Zhou *et al. BMC Bioinformatics* 2018, **19**(Suppl 4):60
https://doi.org/10.1186/s12859-018-2067-8

BMC Bioinformatics

RESEARCH Open Access

CNNH_PSS: protein 8-class secondary structure prediction by convolutional neural network with highway

Jiyun Zhou[1,2], Hongpeng Wang[1], Zhishan Zhao[1], Ruifeng Xu[1*] and Qin Lu[2]

From The Sixteenth Asia Pacific Bioinformatics Conference
Yokohama, Japan. 15-17 January 2018

Abstract

Background: Protein secondary structure is the three dimensional form of local segments of proteins and its prediction is an important problem in protein tertiary structure prediction. Developing computational approaches for protein secondary structure prediction is becoming increasingly urgent.

Results: We present a novel deep learning based model, referred to as CNNH_PSS, by using multi-scale CNN with highway. In CNNH_PSS, any two neighbor convolutional layers have a highway to deliver information from current layer to the output of the next one to keep local contexts. As lower layers extract local context while higher layers extract long-range interdependencies, the highways between neighbor layers allow CNNH_PSS to have ability to extract both local contexts and long-range interdependencies. We evaluate CNNH_PSS on two commonly used datasets: CB6133 and CB513. CNNH_PSS outperforms the multi-scale CNN without highway by at least 0.010 Q8 accuracy and also performs better than CNF, DeepCNF and SSpro8, which cannot extract long-range interdependencies, by at least 0.020 Q8 accuracy, demonstrating that both local contexts and long-range interdependencies are indeed useful for prediction. Furthermore, CNNH_PSS also performs better than GSM and DCRNN which need extra complex model to extract long-range interdependencies. It demonstrates that CNNH_PSS not only cost less computer resource, but also achieves better predicting performance.

Conclusion: CNNH_PSS have ability to extracts both local contexts and long-range interdependencies by combing multi-scale CNN and highway network. The evaluations on common datasets and comparisons with state-of-the-art methods indicate that CNNH_PSS is an useful and efficient tool for protein secondary structure prediction.

Keywords: Protein secondary structure, Convolutional neural network, Highway, Local context, Long-range interdependency

Background

The concept of secondary structure was first introduced by Linderstrøm-Lang at Stanford in 1952 [1, 2] to represent the three dimensional form of local segments of proteins. Protein secondary structure is defined by the pattern of hydrogen bonds between the amine hydrogen and carbonyl oxygen. There are two ways used for the classification of protein secondary structures: three-category classification(Q3) and eight-category classification(Q8). Q3 classifies target amino acid residues into helix(H), strand(E) and coil(C) while Q8 classifies target amino acid residues into (1) 3-turn helix(G), (2) 4-turn helix(H), (3) 5-turn helix(I), (4) hydrogen bonded turn(T), (5) extended strand in parallel and/or anti-parallel β-sheet conformation(E), (6) residue in isolated β-bridge (B), (7) bend(S) and (8) coil(C) [3–5]. Most protein secondary structure prediction studies have been focused Q3 prediction. Q8 prediction is more challenging and can reveal more structural details [6, 7], so we focus the Q8 prediction in this study.

* Correspondence: xuruifeng@hit.edu.cn
[1]School Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, HIT Campus Shenzhen University Town, Xili, Shenzhen, Guangdong 518055, China
Full list of author information is available at the end of the article

Protein secondary structure prediction is secondary structure inference of protein fragments based on their amino acid sequence. In bioinformatics and theoretical chemistry, protein secondary structure prediction is very important for medicine and biotechnology, for example drug design [8] and the design of novel enzymes. Since secondary structure can be used to find distant relationship for proteins with unalignable primary structures, incorporating both secondary structure information and simple sequence information can improve the accuracy of their alignment [9]. Finally, protein secondary structure prediction also plays an important role in protein tertiary structure prediction. Protein secondary structure can determine the structure types of protein local fragments, so the freedom degree of protein local fragments in the tertiary structure can be reduced. Therefore accurate secondary structure prediction is potential for improving the accuracy of protein tertiary structure prediction [4, 7, 10].

Three experimental methods were proposed to determine secondary structures for proteins: far-ultraviolet circular dichroism, infrared spectroscopy and NMR spectrum. Far-ultraviolet circular dichroism predict pronounced double minimum at 208 and 222 nm as α-helical structure and single minimum at 204 nm or 217 nm as random-coil or β-sheet structure, respectively [11]. Infrared spectroscopy uses the differences in the bond oscillations of amide groups for prediction [12] while NMR spectrum predict protein secondary structure by using the estimated chemical shifts [12]. As experimental methods are costly and the proteins with known sequence continue to outnumber the experimentally determined secondary structures, developing computational approaches for protein secondary structure prediction becomes increasingly urgent. Existing computational approaches for protein secondary structure prediction can be divided into 3 categories. The first category is statistical model based methods, which can date back to 1970s. Early, this category uses statistical models to analyze the probability of secondary structure elements for individual amino acid residue [13]. Next, the statistical models were applied for the prediction of segments of 9–21 amino acids. For example, the GOR method [14] used amino acid segment to predict the structure of its central residue. However, the performances (< 60% Q3 accuracy) of this category of methods are far from practical application due to inadequate features.

Due to the lacking of inadequate features for the statistical model based methods, evolutionary information based methods have been proposed. These methods usually used the evolutionary information of proteins from a same structural family [15] extracted by multiple-sequence alignment or position-specific scoring matrices (PSSM) [16] from PSI-BLAST for prediction. An earlier evolutionary information based method was developed based on a two-layered feed-forward neural network, for which the evolutionary information in the form of multiple sequence alignment is used as input instead of single sequences [15]. As SVM [17] is significantly better than neural network in a wide range of pattern recognition problems [18–21], Hua and Sun first proposed a SVM classifier for protein secondary structure prediction [22]. The input for SVM is evolutionary information in the form of multiple sequence alignment. Unbalanced data is a challenging problem in protein secondary structure prediction and existing methods lack the ability to handle it [23, 24]. So Kim and Park proposed a new protein secondary structure prediction method, SVMpsi, by an improved SVM, which reduces the influence of imbalanced data by using different penalty parameters in the improved SVM [23]. By using different penalty parameters, SVMpsi resolved the situation where the recall value of the smaller class is too small. Another SVM based method is PMSVM which was proposed by using dual-layer support vector machine (SVM) and evolutionary information in form of PSSMs [25].

Protein sequence usually contains two types of sequence information: local context and long-range interdependencies [4, 26, 27]. Local contexts denote the correlations between residues with distance less than or equal to a predefined threshold while long-range correlation are the correlations between residues with distance more the threshold. Inspired by the success of convolutional neural networks (CNN) [28] for local context extraction in natural language processing tasks [29, 30], multi-scale CNN has been used to capture local contexts for protein secondary structure prediction [31]. For example, Wang et al. proposed conditional neural fields (CNF) [7] to extract local contexts for prediction by integrating a windows-based neural network with conditional random field (CRF). In addition to local contexts, long-range interdependencies also are important for protein secondary structure prediction(BRNN) [6, 32–34]. In order to extract both local contexts and long-range interdependencies for prediction, Zhou and Troyanskaya proposed GSN [4] by using convolutional architecture and supervised generative stochastic network, which is a recently proposed deep learning technique [26]. In addition to GSN, a novel deep convolutional and recurrent neural network (DCRNN) also has been proposed by Li and Yu [27] for protein secondary structure prediction by extracting both local contexts and long-range interdependencies.

In summary, the statistical model based methods and evolutionary information based methods cannot extract local contexts and long-range interdependencies for prediction. For the deep learning based methods, some methods cannot extract both local contexts and long-range interdependencies for prediction. Although several methods can extract both local contexts and long-range interdependencies, such as GSN and DCRNN, they need extra complex models to extract long-range interdependencies, which are complex and time-consuming. In this paper, we propose a novel method, referred to as CNNH_PSS, by combining multi-scale CNN with highway network, which has ability to extract both local contexts and long-range interdependencies without needing extra models. CNNH_PSS consists of two parts: multi-scale CNN and fully connected and softmax layer. In the multi-scale CNN, any two neighbor convolutional layer contains a highway to deliver information from current layer to the output of the next one to keep local contexts. As the convolutional kernels in higher layer can extract long-range interdependencies by using the local contexts extracted by lower layers, thus with the layer number increasing, CNNH_PSS can extract long-range interdependencies covering more remote residues while keeping local contexts extracted by lower layers by using highway. So CNNH_PSS can extract both local contexts and long-range interdependencies covering very remote residues for prediction. The source code of our proposed method CNNH_PSS is provided for free access to the biological research community at http://hlt.hitsz.edu.cn/CNNH_PSS/ and http://119.23.18.63:8080/CNNH_PSS/.

Methods

As shown by many recently published works [35–37], a complete prediction model in bioinformatics should contain the following four components: validation benchmark dataset(s), an effective feature extraction procedure, an efficient predicting algorithm, a set of fair evaluation criteria. In the following text, we will describe the four components of our proposed CNNH_PSS in details.

Datasets

Two publicly available datasets: CB6133 and CB513 were used to evaluate the performance of our proposed method CNNH_PSS and compare with state-of-the-art methods.

CB6133

CB6133 was produced by PISCES CullPDB [38] and is a larger non-homologous protein dataset with known secondary structure for every protein. It contains 6128 proteins, in which 5600 proteins are training samples, 256 proteins are validation samples and 272 proteins are testing samples. This dataset is publicly available from literature [4].

CB513

CB513 is a public testing dataset and can be freely obtained from [4, 39]. For the testing on CB513, CB6133 is used as the training dataset. As there exists redundancy between CB513 and CB6133, CB6133 is filtered by removing sequences having over 25% sequence similarity with sequences in CB513. After filtering, 5534 proteins left in CB6133 are used as training samples. Since GSN [4] and DCRNN [27] as well as other state-of-the-art methods [31, 40] performed a validation on CB513 to get their best performance, we also perform a validation on CB513 to get the best performance of our method CNNH_PSS to make fair comparisons with them.

Feature representation

Given a protein with L amino acid residues as $X = x_1, x_2, x_3, \cdots, x_L$, where $x_i(\in \mathbb{R}^m)$ is the m-dimensional feature vector of the i^{th} residue, the secondary structure prediction for this protein is formulated as determining $S = s_1, s_2, s_3, \cdots, s_L$ for X where s_i is a Q8 secondary structure label. In this study,x_i is encoded by both sequence features and evolutionary information. Sequence features are used to specify the identity of the target residue. Two methods are used to encode sequence features: one hot and residue embedding. One hot encodes sequence features of each residue by a 21-dimension one-hot vector, in which only one element equals to 1 and the remaining elements are set to 0, where 21 denotes the 20 standard types of residues and one extra residue type which represents all non-standard residue types. However, one-hot vector is a sparse representation and unsuitable for measuring relation between different residues. In order to get dense representation of sequence features, an embedding technique in natural language processing is used to transform 21-dimensional one-hot vector to a 21-dimensional denser representation [41]. The embedding technique maps words or phrases from the vocabulary to vectors of real numbers. Specifically, it maps words from a space with one dimension per word to a continuous vector space with much lower dimension. So the embedding technique provides a real value for every dimension. As the dimension of amino acid representation is already low, we only calculate a real value for every dimension by embedding technique and don’t decrease the dimension. The residue embedding in this paper is implemented by a feedforward neural network layer before multi-scale CNN in CNNH_PSS [42].

Evolutionary information such as position-specific scoring matrix (PSSM) is considered as informative features for predicting secondary structure by previous research [16]. PSSM is a common representation for evolutionary information and has been used in many bioinformatics studies including protein functionality annotation and protein structure prediction [43–47]. In this study, PSSM is calculated by PSI-BLAST [48] against the UniRef90 database with E-value threshold 0.001 and 3 iterations. UniRef90 database contains the known protein sequences with sequence identity less than 90 from almost all known species. So the PSSM calculated from UniRef90 database contains the common sequence information among the known protein sequences of different species. Overtime, scientists have reached a consensus that a protein's structure primarily depends on its amino acid sequence and concluded that the local and long-range interaction are a cause of protein second and tertiary structure. Based on this hypothesis, we can deduce that proteins with similar amino acid sequence tend to have similar secondary structure sequence. Therefore, the common sequence information contained by PSSM can contribute to the secondary structure prediction. For a protein with length L, PSSM is usually represented as a matrix with $L \times 21$ dimensions where 21 denotes the 20 standard types of residues and one extra residue type which represents all non-standard residue types. Before PSSMs are used inputs for CNNH_PSS, they need to be transformed to 0–1 range by the sigmoid function. By concatenating sequence features and evolutional information, each residue in protein sequences can be encoded by a feature vector with dimension of 42.

Multi-scale CNN with highway between neighbor layers

In CNN model, a kernel can examine a local patch in input sequence and extract interdependence among the residues contained in the local patch. With stacking of convolutional layers, the kernels for deep layers have ability to cover correlations among more spread-out residues in the input sequence. So, CNN model with more number of convolutional layers have the ability to extracted long-range interdependencies between residues with more large distance. However, with the number of layers increasing, CNN model will lose the local contexts extracted by lower layers. In this paper, we propose a novel method, referred to as CNNH_PSS, to resolve this problem. CNNH_PSS contains a highway between any two neighbor convolutional layers in multi-scale CNN. As the number of convolutional layers increases, CNNH_PSS can not only extract long-range interdependencies by higher layers, but also obtain the local contexts extracted by lower layers through highway. The frame of CNNH_PSS is shown in Fig. 1. Figure 1 shows that CNNH_PSS contains three parts: input section, multi-scale CNN with highway and output section. In the input section, $x_i \in \mathrm{R}^m$ denotes the feature vector of the i^{th} residue in protein, which is the concatenation of sequence features and evolutional information. Thus a protein of length L is encoded as a $L \times m$ matrix $x_{1:L} = [x_1, x_2, \cdots, x_L]^{\mathrm{T}}$, where L and m denote the length of protein and the number features used to encode residues, respectively. In this study, m equals to 42. In order to keep the output of convolutional layer have the same height with the input, we need to pad $\lfloor h/2 \rfloor$ and $\lfloor (h-1)/2 \rfloor$ m-dimensional zero vectors to the head and the tail of the input $x_{1:L}$, respectively, where h is the length of convolutional kernels in the convolutional layer. The second section contains two parts: multi-scale CNN and highway, where the multi-scale CNN contains n convolutional layers. In the $(t-1)^{\text{th}}$ layer, the convolution operation of the k^{th} kernel $w_k^{t-1} \in \mathrm{R}^{h \times m}$ executed on protein fragment $x_{i:i+h-1}$ is expressed as

$$c_{k,i}^{t-1} = f^{t-1}\left(w_k^{t-1} \cdot x_{i:i+h-1} + b_k^{t-1}\right) \tag{1}$$

where h is the length of convolution kernel,b_k^{t-1} is the bias of the k^{th} kernel, f is activation function and $x_{i:i+h-1}$ denotes the protein fragment x_i, x_{i+1}, x_{i+2}, $\cdots$, x_{i+h-1}. Through executing convolution operation of the k^{th} kernel on all fragments with length h of the padded input, we get a novel feature vector

$$c_k^{t-1} = \left[c_{k,1}^{t-1}, c_{k,2}^{t-1}, c_{k,3}^{t-1}, \cdots, c_{k,L}^{t-1}\right]^{\mathrm{T}} \tag{2}$$

Suppose we have d kernels in the convolutional layer, thus we can get d novel features vectors. By concatenating the d novel feature vectors, we can get a novel feature matrix with dimension $L \times d$

$$c^{t-1} = \left[c_1^{t-1}, c_2^{t-1}, c_3^{t-1}, \cdots, c_L^{t-1}\right] \tag{3}$$

This novel feature matrix is used as the input of the next convolutional layer. If there are n convolutional layers and θ_t is used to denote the kernels and the bias of the t^{th} convolutional layer, then the output of the n^{th} convolutional layer is

$$c^n = f_{\theta_n}^n\left(f_{\theta_{n-1}}^{n-1}\left(\cdots f_{\theta_1}^1(x_{1:L})\right)\right) \tag{4}$$

Finally, the output of the n^{th} convolutional layer is used as the input of the fully connected softmax layer for prediction

$$y_i = argmax(w \cdot c_i{}^n + b) \tag{5}$$

where w and b is the weight and bias of the fully connected softmax layer, respectively. c_i^n is the feature vector of the i^{th} outputted by the n^{th} convolutional layer and y_i is its predicted secondary structure.

CNN has achieved huge progress in many tasks of image processing filed, one common sense is that the successes of CNN are attributed to the multiple convolutional layers in CNN, because CNN with more number of layers can extract correlations covering more residues. However, with the increasing of number of layers in CNN, the information communication between layers will become more difficult and the gradient will disappear [49]. Furthermore, the local contexts extracted by lower layers also will lose. Srivastava et al. [49] have proposed highway network to resolve these problems. So

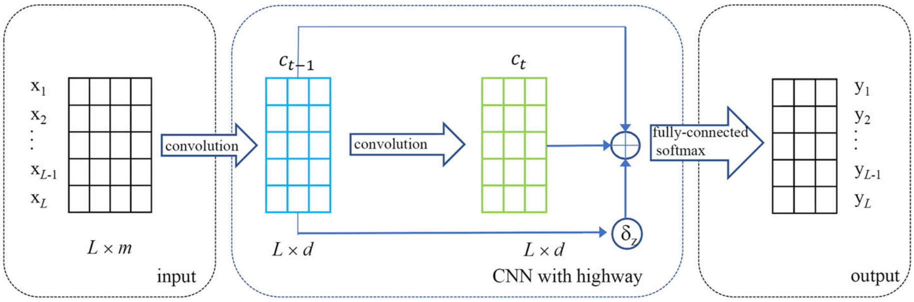

Fig. 1 The frame of CNNH_PSS

CNNH_PSS incorporates highway network and multi-scale CNN to extract both local contexts and long-range interdependencies for secondary structure prediction.

In CNNH_PSS, each convolutional layer except the last layer has three accesses to the next layer (shown in Fig. 1). Two accesses are used to deliver information from the current layer to the output and convolution kernels of the next layer, respectively. The other one is a weight used to determine the share of information in for information from highway. So the output c^t of the t^{th} convolutional layer is the weighted sum of the information delivered by highway from last layer and that outputted by the convolution kernels of current layer

$$z_t = \delta\left(w^z c^{t-1}\right) \quad (6)$$

$$c^t = (1-z_t) \times f^t_{\theta_t}\left(c^{t-1}\right) + z_t \times c^{t-1} \quad (7)$$

where $\delta(\cdot)$ is *sigmoid* function,z_t is the weight of the highway and $f^t_{\theta_t}(\cdot)$ is the convolution operation of current convolutional layer. So the output of thet^{th}convolutional layer contains two portion: information from the $(t-1)^{th}$convolutional layer delivered by highway and that outputted by the convolution kernels of current layer.

Results

The purpose of the evaluation is to examine the effectiveness of our proposed CNNH_PSS over other methods. Four sets of evaluations are conducted here. The first experiment evaluates the performance of multi-scale CNN on CB6133 and CB513. The second experiment evaluates our proposed method CNNH_PSS on CB6133 and CB513. The third experiment compares CNNH_PSS with state-of-the-art methods. Finally, based on CB6133, we analyze the local contexts and long-range interdependencies learned by CNNH_PSS. As we mainly focus the Q8 prediction of protein secondary structure in this study, the performances of prediction methods are measured by Q8 accuracy [4, 27]. The Q8 accuracy is the percentage of the amino acid residues for which the predicted secondary structure labels are correct. The source code of our proposed method CNNH_PSS is provide for free access at http://hlt.hitsz.edu.cn/CNNH_PSS/.

The performance of multi-scale CNN model

In this section, multi-scale CNN is used to predict secondary structure for proteins. The hyper-parameters of the multi-scale CNN for protein secondary structure prediction in this study are listed in Table 1. Note that three kernel lengths are used in the multi-scale CNN model and 80 kernels are used for each kernel length. To conveniently encode and process protein sequences, the length of all protein sequences are normalized to 700. When sequences are shorter than 700, they will be padded with zero vectors. And when sequences are longer than 700, they will be truncated. In order to get best performance, we need to determine how many convolutional layers the multi-scale CNN should contains. We conduct experiments to evaluate the performances of the multi-scale CNNs with different number of convolutional layers on CB513. The performances are shown in Fig. 2, where the *x*-axis is the number of epochs used to train multi-scale CNN and the *y*-axis is Q8 accuracy. Fig. 2 shows the performances for models with number of convolutional layers from 1 to 5. From this figure, we see that the model with 3 convolutional layers gets the best accuracy. When the number of convolutional layers is increased to 4 or 5, the accuracy is decreased obviously. The main reason for this phenomenon may be the loss of extracted local contexts with the increasing of the

Table 1 Hyper-parameters of multi-scale CNN

Layer	Hyper-parameter	Value
Multi-scale CNN	Kernel length	[7, 9, 11]
	Number of kernels	80 for each kernel length
	Batch size	50
	Learning rate	2e-3
	Regularizer	5e-5
	Decay rate	0.05
	Activation function	ReLU

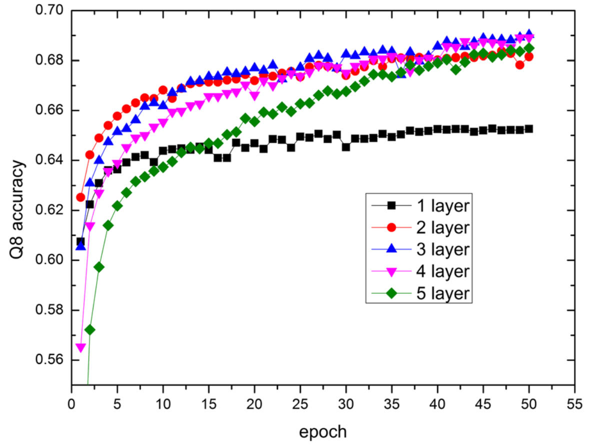

Fig. 2 The performance of multi-scale CNN with different number of convolutional layers

number of convolutional layers in CNN. With the increasing of the number of convolutional layers in CNN, the correlations extracted by higher can cover more residues so that they contains interdependencies between more remote residues. When the number of convolutional layers is increased to 3, the CNN may achieve both local contexts and long-range interdependencies, which is validated by that the CNN with 3 convolutional layers gets the best accuracy in our problem. However, when the number of convolutional layers is more than 3, most local contexts extracted by lower layers are lost in the transport processes of information cross layers, causing the relationships outputted by the last layer in CNN contains less and less local contexts. So the predicting accuracy starts to decrease when the number of convolutional layers is more than 3.

The performances of multi-scale CNN with 3 convolutional layers on CB6133 and CB513 are shown in Table 2, where two sequence features encoding methods for residues are evaluated: one hot and residue embedding. Table 2 shows that residue embedding outperforms one hot on both CB6133 and CB513 by at 0.004 Q8 accuracy, indicating that residue embedding is a better encoding method for sequence features of residues. In the following text, we will use residue embedding method to encode sequence features in the multi-scale CNN model and our proposed method CNNH_PSS.

Table 2 The Q8 accuracy of Multi-scale CNN with 3 convolutional layers

datasets	CB6133	CB513
Multi-scale CNN(one hot)	0.721	0.689
Multi-scale CNN(embedding)	*0.729*	*0.693*

The data in italic denote the best performance

The performance of CNNH_PSS

Local contexts are the relationships among residues at close range while long-range interdependencies are the relationships among remote residues. As there is no strict bounds between local contexts and long-range interdependencies, we specify the information extracted by the first convolutional layers as local contexts and that extracted by all other layers as long-range interdependencies in this study. In CNNH_PSS, any two neighbor convolutional layers have a highway to deliver information from current convolutional layer to the output of the next one, so it can make sure that the output of each layer contains a portion of the local contexts. Furthermore, the convolution kernels of each layer except the first one can extract long-range interdependencies by using the information from previous layer. With the increasing of the number of convolutional layers, CNNH_PSS can extract long-range interdependencies between more remote residues. Therefore, the output the last convolutional layer in CNNH_PSS contains two portion of information: local contexts extracted by the first layer and long-range interdependencies extracted by all other layers.

Similarly, we evaluate the performances of our proposed method CNNH_PSSs with different number of convolutional layers on CB513. The performances are shown in Fig. 3. Figure 3 shows that CNNH_PSS achieves the best performance when the number of convolutional layers is 5. When number of convolutional layers is more than 5, the performance of our method starts to decrease. So CNNH_PSS with 5 convolutional layers is used in the following text. Comparing the multi-scale CNN model with 3 convolutional layers described in above section, our proposed method CNNH_PSS not only contains a highway between any two neigbouring layers, but also have more number of

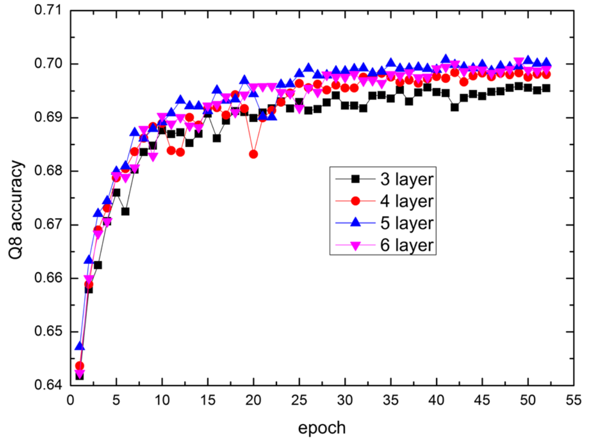

Fig. 3 The performance of CNNH_PSS with different number of convolutional layers

convolutional layers. It makes sure that CNNH_PSS with 5 layers can not only extract local contexts, but also capture long-range interdependencies between more remote residues than the multi-scale CNN model with 3 layers. The performances of CNNH_PSS and the multi-scale CNN model on CB6133 and CB513 are shown in Table 3. Table 3 shows that our proposed method CNNH_PSS outperforms the multi-scale CNN model by 0.011 Q8 accuracy on CB6133 and 0.010 Q8 accuracy on CB513. The outperformance of CNNH_PSS over the multi-scale CNN model on both CB6133 and CB513 validates that the highway in CNN indeed are useful for protein secondary structure prediction.

Comparison with state-of-the-art methods

Protein secondary structure prediction is an important problem in bioinformatics and critical for analyzing protein function and applications like drug design. So many state-of-the-art methods have been proposed for the prediction. SSpro8 is a prediction method proposed by Pollastri et al. [6] by combining bidirectional recurrent neural networks (RNN) and PSI-BLAST-derived profiles. CNF is a Conditional Neural Fields based method which was proposed by Wang et al. [40], which can not only extract relationships between sequence features of residues and their secondary structures, but also capture local contexts [40]. Later, an extension version of CNF (DeepCNF) was proposed by Wang et al. [31] using deep learning extension of conditional neural fields, which is an integration of conditional neural fields and shallow neural networks. It can extract both complex sequence-structure relationship and interdependency between adjacent SS labels. These three methods can only extract local contexts for prediction. Furthermore, GSN is a prediction method proposed by Zhou and Troyanskaya [4] using supervised generative stochastic network and convolutional architectures. Supervised generative stochastic network is a recently proposed deep learning technique [26], which is well suitable for extracting local contexts and also can capture some long-range interdependencies. Finally, DCRNN is the best performing method up to now, which was recently proposed by Li and Yu [27] using multi-scale CNN and three staked bidirectional gate recurrent units (BGRUs) [50]. GSN and DCRNN can extract both local contexts and long-range interdependencies. We first compare our proposed method CNNH_PSS with the three state-of-the-art methods which only can extract local contexts on CB513. The performances of these three methods and our method on CB513 are listed in Table 4. Table 4 shows that CNNH_PSS outperforms the three methods by at least 0.020 Q8 accuracy. The outperformance of CNNH_PSS over the three state-of-the-art methods which only can extract local contexts indicates that the long-range interdependencies extracted by CNNH_PSS are indeed useful for protein secondary structure prediction.

Table 3 The Q8 accuracy of CNNH_PSS

Method	CB6133	CB513
Multi-scale CNN	0.729	0.693
CNNH_PSS	*0.740*	*0.703*

The data in italic denote the best performance

Then, we compare our method CNNH_PSS to GSN and DCRNN by both CB6133 and CB513, which also can extract both local contexts and long-range interdependencies. The performances of these two methods and our method on CB6133 and CB513 are listed in Table 5. The table shows that CNNH_PSS performs better than GSN and DCRNN by at least 0.008 Q8 accuracy on CB6133 and 0.009 Q8 accuracy on CB513. GSN and DCRNN consist of CNN for local context extraction and extra models for long-range interdependencies extraction. As the extra models of the two methods are complex and time-consuming, these two methods need consume much computer resource. Trained on GTX TITANX GPU, CNNH_PSS tends to converge after only a half hour while DCRNN needs more than 24 h to converge [27]. So CNNH_PSS is almost 50 times faster than DCRNN. Although we do not known the exact running time for GSN, we know that GSN needs to be trained for 300 epochs [4] while CNNH_PSS tends to converge after training for less than 35 epochs shown by Fig. 3. It means that CNNH_PSS is almost 9 times faster then GSN. Therefore, the outperformance of our method over GSN and DCRNN further demonstrates that CNNH_PSS can not only cost less computer resource but also achieves better predicting performance.

Discussion

The advantage of our proposed method CNNH_PSS over state-of-the-art methods is that it can extract both local contexts and long-range interdependencies by using multi-scale CNN with highway. In CNNH_PSS, any two neighbor convolutional layers have a highway to deliver information from current convolutional layer to the output of the next one and each layer except the first one have convolution kernels to extract long-range interdependencies by using the information from

Table 4 The Q8 accuracy of CNNH_PSS and state-of-the-art methods containing only local contexts

Method	CB513
SSpro8	0.511
CNF	0.633
DeepCNF	0.683
CNNH_PSS	*0.703*

The data in italic denote the best performance

Table 5 The Q8 accuracy of CNNH_PSS and state-of-the-art methods containing both local contexts and long-range interdependencies

Method	CB6133	CB513
GSN	0.721	0.664
DCRNN	0.732	0.694
CNNH_PSS	*0.740*	*0.703*

The data in italic denote the best performance

previous layer. In this section, we use CNNH_PSS with 5 convolutional layers and the kernel length of 11 to introduce the process for local contexts and long-range interdependencies extraction, which is shown in Fig. 4. First, the target protein are inputted to the first layer and the convolution kernels in the first layer extract local contexts from the inputted protein. So the output of first layer contains local contexts among 11 residues. Then the information in the output of the first layer are delivered to the output of the second layer by two ways: highway between them and the convolution kernels in the second layer. Finally, the output of the second layer is the weighted sum of the information transmitted by the two way. As the convolution kernels in the secondary layer can extract relationships among 21 residues, the output of the second layer contains both local contexts among 11 residues and long-range interdependencies among 21 residues. And so on, the output of the fifth layer contains two parts. One part is the information from the output of the fourth layer by the highway between them, which contains local contexts among 11 and long-range interdependencies among 21, 31 and 41 residues. The other part is the information outputted by the convolutional kernels of current layer, which contains long-range interdependencies among 51 residues. Therefore CNNH_PSS can output local contexts among 11 and long-range interdependencies among 21, 31, 41 and 51 residues while the multi-scale CNN with the same number of convolutional layer as CNNH_PSS outputs only long-range interdependencies among 51 residues.

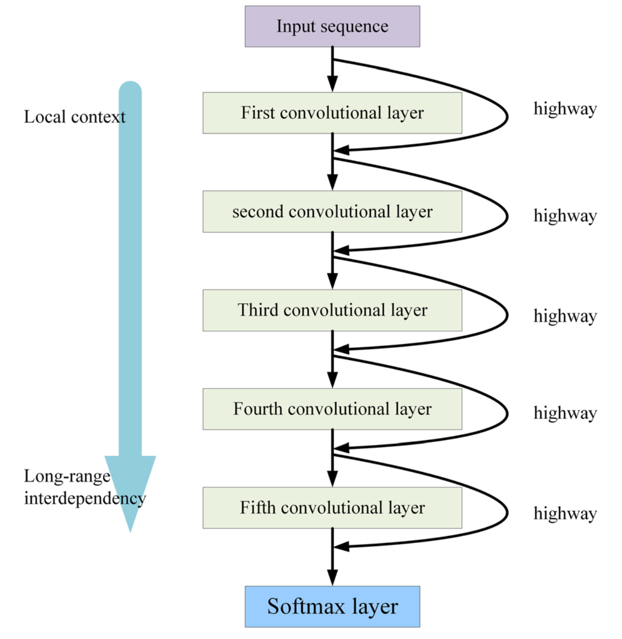
Input sequence
Local context
First convolutional layer
highway
second convolutional layer
highway
Third convolutional layer
highway
Fourth convolutional layer
highway
Long-range interdependency
Fifth convolutional layer
highway
Softmax layer

Fig. 4 Extraction process for local contexts and long-range interdependencies

In order to demonstrate the importance of learned local contexts and long-range interdependencies in protein secondary structure prediction, we show the local contexts and the long-range interdependencies learned in a representative protein PDB 154 L [51], which obtained from the publicly available protein data bank [52]. The learned local contexts and long-range interdependencies by CNNH_PSS in protein PDB 154 L are shown in Fig. 5. In Fig. 5, the five rows correspond to the predicted results by CNNH_PSS with 5 layers, that by CNNH_PSS with 3 layers, that by the multi-scale CNN with 5 layers, real secondary structures and protein sequence, respectively. The reason for why CNNH_PSS with 5 layers, CNNH_PSS with 3 layers and the multi-scale CNN with 5 layers are selected for comparison is that CNNH_PSS with 5 layers can extracted local contexts and long-range interdependencies among up to 51 residues while CNNH_PSS with 3 layers cannot extracted long-range interdependencies among more than 41 residues and the multi-scale CNN with 5 layers cannot extract local contexts. Fig. 5 shows three instances for long-range interdependencies: (1) interdependencies among 24th, 25th and 60th amino acid; (2) that between 60th and 100th and (3) that between 85th and 131th amino acid. As the number of residues covered by these three learned interdependencies is more than 31 and less than 51 residues, both CNNH_PSS with 5 layers and the multi-scale CNN with 5 layers can extract them for correct prediction them while CNNH_PSS with 3 layers cannot capture them. So both CNNH_PSS with 5 layers and the multi-scale CNN with 5 layers make correct prediction for the 24th, 25th, 85th, 100th and 131th residues while CNNH_PSS with 3 layers cannot make correct predictions for them. It validates that CNNH_PSS with more layers indeed can extract long-range interdependencies between more remote residues.

Furthermore, Fig. 4 also shows 4 instances for learned local contexts: (1) contexts from 31th to 35th residues; (2) that from 111th to 115th residues; (3) that from 146th to 149th residues and (4) that from 158th to 163th residues. Both CNNH_PSS with 3 layers and that with 5 layers can learn these four contexts so that the secondary structures of all the residues in the learned contexts can be correctly predicted. However, the multi-

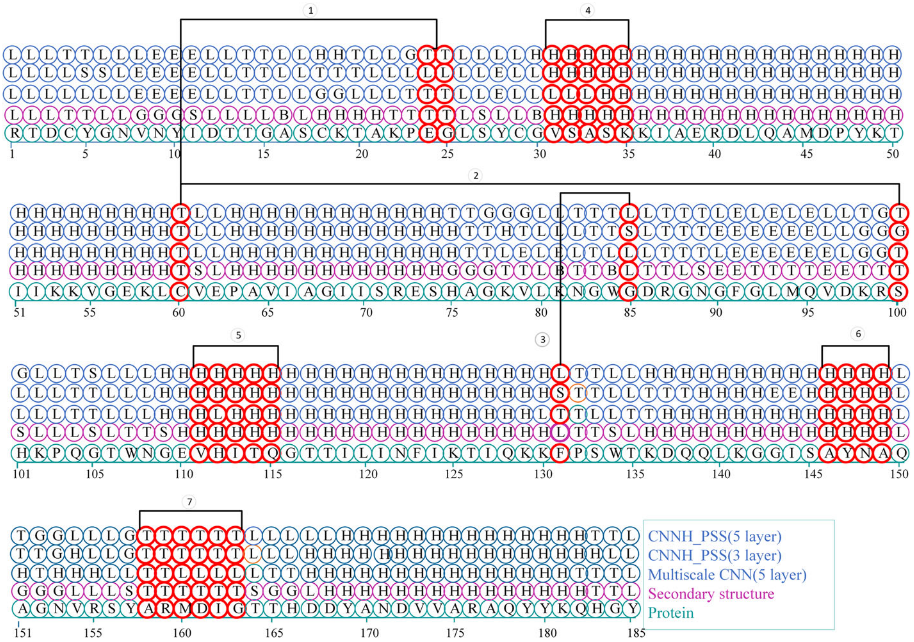

Fig. 5 Prediction results of 154 L by CNNH_PSS

scale CNN with 5 layers cannot learn these four contexts. So it cannot predict the secondary structures correctly for these residues. It validates that the highways in the CNNH_PSS indeed can be used to extract local contexts for prediction.

Conclusion

Protein secondary structure prediction is an important problem in bioinformatics and critical for analyzing protein function and applications like drug design. Several experimental methods have been proposed to determined the secondary structures for proteins, such as far-ultraviolet circular dichroism, infrared spectroscopy and NMR spectrum. However, experimental methods usually are costly and time-consuming. And the proteins with known sequence continues to outnumber the experimentally determined secondary structures. So developing computational approaches that can accurately handle large amount of data becomes increasingly urgent. However, most of these proposed methods cannot extract either local contexts or long-range interdependencies. Although GSM and DCRNN can extract both of them, they are build by combing CNN architecture and extra complex models. Yet CNNH_PSS is developed by only multi-scale CNN with highway. So comparing to GSM and DCRNN, CNNH_PSS may cost less computer resource. We evaluate CNNH_PSS on two commonly used datasets: CB6133 and CB513. CNNH_PSS outperforms the multi-scale CNN without highway on both datasets, which demonstrates that the extracted local contexts through highways are indeed useful for protein secondary structure prediction. CNNH_PSS also outperforms CNF and DeepCNF as well as SSpro8 on CB513, which cannot extract long-range interdependencies. It indicates that long-range interdependencies extracted by CNNH_PSS are useful for protein secondary structure prediction. Furthermore, CNNH_PSS performs better than GSM and DCRNN, demonstrating that CNNH_PSS can not only cost less computer resource but also achieves better predicting performance than state-of-the-art methods. We also analyze the local contexts and long-range interdependencies learned by CNNH_PSS in protein PDB 154 L and show their roles in protein secondary structure prediction. X-ray diffraction crystallography and NMR can measure the structures of proteins, so these methods can be used to calculate the distances between any two residues in a protein sequence. By analyzing the second structures of long-range residues but with short distance in space and short-range residues, we can further demonstrate the importance of long-range interdependencies and local contexts for second structure prediction. Therefore, our future works will validate the conclusions achieved in this paper by using these experimental methods.

Abbreviations

B: residue in isolated β-bridge; C: coil; CNN: convolutional neural network; CRF: conditional random field; E: extended strand in parallel and/or anti-parallel β-sheet conformation; G: 3-turn helix; GSN: generative stochastic network; H: 4-turn helix; I: 5-turn helix; NMR: nuclear magnetic resonance; PDB: Protein Data Bank; PSSM: Position Specific Score Matrix; Q3: three-category classification; Q8: eight-category classification; RNN: recurrent neural networks; S: bend; SVM: Support Vector Machine; T: hydrogen bonded turn

Acknowledgements
Not applicable.

Funding
Publication of this article was supported by the National Natural Science Foundation of China U1636103, 61632011 and Shenzhen Foundational Research Funding 20170307150024907.

Availability of data and materials
All data generated or analysed during this study are included in this published article. The datasets analysed during this study and the source code for CNNH_PSS can be accessed from our webserver of this paper http://hlt.hitsz.edu.cn/CNNH_PSS/ and http://119.23.18.63:8080/CNNH_PSS/.

About this supplement
This article has been published as part of *BMC Bioinformatics* Volume 19 Supplement 4, 2018: Selected articles from the 16th Asia Pacific Bioinformatics Conference (APBC 2018): bioinformatics. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-18-supplement-17.

Authors' contributions
JZ initiated and designed the study. RX made substantial contributions to acquisition of data, analysis and interpretation of data. JZ drafted the manuscript. RX and QL involved in drafting the manuscript or revising it. ZZ and HW provided valuable insights on biomolecular interactions and systems biology modeling, participated in result interpretation and manuscript preparation. All authors read and approved the final manuscript.

Competing interest
The authors declare that they have no conflicting interests.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Publisher's Note

Author details
[1]School Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, HIT Campus Shenzhen University Town, Xili, Shenzhen, Guangdong 518055, China. [2]Department of Computing, the Hong Kong Polytechnic University, Hung Hom, Hong Kong.

Published: 8 May 2018

References

1. Linderstrøm-Lang KU. Lane medical lectures: proteins and enzymes. California: Stanford University Press; 1952. p. 115.
2. Schellman JA, Schellman CG. Kaj Ulrik Linderstrøm-Lang (1896-1959). Protein Sci. 1997;6(5):1092–100.
3. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577–637.
4. Zhou J, Troyanskaya O. Deep supervised and convolutional generative stochastic network for protein secondary structure prediction. In: Proceedings of the 31st international conference on machine learning (ICML-14); 2014. p. 745–53.
5. Yaseen A, Li Y. Template-based c8-scorpion: a protein 8-state secondary structure prediction method using structural information and context-based features. BMC Bioinformatics. 2014;15(Suppl 8):S3.
6. Pollastri G, Przybylski D, Rost B, Baldi P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins: Structure, Function, and Bioinformatics. 2002;47(2):228–35.
7. Wang Z, Zhao F, Peng J, Xu J. Protein 8-class secondary structure prediction using conditional neural fields. Proteomics. 2011;11(19):3786–92.
8. Noble ME, Endicott JA, Johnson LN. Protein kinase inhibitors: insights into drug design from structure. Science. 2004;303(5665):1800–5.
9. Simossis VA, Heringa J. Integrating protein secondary structure prediction and multiple sequence alignment. Curr Protein Pept Sci. 2004;5(4):249–66.
10. Ashraf Y, Yaohang L. Context-based features enhance protein secondary structure prediction accuracy. Journal of chemical information and modeling. J Chem Inf Model. 2014;54(3):992–1002.
11. Pelton JT, McLean LR. Spectroscopic methods for analysis of protein secondary structure. Anal Biochem. 2000;277(2):167–76.
12. Meiler J, Baker D. Rapid protein fold determination using unassigned NMR data. Proc Natl Acad Sci U S A. 2003;100(26):15404–9.
13. Chou PY, Fasman GD. Prediction of protein conformation. Biochemistry. 1974;13(2):222–45.
14. Gascuel O, Golmard JL. A simple method for predicting the secondary structure of globular proteins: implications and accuracy. Computer Appl Biosci. 1988;4(3):357–65.
15. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol. 1993;232(2):584–99.
16. Jones TD. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 1999;292(2):195–202.
17. Cortes C, Vapnik V. Support vector networks. Mach Learn. 1995;20:273–93.
18. Scholkopf B, Burges C, Vapnik V. Extracting support data for a given task. In: Proceedings, first international conference on knowledge discovery and data mining. Menlo Park, CA: AAAI Press; 1995. p. 252–7.
19. Roobaert D, Hulle MM. View-based 3D object recognition with support vector machines. In: Proceedings of the IEEE neural networks for signal processing workshop. NJ: IEEE Press; 1999. p. 77–84.
20. Schmidt M, Grish H. Speaker identification via support vector classifiers. In: The proceedings of the international conference on acoustics, speech and signal processing, 1996. Long Beach, CA: IEEE Press; 1996. p. 105–8.
21. Drucker H, Wu D, Vapnik V. Support vector machines for spam categorization. IEEE Trans Neural Netw. 1999;10:1048–54.
22. Hua S, Sun Z. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J Mol Biol. 2001;308(2):397–407.
23. Kim H, Park H. Protein secondary structure prediction based on an improved support vector machines approach. Protein Eng Des Sel. 2003;16(8):553–60.
24. Zhou J, Lu Q, Xu R, He Y, Wang H. EL_PSSM-RT: DNA-binding residue prediction by integrating ensemble learning with PSSM relation transformation. BMC Bioinformatics. 2017;18:379.
25. Guo J, Chen H, Sun Z, Lin Y. A novel method for protein secondary structure prediction using dual-layer SVM and profiles. Proteins: Structure, Function, and Bioinformatics. 2004;54(4):738–43.
26. Bengio Y, Thibodeau-Laufer É, Alain G, Yosinski J, preprint arXiv:.1091. Deep generative stochastic networks trainable by backprop. Computer Sci. 2013;2:226–34.
27. Li Z, Yu Y: Protein secondary structure prediction using cascaded convolutional and recurrent neural networks. 2016.
28. Lawrence S, Giles CL, Tsoi AC, Back AD. Face recognition: a convolutional neural-network approach. IEEE Trans Neural Netw. 1997;8(1):98–113.
29. Yih W, Toutanova K, Platt JC, Meek C. Learning discriminative projections for text similarity measures. In: Proceedings of the fifteenth conference on computational natural language learning; 2011. p. 247–56.
30. Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification. In: Advances in neural information processing systems; 2015. p. 649–57.
31. Wang S, Peng J, Ma J, Xu J. Protein secondary structure prediction using deep convolutional neural fields. Sci Rep. 2016;6:18962.
32. Baldi P, Brunak S, Frasconi P, Soda G, Pollastri G. Exploiting the past and the future in protein secondary structure prediction. Bioinformatics. 1999;15(11):937–46.
33. Schmidler SC, Liu JS, Brutlag DL. Bayesian segmentation of protein secondary structure. J Comput Biol. 2000;7(1–2):233–48.
34. Chu W, Ghahramani Z, Wild DL. A graphical model for protein secondary structure prediction. In: Proceedings of the twenty-first international conference conference on machine learning (ICML); 2004. p. 161–8.
35. Xu R, Zhou J, Liu B, He Y, Zou Q, Wang X, Chou KC. Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach. J Biomol Struct Dyn. 2015;33(8):1720.

36. Xu R, Zhou J, Wang H, He Y, Wang X, Liu B. Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. BMC Syst Biol. 2015;9(S1):1–12.
37. Zhou J, Xu R, He Y, Lu Q, Wang H, Kong B. PDNAsite: identification of DNA-binding site from protein sequence by incorporating spatial and sequence context. Sci Rep. 2016;6:27653.
38. Wang G, Jr DR. Pisces: a protein sequence culling server. Bioinformatics. 2003;19(12):1589–91.
39. Cuff JA, Barton GJ. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins: Structure, Function, and Bioinformatics. 1999;34(4):508–19.
40. Wang Z, Zhao F, Peng J, Xu J. Protein 8class secondary structure prediction using conditional neural fields. IEEE Int Conf Bioinformatics Biomed. 2011; 11(19):3786–92.
41. Mesnil G, Dauphin Y, Yao K, Bengio Y, Deng L, Hakkani-Tur D, He X, Heck L, Tar G, Yu D, et al. Using recurrent neural networks for slot filling in spoken language understanding. IEEE/ACM Trans Audio Speech Lang Process. 2015; 23(3):530–9.
42. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems; 2013.
43. Kumar M, Gromiha M, Raghava G. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. BMC Bioinformatics. 2007;8(1):563.
44. Jones DT. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics. 2007;23:538–44.
45. Biswas AK, Noman N, Sikder AR. Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information. BMC Bioinformatics. 2010;11(1):273.
46. Ruchi V, Grish CV, Raghava GPS. Prediction of mitochondrial proteins of malaria parasite using split amino acid composition and PSSM profile. Amino Acids. 2010;39:101–10.
47. Zhao XW, Li XT, Ma ZQ, Yin MH. Prediction of lysine ubiquitylation with ensemble classifier and feature selection. Int J Mol Sci. 2011;12:8347–61.
48. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. 2001;29(14):2994–3005.
49. Srivastava RK, Greff K, Schmidhuber J. Training very deep networks. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. Advances in neural information processing systems, vol. 28; 2015. p. 2377–85.
50. Cho K, Merrienboer BV, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Computer Sci. 2014;
51. Simpson RJ, Morgan FJ. Complete amino acid sequence of embden goose (anser anser) egg-white lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology. 1983;744(3):349–51.
52. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res. 2000;28:235–42.