ATLANTIS PRESS

Advances in Biological Sciences Research, volume 5

International Workshop on Bioinformatics, Biochemistry, Biomedical Sciences (BBBS 2018)

Construction of Protein-Protein Interactions Model by Deep Neural Networks

Yuanmiao Gui[1, 2, a], Rujing Wang [1, 2, b], Yuanyuan Wei [1, c] and Xue Wang[1, 2, 3, d, *]

[1.] Institute of Intelligent Machine, Hefei Institutes of Physical Science, Chinese Academy of Sciences, HeFei City, AnHui Province, 230031, China;

[2.] University of Science and Technology of China, HeFei City, AnHui Province, 230026, China;

[3.] Institute of Technical Biology & Agriculture Engineering, Hefei Institutes of Physical Science, Chinese Academy of Sciences, HeFei City, AnHui Province, 230031, China;

[a] smalltalkman@foxmail.com, [b] rjwang@iim.ac.cn, [c] jsjwyy@126.com, [d] 181543681@qq.com

Keywords: Deep neural networks, Protein-protein interaction, Construction.

Abstract. In order to improve the effectiveness of the network prediction result of protein-protein interaction, the network prediction model of protein-protein interaction based on deep neural network has been proposed. Here, we present here a novel DPPI model with deep neural network and conjoint triad (CT) descriptors to predict protein-protein interactions only using the information of protein sequences. The best DPPI model achieved an accuracy of 97.65%, recall of 98.96% and area under the curve (AUC) of 98.51% with 10-fold cross-validation, respectively, which means that the model of predicting the interaction of protein-protein by deep neural network algorithm is accurate and effective.

1. Introduction

The identification of protein-protein interactions (PPI) plays an important role in many cellular biological processes. In the past few years, some high-throughput methods have been implemented to identify PPI [1-8], such as yeast two-hybrid screens [4] and mass spectrometric protein complex identification [5], mass spectrometry [6], protein chips [7] and hybrid approaches [8], have generated massive data, however, they are expensive and time consuming. How to use high-throughput computational approaches based on these data to dig out effective information is the important problem to be solved during life exploration process. Up to now, a large number of computer computational methods have been developed for the large-scale PPI prediction based on protein sequence, structure and evolutionary relationships in complete genomes. For example, Huang et al. [9] present a novel computational model combining weighted sparse representation based classifier and global encoding of amino acid sequence. Shen et.al [10] proposed a descriptor named conjoint triad based on support vector machine combined with a kernel function. Guo et al. [11] using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. Some methods use the structural information of proteins. Aloy et al. [12] present a method to model putative interactions on known 3D structure information and this method permits to score all possible pairs between two protein families. Zhang et al. [15] proposed a method based on three-dimensional protein structure to predict PPI. Some other methods are based on the amino acid index distribution, such as phylogenetic profiles [16], gene neighbourhood [17], gene fusion events [18, 19] and amino acid index distribution [20] to extract features.

Deep learning is now one of the most active fields in machine learning, and have received considerable attention due to their successful applications in speech and image recognition[21, 22], natural language understanding [23], decision making[24] and most recently in computational biology[25-28]. Deep learning has also successfully applied to several problems in bioinformatics [29-33]. Recently, deep neural network, a type of deep-learning algorithm, have achieved successful results for PPI prediction. Tian et al. [13] propose a method called DL-CPI (the abbreviation of Deep

Learning for Compound-Protein Interactions prediction), which employs deep neural network to effectively learn the representations of compound-protein pairs. Du et al. [26] used deep neural networks, to study the sequence-based PPI prediction, and acquired the accuracy of 97.07%, precision of 94.38%, respectively. Although, deep neural network algorithms have achieved successful results in PPI prediction, large-scale application in PPI prediction has not been found.

Here, we developed a computational approach combining DNN with CT feature extraction methods to predict genome-wide novel PPIs. Concretely, we propose DPPI model, to predict new PPI. Firstly, protein sequence features are extracted with Con joint triad (CT). Then, we input the feature vectors of both positive and negative examples to the DNN model. After hyper-parameter adjustment, we train the DNN model and get the DPPI predictor. Finally, we evaluate the prediction performance of the DPPI predictor using a set of performance metrics and compare our method with existing prediction approaches. The best model achieved an average accuracy of 97.65% for the whole training dataset.

2. Materials and Methods

2.1 Preparation of Data Set.

In our experiments, human PPI data set was select a valid BenchMark data set to evaluate the predictor. We get this data set from http://www.csbio.sjtu.edu.cn/bioinf/LR_PPI/Data.htm (HPRD, 2007 version) [34]. This data set contains 36,630 positive pairs and 36,480 negative pairs. We removed protein pairs with fewer than 50 amino acids and unusual amino acids, such as B, J, O, U, X and Z, and obtained 36,591 positive samples and 36,324 negative samples. The positive and negative samples were mixed and selected randomly 30,000 positive samples and 30,000 negative samples as training dataset for model, the remainder of which formed the raining set as hold-out test set for model validation.

2.2 Conjoint Triad (CT).

The CT method was first proposed by Shen et al. [10], they considered the properties of one amino acid and its vicinal amino acids and regarded any three continuous amino acids as a unit. The PPI information of amino acids sequence can be projected into a homogeneous vector space by counting the frequency of each triad type. All 20 amino acids are clustered into seven groups based on the dipoles and volumes of the side chains. The classification of amino acids is listed in Table 1.

Table 1. Classification of amino acids of CT coding method

Number	Amino Acids
1	Ala(A), Gly(G), Val(V)
2	Ile(I), Leu(L), Phe(F), Pro(P)
3	Tyr(Y), Met(M), Thr(T), Ser(S)
4	His(H), Asn(N), Gln(Q), Trp(W)
5	Arg(R), Lys(K)
6	Asp(D), Glu(E)
7	Cys(C)

Firstly, we replaced each amino acid in the protein sequence by the index depending on its grouping. For instance, protein sequence “RLASCTELRTLNLARN” is replaced by 5213736253242154. Then, we use a binary space (V, F) to represent a protein sequence. Here V is the vector space of the sequence features, and F is the frequency vector corresponding to V [10]. A protein sequence have been catalogued into seven classes, the size of V should be 777; thus, i =1, 2, ..., 343. If a set of data is $i^1 i^2 i^3$, the corresponding value of f is $i_1 + (i_2 - 1) \times 7 + (i_3 - 1) \times 7 \times 7$. The detailed description for vector space of a sequence features are illustrated by the following formula:

$$\left\{ \begin{array}{l} 111 = f1 \;\; 121 = f8 \ldots 177 = f337 \\ 211 = f2 \;\; 221 = f9 \ldots 277 = f338 \\ \qquad\qquad \ldots \\ 711 = f7 \;\; 721 = f14 \ldots 777 = f343 \end{array} \right\} \tag{1}$$

Schematic diagram for constructing the vector space (V, F) of protein sequence is shown in Figure 1. And thus the dimensions of a protein sequence were dramatically reduced to 7×7×7 =343. Finally, a total 686-dimensional vector has been built to represent each protein pair.

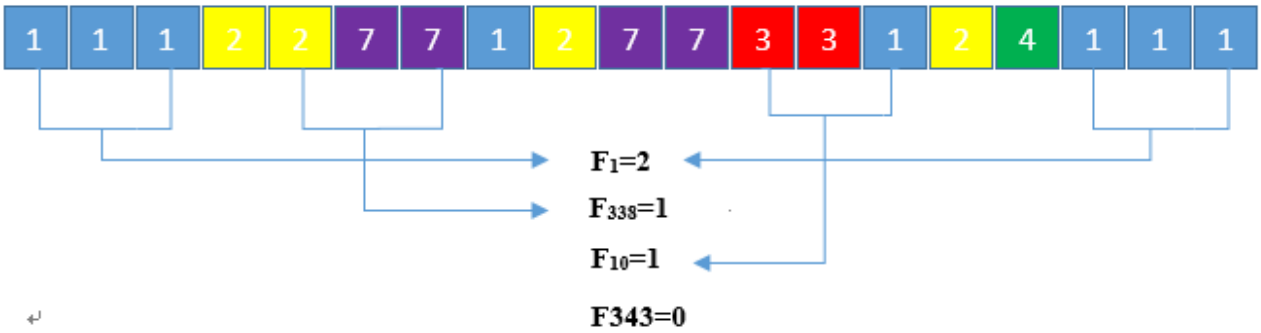

Fig. 1 Classification of amino acids of CT coding method

2.3 Deep Neural Network (DNN).

The DNN model is comprised of three components: the input layer, two or more hidden layers, and the output layer, which are similar to general artificial neural networks but differ from them in the number of hidden layers and the training procedure [14]. Figure 2 illustrates the architecture of the DNN model and its training procedure. The depth of a neural network corresponds to the number of hidden layers, and the width to the maximum number of neurons in one of its layers [35, 36]. A deep neural networks architecture is a multilayer stack of simple modules, the input layer receives data, and then data information is transformed in a nonlinear way through multiple hidden layers. Computing the average gradient and adjusting the weights accordingly, before final outputs are computed in the output layer. Mathematically, let *x* denote the input data, *z(l)* denote the input of the *l-th* layer, and *a(l)* denote the activation of the *l-th* layer, we have the following formulation:

$$z^{(l+1)} = w^l a^l + b^l \tag{2}$$

Where a^l is the connection weight matrix between the *l-th* layer and the *(l+1)-th* layer, b^l is the bias term in the *(l+1)*-th layer and *a(l+1)=f(z(l+1))*. Here *f()* means the activation function, we adopted ReLU as the activation function in the DPPI model

$$\delta(z) = \max(0, z) \tag{3}$$

ReLU mainly to solve the problem of gradient disappears, that thresholds negative signals greater than 0 will remain unchanged, less than 0 will be activated to 0 and passes through positive signal. This type of activation function allows faster learning compared to alternatives (e.g., sigmoid or tanh unit) [37]. In our study, we employed cross entropy [46] as the loss function,

$$J(w, b) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{2}\left[y_j^i \ln(t_j^i) + (1 - y_j^i)\ln(1 - t_j^i)\right] \tag{4}$$

Where n is the number of the training examples, t^i is the predicted class of the *i-th* example, and y^i is the real class of this sample.

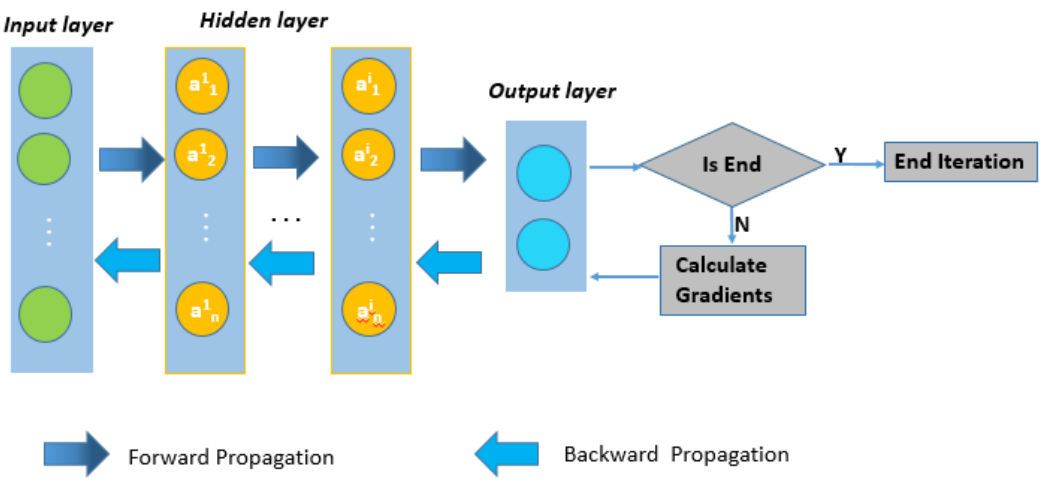

Fig. 2 Neural network architecture and training procedure

2.4. Model Architecture.

Fig. 3 shows the training framework of deep neural networks for protein-protein interactions prediction. This framework consists of the following five steps.

(1) Obtaining positive and negative set from BenchMark database.

(2) The feature of protein sequence are extracted using CT.

(3) Select randomly 30000 pairs from positive set and negative set as training set, respectively, the remainder of dataset as test set.

(4) Then, training set is used to train DNN models, and test set is used to test the performance of the models.

(5) We evaluate the prediction performance of the models using a set of performance metrics.

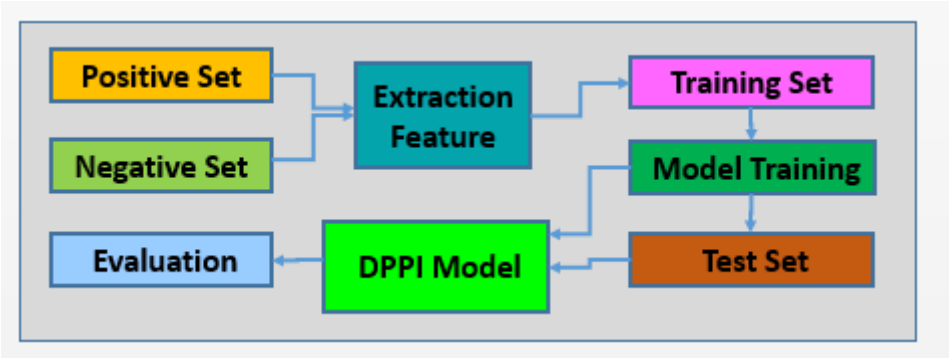

Fig. 3 The training framework of deep neural networks for protein−protein interactions prediction

2.5. Evaluation Criteria.

To measure the performance of the proposed method, we adopted 10-fold cross validation and four parameters, accuracy, recall, loss and area under the receiver operating characteristic curve (AUC). Some are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

Where true positive (TP) is the number of true PPIs that are predicted correctly; false negative (FN) is the number of true PPIs that are predicted to be non-interacting pairs; false positive (FP) is the number of true non-interacting pairs that are predicted to be PPIs, and true negative (TN) is the number of true non-interacting pairs that are predicted correctly.

3. Results

3.1. DPPI Model Construction and Parameter Optimization.

In order to choose the best hyper-parameter configuration, models with different configurations should be trained and their performance evaluated by a set of performance metrics. The learning rate and batch size can strongly impact training speed and model performance. Different learning rates are usually explored on a logarithmic scale such as 0.1, 0.01, 0.001 or 0.0001, with 0.01 as the recommended default value but it would be foolish to rely exclusively on this default value [38]. In this experiment, training was started with 0.00001 as the initial value of the learning rate, and increase the learning rate in increments of 0.00001 in each batch. The trends of the learning rate and loss function are shown Fig. 4. From Fig. 4, we can see that when the learning rate is between 0.001 and 0.01, the loss function has the fastest decline.

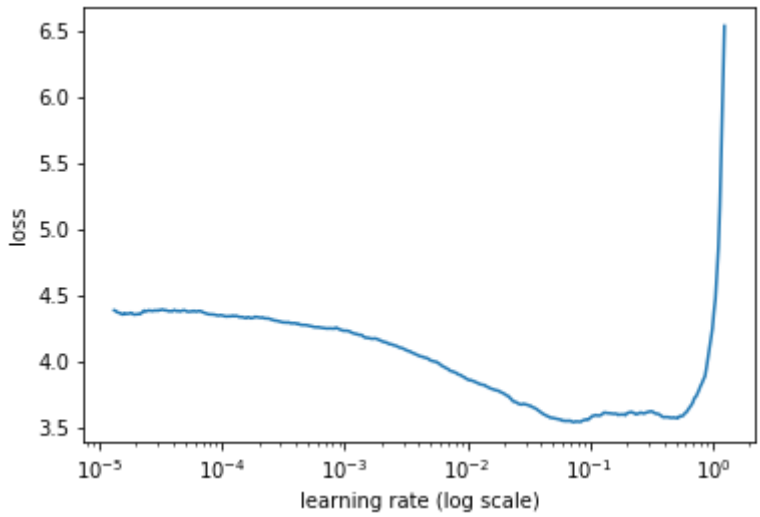

Fig.4 The trends of the learning rate and loss function

Per-parameter adaptive learning rate methods, such as RMSprop, Adagrad [39] and Adam [40], have been developed in order to appropriately adapt the learning rate per-parameter during training. Adam combines the strengths of previous methods RMSprop and Adagrad and is generally recommended for many applications. ReLU mainly to solve the problem of gradient disappears, ReLU allows faster learning compared to alternatives [37]. We found that sigmoid and tanh slow down the gradient descent by experiments, so we choose ReLU and ELU as activate function. Table 2 provides the recommended hyper-parameters that are chosen by a large number of experiments. The recommended parameters such as the learning rate, which is exploring different values while keeping all other hyper-parameters constant.

Table 2. Recommended parameter of DNN-PPI models in the experiments

Name	Range	ecommendation
Learning rate	0.01, 0.001, 0.0001, 0.00001	0.001
Batch size	16, 32, 64, 128, 256	128
Activation function	ReLU, tanh, sigmoid, softmax, ELU	ReLU, ELU
Dropout rate	0.5, 0.6, 0.7	0.6
Cost function	cross-entropy	cross-entropy
Optimizer	Adadelta, RMSprop, Adam	Adam
Depth	2, 3, 4, 5, 6	3, 4
Width	128, 256, 1024, 2048	256

3.2. Performance of DPPI Model.

The performance of DPPI model was shown in Table 3, the accuracy, recall, AUC and loss values of DPPI model are 97.65%, 98.96%, 98.51% and 26.69%, respectively. The average accuracy, recall, AUC and loss values are 97.11%, 98.91%, 98.35 and 26.52, respectively. Sun et al. [25] applied Stacked auto-encoder (SAE) with CT to study sequence-based human PPI predictions, which yielded the accuracy of 94.52%. Sun et al. is the first to use a deep-learning algorithm for sequence-based PPI prediction, and they achieved prediction performance that surpassed previous methods. Compared to Sun's results, our accuracy metrics are outstanding, which indicate that DPPI model is successful in predicting PPIs.

Table 3. The best prediction performance of the DPPI model.

	Accuracy(%)	Recall(%)	AUC(%)	Loss(%)
DPPI	97.65	98.96	98.51	26.69
Average	97.11	98.91	98.35	26.52

AUC: Area under the receiver operating characteristic curve

3.3. Comparison with Existing Methods.

Recently, a large number of computational methods have been proposed for predicting PPIs due to the continually development of the high-throughput technologies. Here, we use the same human data set, and compare our experimental result with those of Shen's work [10], Zhang's work [44], Huang's work [45] and so on, compared to seven other individual methods, Table 4 shows the results performed by other methods and we can see that the accuracies obtained by these methods are between 83.90% and 97.65%. Our work achieves the best performance, none of these methods gets higher accuracy than that of our proposed method, which yielded the accuracy of 97.65%. Considering these comparisons, it is demonstrated that our model can improve the prediction accuracy compared with the current methods.

Table 4. Comparison of different methods performance on the human dataset.

References	Method	Average Accuracy
Shen's work [10]	SVM+CT	0.8390
Zhang's work [44]	CS+SVM	0.9410
Huang's work [45]	DCT+SMR	0.9630
Sun's work [25]	SAE+CT	0.9452
You's work [41]	ELM	0.8480
Guo's work [42]	SVM	0.9067
Tian's work [43]	DNN	0.9320-0.9380
Our model	DNN+CT	0.9765

4. Summary

Deep neural network algorithms have been used in many fields and obtained a series of achievements. However, these powerful methods it is rarely used in PPI prediction. Thus, in this work, we used DNN with a widely-used protein sequence coding methods CT, to study human PPI. From Table 4, we know that the performance of our model may produce a stable prediction model with higher accuracy than other similar methods. Our work yielded the accuracy of 97.65%, none of these methods gets higher accuracy than that of our proposed method. In this work, the first contribution to the good performance of DPPI model is that we adjust various parameters such as learning rate, batch size, activation function and dropout rate etc., and provides the recommended hyper-parameters that are chosen by a large number of experiments. Then, the application of DNN can learn an internal distributed feature representation automatically from the data. The experimental result shows that our model is superior in prediction accuracy improvement in comparison with existing prediction method.

References

[1]. Uetz P, Giot L, Cagney G, Mansfield T A, et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature.Vol.403 (2000) No. 6770, p. 623-627.

[2]. La Count DJ, Vignali M, Chettier R, et al. A protein interaction network of the malaria parasite Plasmodium falciparum. Nature. Vol. 438(2005) No.7064, P. 103-107.

[3]. Parrish JR, Yu J, Liu G, et al. A proteome-wide protein interaction map for Campylobacter jejuni. Genome Biol. Vol. 8(2007) No.7, p.R130.

[4]. Fields S, Song O. A novel genetic system to detect protein protein interactions. Nature. Vol. 340(1989) No.6230, P. 245-246.

[5]. Ho Y, et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature. Vol. 415(2002) No. 6868, p.180-183.

[6]. Gavin AC, Bosche M, Krause R, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature. Vol. 415(2002), P. 141-147.

[7]. Heng Z, Metin B, Rhonda B, et al. Global Analysis of Protein Activities Using Proteome Chips. Science. Vol. 293(2001) No. 5537, p. 2101-2105.

[8]. Tong AHY, Becky D, Giuliano N, et al. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science. Vol. 295(2002) No. 5553, p. 321-324.

[9]. Huang YA, You ZH, Chen X, et al. Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding. BMC Bioinformatics. Vol. 17(2016) No. 1, p. 1-11.

[10]. Shen JM, Zhang J, Luo XM, et al. Predicting protein-protein interactions based only on sequences information. Proc. Natl Acad. Sci. Vol. 104(2007) No.11, P. 4337-4341.

[11]. Guo YZ, Yu LZ, Wen ZN, et al. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic Acids Research. Vol. 36(2008) No.9, p. 3025-3030.

[12]. Aloy P, Russell RB. Interrogating protein interaction networks through structural biology. Proc. Natl Acad. Sci. Vol. 99(2002) No. 9, p. 5896-5901.

[13]. Tian K, Shao MY, Wang Y, et al. Boosting compound-protein interaction prediction by deep learning. Methods. Vol. 110(2016) p. 64-72.

[14]. Spencer M, Eickholt J, Cheng J, A deep learning network approach to ab initio protein secondary structure prediction. IEEE/ACM Trans. Comput. Biol. Bioinf. Vol. 12(2015) No. 1, p.103-112.

[15]. Zhang QFC, Petrey D, Deng L, et al. Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature. Vol. 490(2012) No. 7421, p.556-60.

[16]. Pellegrini M, Marcotte E M, Thompson M J, et al. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. Vol. 96(1999) No. 8, p. 4285-4288.

[17]. Overbeek R, Fonstein M, D'Souza M, et al. Use of contiguity on the chromosome to predict functional coupling. Silico Biol. Vol. 1(1999) No. 2, p.93-108.

[18]. Marcotte EM. Detecting protein function and protein-protein interactions from genome sequences. Science. Vol. 285(1999) No. 5428, p. 751-753.

[19]. Enright AJ, Iliopoulos I, Kyrpides N C, et al. Protein interaction maps for complete genomes based on gene fusion events. Nature. Vol. 402(1999) No. 6757, p. 86-90.

[20]. Zhang SW, Hao LY, Zhang TH, et al. Prediction of protein-protein interaction with pairwise kernel Support Vector Machine. International Journal of Molecular Sciences. Vol.15(2014) No. 2, p. 3220-3233.

[21]. Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process Magazine. Vol. 29(2012) No. 6, p.82-97.

[22]. Krizhevsky A, Sutskever I, Hinton GE, et al. Imagenet classification with deep convolutional neural networks. International Conference on Neural Information Processing Systems. Vol. 60(2012) No. 2, p. 1097-1105.

[23]. Lipton Z C, Berkowitz J, Elkan C, et al. A critical review of recurrent neural networks for sequence learning. Computer Science 2015; arXiv preprint arXiv:150600019.

[24]. Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. Nature. Vol. 529(2016) No. 7587, p.484-489.

[25]. Sun TL, Zhou B, Lai HH, et al. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. Bmc Bioinformatics. Vol. 18(2017) No. 1, p. 277-285.

[26]. Du XQ, Sun SW, Hu CL, et al. DeepPPI: Boosting prediction of protein-protein interactions with deep neural networks. Journal of Chemical Information & Modeling. Vol. 57(2017) No. 6, p. 1499-1510.

[27]. Angermueller C, Pärnamaa T, Parts L, et al. Deep learning for computational biology. Molecular Systems Biology. Vol. 12(2016) No. 7, p. 878-894.

[28]. Chicco D, Sadowski P, Baldi P, et al. Deep Autoencoder Neural Networks for Gene Ontology Annotation Predictions. Acm Conference on Bioinformatics. Vol. 21(2014) No. 2, p. 533-540.

[29]. Spencer M, Eickholt J, Cheng J, et al. A deep learning network approach to ab initio protein secondary structure prediction. Computational Biology & Bioinformatics IEEE/ACM Transactions on 2015. Vol. 12 (2015) No. 1, p.103-112.

[30]. Lena PD, Nagata K, Baldi PF, et al. Deep spatio-temporal architectures and learning for protein structure prediction, Advances in Neural Information Processing Systems. Vol. (2012) No.1, p. 512-520.

[31]. Chicco D, Sadowski P, Baldi P, et al. Deep autoencoder neural networks for gene ontology annotation predictions. Acm Conference on Bioinformatics. Vol. 21(2014) No.2, p. 533-540.

[32]. Sheng W, Peng J, Ma JZ, et al. Protein secondary structure prediction using deep convolutional neural fields. Scientific Reports. Vol. 6(2016) No.18962.

[33]. Leung MK, Xiong HY, Lee LJ, et al. Deep learning of the tissue-regulated splicing code. Bioinformatics. Vol. 30(2014) No. 12, p. i121-i129.

[34]. Pan XY, Zhang YN, Shen HB, et al. Large-Scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features. Journal of Proteome Research. Vol. 9(2010) No. 10, p. 4992-5001.

[35]. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. Vol. 313(2006), p. 504-507.

[36]. Hinton GE, Osindero S, Teh YW, et al. A fast learning algorithm for deep belief Nets. Neural Computation. Vol. 18(2014) No. 7, p. 1527-1554.

[37]. Glorot X, Bordes A, Bengio Y, et al. Deep Sparse Rectifier Neural Networks. International conference on artificial intelligence & statistics. Vol. 15(2011), p. 315-323.

[38]. Bengio Y. Practical Recommendations for Gradient-Based Training of Deep Architectures. 2012; arXiv:1206.5533v2.

[39]. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from over fitting. Journal of Machine Learning Research. Vol. 15(2014) No.1, p. 1929-1958.

[40]. Kingma D, Ba J. Adam: a method for stochastic optimization. Computer Science 2015; arXiv:1412.6980V8.

[41]. You ZH, Li S, Gao X, Luo X, et al. Large-scale protein-protein interactions detection by integrating big biosensing data with computational model. Biomed Res Int. Vol. 2014 No.2, p. 598129.

[42]. Guo YZ, Li ML, Pu XM, et al. PRED_PPI: a server for predicting protein-protein interactions based on sequence data with probability assignment. Bmc Research Notes. Vol. 3(2010) No. 1, p. 145-152.

[43]. Tian K, Shao MY, Wang Y, et al. Boosting compound-protein interaction prediction by deep learning. Methods. Vol. 110(2016), p. 64-72.

[44]. Zhang YN, Pan XY, Huang Y, et al. Adaptive compressive learning for prediction of protein-protein interactions from primary sequence. Journal of Theoretical Biology. Vol. 283(2011) No. 1, p. 44-52.

[45]. Huang YA, You ZH, Gao X, et al. Using weighted sparse representation model combined with discrete cosine transformation to predict protein-protein interactions from protein sequence. Biomed Res Int. Vol. 2015(2015), p. 902198.

[46]. Golik P, Doetsch P, Ney H. Cross-entropy vs. squared error training: a theoretical and experimental comparison.in: Interspeech. Vol.(2013), p. 1756-1760.