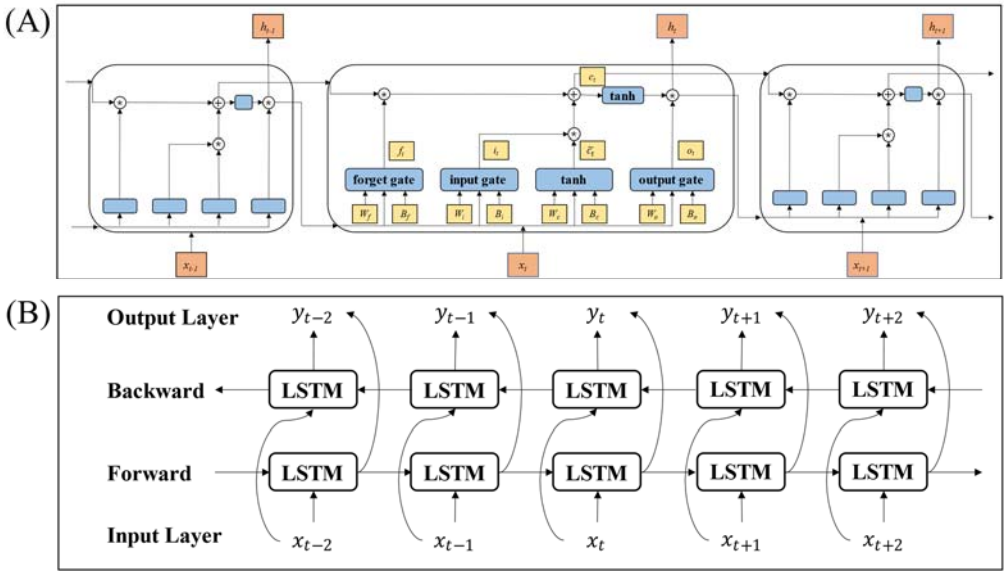

Figure S1. Architecture of LSTM unit and BiLSTM. (A) LSTM unit; (B) BiLSTM.

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + B_i) \tag{S3}$$

$$f_t = \sigma\left(W_f \times [h_{t-1}, x_t] + B_f\right) \tag{S4}$$

$$\widetilde{c_t} = tanh(W_c \times [h_{t-1}, x_t] + B_c) \tag{S5}$$

$$c_t = f_t * c_{t-1} + i_t * \widetilde{c_t} \tag{S6}$$

$$o_t = \sigma(W_O \times [h_{t-1}, x_t] + B_o) \tag{S7}$$

$$h_t = o_t * tanh(c_t) \tag{S8}$$

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. Since there are interrelationships between adjacent sequences of a protein, we chose BiLSTM as the main structure of the BiBagPSS model.

Text S3. Architecture of Fully connected neural network

Fully connected neural network (FC) is a classical deep learning network with strong nonlinear fitting ability.

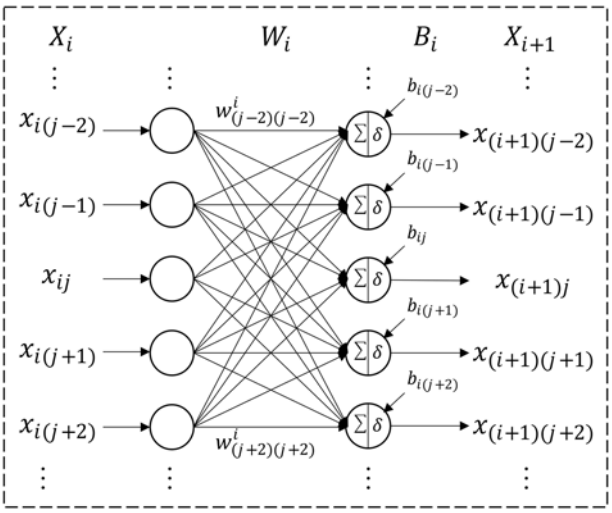

Figure S2. Architecture of fully connected neural network

As shown in Figure S2, the formula can be expressed as follow:

$$X_{i+1} = \delta(W_i * X_i + B_i) \tag{S9}$$

where X_i is the input vector of i-th FC 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 FC layer.

Text S3. Evaluation Indices

A. Q score

Q score is the most direct and important evaluation to the prediction results, and it is the ratio of the number of residues which correctly be predicted to the total residues. The formula (Eq. S10) for calculating Q score is as follows:

$$Q_m = 100 \times \frac{1}{N} \sum_{i=1}^{m} M_{ii} \tag{S10}$$

where m=3 and m=8 is referred as Q_3 and Q_8 score, respectively. N is the total number of residues in a protein sequence. M_{ii} is correctly predicted number of residues in state i.

The per-state accuracy is the percentage of correctly predicted residues in a particular state. The Q score of per-state accuracy is as follows:

$$Q_i = 100 \times \frac{M_{ii}}{obs^i} \tag{S11}$$

where obs^i is the total number of residues in state i. The Q score of the entire dataset is obtained by calculate the mean of Q scores of all protein sequences.

B. Sov score

Different from the Q score, which considers only a single residue, Sov score considers sequence continuity. For consecutive segments that are predicted correctly, Sov score will be high. For discontinuous segments, the Sov score will be very low. The accuracy of fragment prediction is extremely important because protein sequences are composed of continuous residues. The measure was originally proposed by Rost *et al.* [11] called segment overlap measure (Sov). Later Zemla *et al.* [12] improved it. In this study, we used improved version of the latter. As with the Q score, we averaged the Sov score for each protein to get the average Sov score for the entire dataset. The formula of Sov is as follows:

$$Sov = 100 \times \frac{1}{N} \sum_{i \in [H,E,C]} \sum_{S(i)} \frac{\min ov(s_1,s_2) + \delta(s_1,s_2)}{\max ov(s_1,s_2)} \times len(s_1) \tag{S12}$$

where N is a sum of $N(i)$ over all three states (H, E and C):

$$N = \sum_{i\in[H,E,C]} N(i) \tag{S13}$$

s_1 and s_2 are observed and predicted structure segments, respectively. $\min ov(s_1, s_2)$ represents the overlap segment of s_1 and s_2. $\max ov(s_1, s_2)$ represent the total extent for which either of the segments s_1 and s_2 has a residue in state i. The formula of $\delta(s_1, s_2)$ is as follows:

$$\delta(s_1, s_2) = \min\begin{Bmatrix} (\max ov(s_1, s_2) - \min ov(s_1, s_2)), \\ \min ov(s_1, s_2), int\left(\frac{len(s_1)}{2}\right), int\left(\frac{len(s_2)}{2}\right) \end{Bmatrix} \tag{S14}$$

REFERENCES

1. Altschul SF, Madden TL, Schäffer AA et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs 1997;25:3389-3402.
2. Henikoff S, Henikoff JGJPotNAoS. Amino acid substitution matrices from protein blocks 1992;89:10915-10919.
3. Lin W-Z, Fang J-A, Xiao X et al. iDNA-Prot: identification of DNA binding proteins using random forest with grey model 2011;6:e24756.
4. Hu J, Rao L, Zhu Y-H et al. TargetDBP+: Enhancing the Performance of Identifying DNA-Binding Proteins via Weighted Convolutional Features 2021;61:505-515.
5. Chen K, Mizianty MJ, Kurgan LJB. Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors 2012;28:331-341.
6. Chauhan JS, Mishra NK, Raghava GPJBb. Identification of ATP binding residues of a protein from its primary sequence 2009;10:1-9.
7. Hanson J, Paliwal K, Litfin T et al. Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks 2019;35:2403-2410.
8. Remmert M, Biegert A, Hauser A et al. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment 2012;9:173-175.
9. Joo K, Lee SJ, Lee J. Sann: solvent accessibility prediction of proteins by nearest neighbor method, Proteins: Structure, Function, and Bioinformatics 2012;80:1791-1797.
10. Hochreiter S, Schmidhuber J. Long Short-Term Memory, Neural Computation 1997;9:1735-1780.
11. Rost B, Sander C, Schneider RJJomb. Redefining the goals of protein secondary structure prediction 1994;235:13-26.
12. Zemla A, Venclovas Č, Fidelis K et al. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment 1999;34:220-223.