PS8-Net: A Deep Convolutional Neural Network to Predict the Eight-State Protein Secondary Structure

Md Aminur Rab Ratul, Maryam Tavakol Elahi, M. Hamed Mozaffari, WonSook Lee
School of Electrical Engineering and Computer Science (SITE)
University of Ottawa, Ottawa, Canada
{mratu076, mtava020, mmoza102, wslee}@uottawa.ca

***Abstract*—Protein secondary structure is crucial to creating an information bridge between the primary and tertiary structures. Precise prediction of eight-state protein secondary structure (PSS) has been significantly utilized in the structural and functional analysis of proteins. Deep learning techniques have been recently applied in this area and raised the eight-state (Q8) protein secondary structure prediction accuracy remarkably. Nevertheless, from a theoretical standpoint, there are still many rooms for improvement, specifically in the eight-state PSS prediction. In this study, we have presented a new deep convolutional neural network called PS8-Net, to enhance the accuracy of eight-class PSS prediction. The input of this architecture is a carefully constructed feature matrix from the proteins sequence features and profile features. We introduce a new PS8 module with skip connection to extracting the long-term inter-dependencies from higher layers, obtaining local contexts in earlier layers, and achieving global information during secondary structure prediction. This architecture enables the efficient processing of local and global interdependencies between amino acids to make an accurate prediction of each class. To the best of our knowledge, our proposed PS8-Net experiment results demonstrate that it outperforms all the state-of-the-art methods on the benchmark CullPdb6133, CB513, CASP10, and CASP11 datasets.**

***Index Terms*—Deep Convolutional Neural Network, Protein Secondary Structure Prediction, Bioinformatics, Skip Connection**

I. METHODOLOGY

In our proposed PS8-Net, we have different convolutional layers with distinct convolution operations, three PS8 modules, skip connections, and fully connected layers. PS8 modules and skip connection accommodate several convolutional layers.

A. Input Feature Representation of PS8-Net

We cautiously design initial feature representation, which consists of sequence features and rich information of profile features because feature representation is valuable for this prediction task. In this study, 21-dimensional sequence features have been used to encode the types of target residues. 21-dimensional profile features acquired from the PSI-BLAST log file [1] and later logistic function [2] used to re-scale it. Sequence features are represented as a 21-dimensional one-hot vector, and the profile features have dense representation. Therefore, to maintain the consistency in all feature representations of proteins, we apply an embedding operation to convert sparse sequence features into the dense vector. This embedding technique maps the 21-dimensional dense vector from the 21-dimensional sparse vector. Finally, a 42-dimensional initial protein feature can be obtained by concatenating both the dense representation of the sequence and the profile features and representing a feature vector x_n.

The input of the PS8-Net is the $I \times d$ matrix, where $I \times d$ matrix can be encoded as $x_{1:I} = [x_1, x_2, , x_I]^T$. Here, I is the length of protein and d denote the number of input features utilized to encode residues. The value of $I = 700$ and $d = 42$, and the output of this network is eight-state secondary structure labels of a specific protein (X) which can be formulated as, $S = s_1, s_2, s_3, ..., s_{I-1}, s_I$.

B. PS8-Net with PS8 Modules

In a deep CNN (DCNN), a kernel or sliding window can scrutinize the local patch from the provided input sequences. In the local patch, this kernel operation extract interdependence among the amino acid residues. To get local dependencies of adjoining amino acids throughout the system, we apply four different kernel sizes: 1, 3, 5, 11. In Figure 1, "CONV 1", "CONV 3", "CONV 5", and "CONV 11" denote convolutional layers with kernel size 1, 3, 5, and 11, respectively. Moreover, we employ the $ReLU$ activation in these layers to extract long-term inter-dependencies between residues with more long-distance. In the output, to retain the same height as the input, in the head and in the tail of the input $x_{1:I}$ respectively, we require to pad $[L/2]$ and $[(L-1)/2]$, where L denote the length of kernels. In PS8 modules and "CONV5" after applying convolution operation of the k^{th} kernel on all input sequences incorporate length L of the padded input, a feature vector is acquired. Besides, If we have p numbers of kernels, then we can produce p numbers of the feature vector. Next, we obtain a new feature matrix from a convolutional layer with a $I \times p$ dimension by concatenating p feature vectors.

There are four series of the convolutional layer in the PS8 module. The first and fourth series contain "CONV 3" followed by the "CONV 1" layer. Furthermore, the second and third series include three "SK Block 1" with these "CONV 1" and "CONV 3" layers. These four series follow the same methods to generate the outcome. Finally, we can concatenate the output of these four series of convolutional layers to get the final output of the PS8 module. The number of hidden units throughout the PS8 Module A is 256, whereas Module B and C contain 128. Finally, throughout these modules, we apply a 0.25 dropout rate to prevent overfitting.

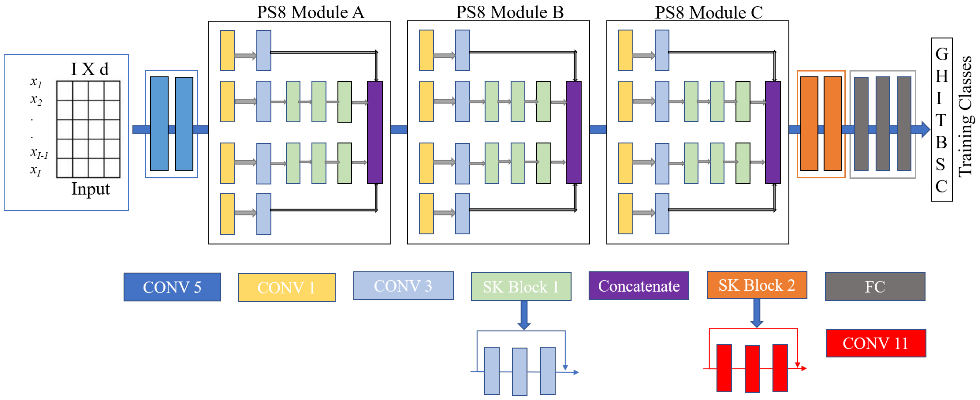

Fig. 1. The schematic of our proposed PS8-Net model. It is comprised of three PS8 modules, two skip connection blocks, and three fully connected layers.

C. Skip Connection

A CNN with many convolutional layers can extract long-term interdependencies from long sequences of amino acid residues. However, when we enhance the number of CNN layers, the network will drop the important local contexts information learned during the initial training process. Besides, the vanishing gradients problem is a colossal issue in CNN, which is integrated with a large number of layers [3]. To tackle this issue, we employ ”SK Block 1” and ”SK Block 2,” which are based on the skip connections mechanism [4] [5]. The PS8 module and ”skip connection” help us achieve local contexts from lower layers for PSS prediction.

”SK Block 1” and ”SK Block 2” contains the kernel size of 3 and 11, respectively. Additionally, these skip connection blocks have three convolutional layers with BatchNormalization and $ReLU$ activation. In these blocks, “skip connection” has been utilized to backpropagate the gradient to the initial layers so that any SK Blocks can suitably learn the identity mapping [5]. The output of the SK Blocks is based on the knowledge gain from the input of the block and the outcome of convolutional kernels of the final layer.

D. PS8-Net Architecture

”CONV 5” and ”SK Block 2” respectively incorporate the hidden units 512 and 128. The classifier part has three fully connected (FC) layers with 512, 256, and 8 (eight classes) hidden units, respectively. We used the $ReLU$ activation function in the first two FC layers and SoftMax in the last one. We trained our proposed PS8-Net for 120 epochs with Adam optimizer when the learning rate is $2e-4$. The mini-batch size is 64, and we apply cross-entropy loss. We have used one callback function to reduce the learning rate factor by $\sqrt{0.1}$ if learning stagnates for 7 epochs.

II. Results and Discussion

Here, we focused on the overall accuracy of eight classes of protein secondary structure prediction on the five public datasets and compared the performance with popular state-of-the-art methods. We train our model in two different manners. Firstly, train on 6128 protein sequences of CullPdb6133, validated on 256 protein sequences, and tested on the remaining 272 protein sequences. Secondly, picked 5234 protein sequences from the CullPdb6133-filtered dataset for training, validated on the remaining 300 sequences, and tested on CB513, Casp10, Casp11 datasets. To inspect the importance of both sequence and profile features, we trained our network with CullPDB6133-filtered and tested on CB513.

TABLE I
Performance analysis for PS8-Net on various input features

Sequence Features	✓		✓
Profile Features		✓	✓
Q8 (%) on CB513	61.57	69.72	**71.94**

TABLE II
Comparison of Q8 (%) accuracy for PSS of PS8-Net and some state-of-the-art method where best results marked in **bold**

Method	CullPdb6133	CB513	CASP10	CASP11
SSPro-8 [6]	66.6	63.5	64.9	65.6
DeepCNF [7]	75.2	68.3	71.8	72.3
DCRNN [8]	73.2	69.4	-	-
CNNH-PSS [9]	74.0	70.3	-	-
DeepACLSTM [10]	-	70.5	75.0	73.0
MUFOLD-SS [11]	-	70.63	76.47	74.51
2DConv-BLSTM [12]	75.7	70.2	74.5	72.5
2DCNN-BLSTM [12]	74.3	70.0	74.5	72.6
PS8-Net (ours)	**76.9**	**71.94**	**76.86**	**75.26**

According to TABLE I, we can exhibit that our proposed network acquire superior results when we take both sequence and profile features as the input. In TABLE II, we provide the test result of our two different training approaches together. PS8-Net shows superior overall accuracy than any other state-of-the-art models for eight-state PSS on CULLPdb6133, CB513, CASP10, CASP11, to the best of our knowledge.

To conclude, we have introduced a new CNN architecture, namely PS8-Net, to predict the eight-state PSS. Our newly proposed PS8 module with skip connection efficiently processed the long-term inter-dependencies extraction from higher layers, obtaining local contexts in earlier layers, and achieving global information during secondary structure prediction. Our proposed PS8-Net is trained and tested on five benchmark datasets and display that it surpasses all state-of-the-art methods.

References

[1] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic acids research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[2] 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.

[3] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Advances in neural information processing systems*, 2015, pp. 2377–2385.

[4] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4799–4807.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[6] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, "Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles." *Proteins*, vol. 47 2, pp. 228–35, 2001.

[7] S. Wang, J. Peng, J. Ma, and J. Xu, "Protein secondary structure prediction using deep convolutional neural fields," *Scientific reports*, vol. 6, no. 1, pp. 1–11, 2016.

[8] Z. Li and Y. Yu, "Protein secondary structure prediction using cascaded convolutional and recurrent neural networks," *arXiv preprint arXiv:1604.07176*, 2016.

[9] J. Zhou, H. Wang, Z. Zhao, R. Xu, and Q. Lu, "Cnnh_pss: protein 8-class secondary structure prediction by convolutional neural network with highway," *BMC bioinformatics*, vol. 19, no. 4, pp. 99–109, 2018.

[10] Y. Guo, W. Li, B. Wang, H. Liu, and D. Zhou, "Deepaclstm: deep asymmetric convolutional long short-term memory neural models for protein secondary structure prediction," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–12, 2019.

[11] C. Fang, Y. Shang, and D. Xu, "Mufold-ss: New deep inception-inside-inception networks for protein secondary structure prediction," *Proteins: Structure, Function, and Bioinformatics*, vol. 86, no. 5, pp. 592–598, 2018.

[12] Y. Guo, B. Wang, W. Li, and B. Yang, "Protein secondary structure prediction improved by recurrent neural networks integrated with two-dimensional convolutional neural networks," *Journal of bioinformatics and computational biology*, vol. 16, no. 05, p. 1850021, 2018.