ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp. 1345 -1352
doi: 10.18178/wcse.2017.06.233

Feature Extraction Based on Stacked Auto Encoder for Protein Secondary Structure Prediction

Yehong Chen[1, 2], Jinyong Cheng[1+] and Yihui Liu[1]

[1]Institute of Intelligent Information Processing, [2]School of Printing & Packaging,

Qilu University of Technology

Jinan, China

Abstract. In this paper, a novel sequence feature extraction method based on the deep learning network is proposed for protein secondary structure prediction. This deep learning architecture, mainly composed of two layers stacked auto encoder and a fully connected softmax classifier. Position-specific scoring matrix (PSSM) profiles are used as raw data for feature extraction. The stacked auto encoder structure could learn the second order feature parameters by the importance on massive PSSM profiles of polypeptide unaware of secondary structure, which does improve the performance of the encoder in general. Compared to the representation of original PSSM profiles, the extracted feature not only reflects the evolutionary information, but also the sequence interaction of residues. Finally, the extracted features are fed into a fully connected softmax layer as a classifier for the secondary structure prediction. The experimental results indicate that this method can achieve an overall accuracy (Q3) above 78% on 25PDB. This is comparable with that of the art-of-the-state PSSM+SVM methods, at the same time, in relatively short prediction period.

Keywords: Sparse auto-encoder, Stacked auto encoder, Protein secondary structure prediction, Deep learning neural network.

1. Introduction

Protein structure prediction is very critical for analyzing protein function and its applications such as drug design [1]. Furthermore, protein secondary structure prediction plays an important role in the further three-dimensional structure analysis. It is widely accepted that the amino acid sequence (AAs) contains sufficient information to determine the three dimensional structure of a protein, however, it is extremely difficult to directly predict protein structure based on a whole sequence of amino acid residues [2]. Hence, prediction of protein secondary structure from sequence-known protein by a fast computational method is very fundamental and challenging [3]. In the past decade, numerous efficient methods had been proposed, such as, methods based on probabilistic model (HMM) [4, 5], dynamic Bayesian networks (DBN) [6], or machine learning-based methods mainly including neural networks (NN) [5, 7] and support vector machines (SVM) [8, 9, 10]. However the accuracy of prediction was below 80% [11], which has not been improved in the past decades. In general, the feature extraction of amino acid sequence information is a key step to improve the performance of predicting protein secondary structure. Artificial designed statistical features based on contents, such as the frequency of each AA（amino acid) in given proteins [8], normally can only achieve low prediction accuracy since they ignore the sequential order of AAs and the relationships among the distant AAs. According to Q. Dai et al. [12], the position based features and the contents based features

+ Corresponding author. Tel.: + 15154115705.
E-mail address: chenyh @spu.edu.cn

must work closely to make significant and complementary contributions. To date, more and more feature extraction and feature selection methods were proposed to offer diversity and inherence features for this prediction problem [13, 14, 15, 16].

The position-specific scoring matrix (PSSM) [13] which encodes evolutionary information as the profile of the protein sequence has been proven most helpful for building prediction model by SVM [13,17,18]. However, when low-homology datasets with pairwise sequence identity below 40% were tested, these methods were not effective any more. For instance, the reported overall accuracy for the widely used dataset 25PDB whose sequence homology is about 25%, were about 60~70% only [12,18]. The compound pyramid model adopts a gradually refining, multi-hierarchical configuration, in which the layers focus on independent functions, so that this model gets the higher prediction accuracy comparatively than before [7, 10, 11, 19], however, beyond 80% of Q3 is difficult. Recently, much attention has been paid to use deep learning network to predict protein secondary structure, including Porter 4.0, SCORPION, SPIDER 2, and DeepCNF [20, 21, 22, 23, 24, 25, 26]. Deep learning networks are the revolutionary development of neural networks, and the results show that can create more powerful predictors. The deep neural networks also can give the promise of self-taught feature learning from massive amounts of unlabeled data, which means more inherence information for protein structure could be encoded into the descriptors. Sparse auto encoder is developed from neural network for special self-taught feature learning. Using sparse auto-encoder for unsupervised learning is not new, however, for protein secondary structure protection is rarely in literatures at present.

In theory, the AAs of a whole protein will determine the three dimensional structure of a protein. In practice, a secondary structure prediction problem is usually formulated under the concept of slide window [7, 20]. By far, still no researchers had claimed that how long the window size should be enough for the prediction of the center position. The longer the sequence is, the more information is involved, however, at the same time more interferences are unavoidable. Similar machine learning approaches to secondary structure prediction have reported success using a variety of window sizes from 13 to 21 [7,20]. The window sizes from 11 to 25 were tested in [20], and the authors found that the average evaluation scores generally increased to a window size of 19, and then sharply dropped off for windows larger than 20.

In this research, a long window feature extraction method is proposed based on the deep learning architecture to extract the new presentation of protein sequences. The extracted new presentation is expected to reflect reasonably more related information of AAs. Compared to the original PSSM profile, the new presentation of our proposed sequences features extraction (SFE) method can obtain comparable results than before as well as relatively short-term prediction period.

2. Feature Extraction Based on Deep Learning Architecture

In this study, a sequences feature extraction (SFE) method is proposed for the prediction of protein secondary structure. It is a deep learning architecture that takes advantages of self-taught feature learning. In this section a clear explanation about the data and methods is given to every detail.

2.1. Datasets

Datasets used in this paper include: RS126 [27] that comprises 126 protein of about 25% sequence identity; 25PDB [28] that comprises 1672 proteins of about 25% sequence identity; and CB513 [19] that comprises 513 proteins of less than 25% sequence identity. All above protein datasets are encoded in PSSM by PSI-BLAST [30].

PSSM introduces evolution information of protein for prediction model learning. The theoretical basis is that the most reliable way to predict protein secondary structure is homologous with a known structural protein [13]. Each protein sequence is used as a seed to search and align homogenous sequences from NCBI's NR database (ftp://ftp.ncbi.nih.gov /blast/db/nr) by the iterative databank-searching tool BLAST (PSI-BLAST) (http://blast.ncbi.nlm.nih.gov/Blast.cgi) with three iterations and a cut off of E-value 0.001. In our experiments, BLOSUM62 Substitution Matrix is the adopted measures as a score matrix to reflect the similarity among the amino acids. Finally, the obtained PSSM profile of a protein sequence is an $L \times 20$ matrix, in which L is the length of the protein instance.

A secondary structure prediction is usually formulated under the concept of the sliding window. The secondary structure of the center position can be predicted by the information extracted from amino acid sequence inside the window. The head-part and tail-part of one protein sequence is extrapolated by reflecting across edge method so that these amino acids located at head and tail parts of protein sequences can be predicted by the same window size. Under the concept of the sliding window, a whole protein sequence with L amino acid is split in L non-overlapping intervals of n base pairs for the study. n is the window size. Protein secondary structure is assigned from the experimentally determined tertiary structure by Dictionary of Secondary Structure of Proteins (DSSP) [14]. The defined contents of the secondary structures using the DSSP file of the proteins have eight classes: H (α-helix), G (310-helix), I (Л -helix), E (β-strand), B (β-bridge), T (turn), S (bend) and C (rest random coil). The H, E and C can be denoted as α -helix, β -strand and all other elements including coil. This strategy has been widely accepted in secondary structure prediction problems [7, 12, 28].

2.2. Self-taught Feature Learning Based on the Sparse Auto - encoder

In order to improve the accuracy of the prediction, one method is to get more labeled data, but this will be expensive. Sparse auto encoder is developed from neural network for special self-taught feature learning [29]. So we try to obtain and learn generative features from massive amounts of unlabeled data.

2.2.1 Sparse Auto Encoder

Given a set of unlabeled training examples {x (1), x (2), x (3),……}, where x(i)∈Rn. x is a training sample in n dimension. i is from 1 to m and m is the number of the samples. The architecture of an auto encoder is a three layer neural network, yet the output layer nodes are set as the same as the input layer nodes. It is an unsupervised learning algorithm. Figure 1 shows the theoretical structure of a sparse auto encoder.

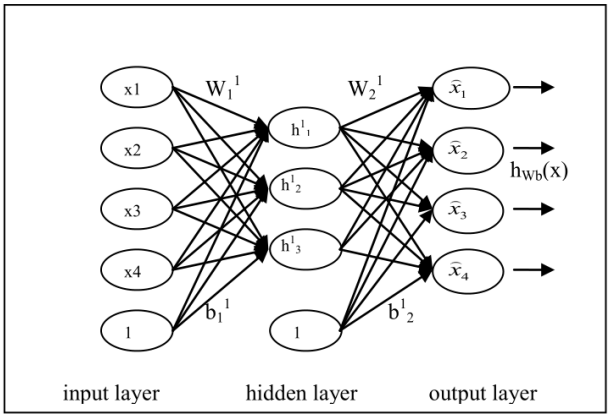

Fig. 1: The theoretical illustration of a sparse auto encoder with 3 layers， from left to the right being: input layer, hidden layer and output layer. Specially, output units are equal to the input units, as $\hat{x}_i = x_i, i = 1,2,3,4,\cdots$. In this paper, the input layer is L×20 units for PSSM patch; The number of the hidden layer is 400 and the output layer is the same as the input layer.

As same as the neural network with one hidden layer, auto encoder will find a way of defining a complex, non-linear form of hypotheses hW,b(x), with parameters W, b that are fitted to training data.

$$h_{w,b}(\mathrm{x}) = f(W^T x) = f(\sum_{\mathrm{i}=1}^{n} W_i x_i + b) \tag{1}$$

W is a *s2*×*s1* transfer matrix. *s1* is the number of the nodes in the first layer and *s2* is the number of the nodes in the second layer. Specially, W_1, b_1 is parameters transfer input values of n dimensions into the hidden layer (2). The activation function of the hidden layer is the sigmoid function, which can map all values appeared in hidden unit to [0,1] (3). z_j is the values appeared in the j-th hidden unit.

$$h_{\mathrm{j}} = \sum_{\mathrm{i}=1}^{n} W1_{ji} x_i + b1_j; \tag{2}$$

$$f(h) = \frac{1}{1+\exp(-h)} \tag{3}$$

In figure 1, W_1,b_1 are parameters mapping input data into hidden nodes values z, and W_2,b_2 are transfer parameters mapping hidden values f(h) into output values $\hat{x}$. This network applies back propagation, by setting the output layer values to be equal to the inputs. I.e., it uses $y^{(i)} = x^{(i)}$. The auto encoder uses the back propagation to perform gradient descent exactly on the objective $J_{sparse}(W,b)$ that contains 3 terms: the squared error term, the weight decay term, and the sparsity penalty which imposes a sparsity constraint on the hidden units.

$$J(W,b) = \left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\|h_{W,b}(x^{(i)}) - y^{(i)}\right\|^2\right)\right] + \frac{\lambda}{2}\sum_{k,j}^{n,s2}(W_{jk})^2 \tag{4}$$

$$J_{sparse}(W,b) = J(W,b) + \beta\sum_{j=1}^{s2} KL(\rho \,\|\, \hat{\rho}_j) \tag{5}$$

In (4), the weight decay parameter λ controls the relative importance of the two terms. Parameter β controls the relative importance of the two terms in (5). A sparsity parameter ρ is typically a small value close to zero (say ρ=0.05). $\hat{\rho}_j$ is the average activation of hidden u.t over all train dataset . When $\hat{\rho}_j = \rho$, $KL(\rho \,\|\, \hat{\rho}_j)$ equal to 0. In other words, we would like the average activation of each hidden neuron to be close to 0.05. Training the auto encoder is the procession of calculating the optimal W, b, and by minimizing $J_{sparse}(W,b)$. In practice, gradient descent algorithm L-BFGS usually can work fairly well and is used for this optimization problem in this paper.

2.2.2 Features Extraction Based on Two Stacked Autoencoders

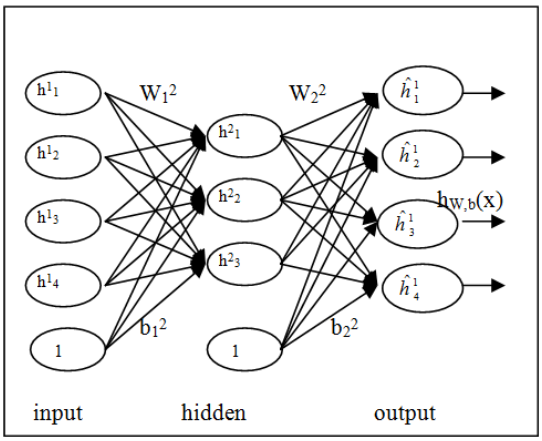

Fig. 2: The training of the second layer of a stacked autoencoder. The input raw data is from the hidden layer of the first autoencoder, and that data also link to the output layer as: $\hat{h}_i = h_i, i = 1,2,3,4,\cdots$.

From observation, one knows the adjacent secondary structure has obvious interdependence. A stacked auto encoder used in this paper compose of two layers, auto encoders in which the outputs of the first layer are wired to the inputs of the successive layer. The greedy layerwise approach is the way we used here to train our stacked autoencoder. By slide window concept, PSSM profiles are the raw data feed to the input layer of the first auto encoder, and the output layer of this autoencoder is set as the same as the input(Figure 1).Through training this auto encoder, the hidden layer recover the first order feature h1 (1~400) of the L windows PSSM, and the output layer is throw away.Then the first order features h1 () extracted features is used as the raw data to train the second autoencoder(Figure 2), and after training, the second order feature h2 () will appear at the hidden layer of the second layer, and the output layer which is set equal to the input is also thrown after training.

The features from the stacked autoencoder can be used for classification problems by feeding them to a softmax classifier. These learned features could encode the probability distribution of longer polypeptides through learning massive unlabeled polypeptides, and this probability distribution can be most consist with natural scenes [30].

2.3. Classification Layer

Finally, we combine these two trained auto encoder layers together to form a stacked autoencoder with 2 hidden layers and a final softmax classifier layer capable of classifying the three types of protein secondary structure: H, E, C (Figure 3). The softmax function is considered as the multi-class generalization of the logistic sigmoid function (3). To adjust the weights for training, a back propagation algorithm called fine tuning is used which is an application of the gradient descent algorithm. According to fine tunning, the weights in the network will move along the negative gradient of the response logistic sigmoid function to get the fastest way for adjustment. From our experiments, the prediction accuracy results got great improvement after fine tunning procession.

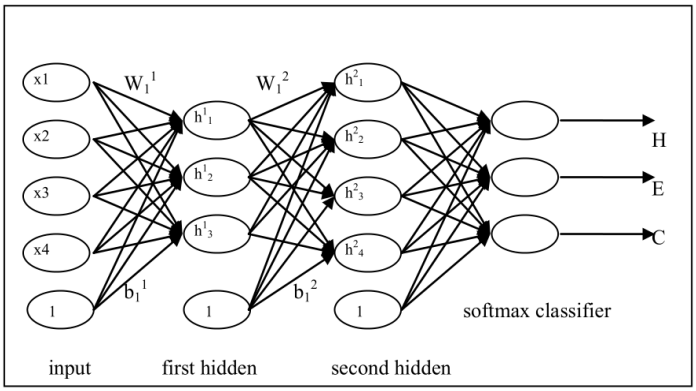

Fig. 3: The whole deep learning network with a fully connected softmax classification layer. The units of the input layer are L×20 for L-polypeptides in PSSM patch. The first hidden layer (400 notes) is the first order features and the second layer (400 notes) is the second order features. The output layer is a fully connected classification layer with three output nodes.

3. The Experimental Results and Discussions

Our experiments are implemented in MATLAB 2014a, which runs on an Intel® Core™ i7-4790 3.60 GHz CPU with 32 GB RAM. The generated PSSM matrix of RS126, CB513, and 25PDB within the search scope of nr datasets are captured as original protein sequence input data, as well as the protein secondary structure of the three states are respectively H [1, 0, 0], E [0, 1, 0] and C [0, 0, 1]. The concept of sliding window is used for pilling up these PSSM as protein sequence samples, and the window sizes used in this experiment are from 13 up to 35. Protein secondary structure prediction is usually evaluated by Q3, which measures the percent of residues for which 3-state secondary structure is correctly predicted [19]. In this work, Q3 is used to evaluate the top layer softmax classifier. Every evaluation result is the mean of Q3 value over three times practical running under five fold partition, actually 4 in 5 of proteins are randomly picked up as training data, and the remaining are all the test samples.

In general, unsupervised learning favors a large amount of training data, so that the nature of the data can be better captured, however, the bigger the dataset is, the longer is the training time. The hidden layer units' number is 400 in both hidden layers of the stacked auto encoder, and the other parameters of a sparse auto encoder in (4) are set as: $\rho = 0.035$; $\lambda = 3e-3$; β=5 . Polypeptides come from CB513, RS126 and 25PDB by the slide window method are used for building prediction model. In our proposed SFE method a new representation of these polypeptides should be extracted as the output of a stacked two layers auto encoder. Compared to the representation of PSSM, the representations of SFE features take into account both local variants and global variation in the balance. It not only reflects the evolutionary information, but also the sequence interaction of residues. Hence SFE is a powerful feature extraction method for secondary structure prediction.

SVM is thought as the best shallow network for classification problem, however, the time consumption will increase very much when the dataset is becoming more and more big. In our experiments, the softmax classifier with fine-tuning as the shallow network achieves the comparable accuracy and with a relatively short production period.

Table1: Experimental results of prediction accuracy Q_3 by SFE method on different datasets.

Train Data	Window size	Prediction of H; E; C (%)	Accuracy (Q3) (%) After / before (Fine tuning)
RS126	13	76.92; 64.88; 70.06	71.16 / 62.03
	23	76.79;63.23;74.14	72.87 / 66.88
CB513	13	78.29;69.29;76.33	75.69 / 66.76
	23	79.22;67.51;76.85	75.86 / 64.25
25PDB	13	81.11;74.82;76.43	77.57 / 66.41
	23	81.51; 74.45; 77.39	78.11 / 65.48
	27	82.59;74.41;76.19	77.79 / 64.23
	35	81.5 ; 74.9 ; 76.6	77.83 / 63.92

Table 1 shows the performance based on SFE features. Variable H, E, C and Q3 represent the accuracy of Helix, Sheet, Coil, and overall three classes. Accuracy (Q3) is calculated under 5 fold cross-validation.

When the sliding window of 23 is used for PSSM profile, the best prediction accuracy Q3 for each dataset from our comparative experiments are: around 72% for RS126, around 75% for CB513 and around 78% for 25pdb. To notice, on 25PDB the prediction accuracy of the Helix H structure is close to 81% and of Sheet E structure is passing 74% by using SFE features. It shows that SFE is powerful on recognizing these two main structures in longer polypeptides. Model trains and testing is relatively fast. The more important thing is that SFE is more suitable for a big protein dataset, that is proved by these experiments. Table 1 also recover that the fine-tuning is a very important part of the training of this deep learning network.

Our proposed SFE close to the method proposed in [20]. They both learn features in the input data and initializes weights for the next network layers. However, [20] uses RBMs (restricted Boltzmann Machines) for initializing the weights in a DN (Deep learning (belief) networks) via training approach, while SFE uses stacked sparse auto encoders for initializing the weights in a deeper learning network. In paper [22], an integration of Conditional Random Fields (CRF) and shallow neural networks achieves 84% Q3 on the CASP and CAMEO test proteins. The most obvious improvement of SFE is that the window size used in SEF is 23, while 11 in [22] and 17 in [20]. In paper [20], three kinds of features: the amino acid residues (RES), the PSSM information (PSSM), and the Atchley factors (FAC) are selected as the input profiles. In paper [22], there are 42 input features for each residue, 21 from PSSM and the other 21 from the primary sequence. The model [21] based on a deep supervised generative stochastic network (GSN), using PSSM, protein sequence of amino-acid residues, and start and end positions of the protein sequence. In our research, we only use PSSM profile for SFE. In the future work, more information will be added to improve the performance of SFE.

The contributions of SFE method are as follows:

1. A new feature extraction method by a stacked auto encoder is proposed to discover an improved second order presentation for protein secondary structure prediction.

2. The longer distance interactions up to 23 are encoded into the description of protein polypeptides based on PSSM. This representation is with the potential power for learning more complex model in the field of protein secondary structure prediction.

3. The proposed method achieves comparable prediction accuracy (Q3) than art-of-the-state PSSM+SVM method.

Thus, it is a valuable method to predict protein structure, particularly for low-homology amino acid sequences and may at least play an important complementary role to existing methods. We highlight that the deep learning is an emerging technique in this area, and study of deep learning network for protein secondary structure prediction is our main direction in the near future.

4. Acknowledgements

The research work is supported by the National Natural Science Foundation of China (Grant No. 61375013), and Natural Science Foundation of Shandong Province (ZR2013FM020), China.

5. References

[1] J.M. Thorton. From Genome to Function. Science.2001, vol.292 (5524): 2095-2097.

[2] P.H. Raven and G.B. Johnson. How Scientists Think. WCB/McGraw-Hill, 1997

[3] Shengli Zhang, Yunyun Liang and Xiguo Yuan. Improving the prediction accuracy of protein structural class: Approached with alternating word frequency and normalized Lempel–Ziv complexity. journal of theoretical biology.2014,vol. 341:71-77.

[4] J. Martin, J.F. Gibrat and F. Rodolphe. Analysis of an optimal hidden Markov model for secondary structure prediction. BMC Structural Biology. 2006, vol. 6 (25).

[5] J. Kunal. Prediction of Ubiquitin Proteins using Artificial Neural Networks, Hidden Markov Model and Support Vector Machines. SilicoBiology.2007,vol. 7 (6): 559-568.

[6] X.-Q. Yao, H. Zhu and Z.-S. She. A dynamic Bayesian network approach to protein secondary structure prediction. BMC Bioinformatics. 2008,vol. 9 (49).

[7] W. Qu, H. Sui, B. Yang and W. Qian. Improving protein secondary structure prediction using a multi-modal BP method. Computers in Biology and Medicine. 2011, 41 (10): 946-59.

[8] C. Chen, Y. Tian, X. Zou, P Cai. and J. Mo. Prediction of protein secondary structure content using support vector machine.Talanta. 2007,vol. 71(5):2069-73.

[9] S. Teng, A.K. Srivastava and L. Wang. Sequence feature-based prediction of protein stability changes upon amino acid substitutions . BMC Genomics. 2010, vol.11(Suppl 2):S5.

[10] Haifeng Sui. predicting protein secondary structure using a mixed-modal SVM method in a compound pyramid model. 2011,vol. 24(2):304–313.

[11] Bingru Yang, Wei Hou, Zhun Zhou and Huabin Quan. KAAPRO: an approach of protein secondary structure prediction based on KDD in the compound pyramid prediction model. Expert Systems with Applications. 2009,vol. 36,p. 9000–9006.

[12] Qi Dai, Yan Li, Xiaoqing Liu, Yuhua Yao, Yunjie Cao and Pingan He. Comparison study on statistical features of predicted secondary structures for protein structural class prediction: From content to position. BMC Bioinformatics.2013, p.14:152

[13] J. A. Cuff and G. J. Barton. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins: Structure, Function and Genetics. 2000, vol.40(3):502-511.

[14] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983, vol.22(12):2577-637.

[15] Kou chen chou. Some remarks on protein attribute prediction and pseudo amino acid composition. Journal of Theoretical Biology. 2011, vol.273:236-247.

[16] Xian Xu. Integrated feature subset selection extraction within applications in bioinformatics[D] Buffalo: State University of New York at Buffalo. August 16, 2006.

[17] D.T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. Journal of molecular biology. J Mol Biol. 1999,vol.292(2):195-202.

[18] Bingru Yang,Wu Qu,Yonghong Xie and Yun Zhai. Predicting protein second structure using a novel hybrid method. Expert Systerms with Applications. 2011, vol.38, p.11657-11664.

[19] JA Cuff and GJ Barton. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins. 1999, vol.34:508-519.

[20] M. Spencer, J. Eickholt, and J. Cheng. A deep learning network approach to ab initio protein secondary structure. IEEE/ACM Transactions on Computational Biology & Bioinformatics. 2015, vol. 12, pp. 103-112.

[21] J. Zhou, and O. Troyanskaya. Deep supervised and convolutional generative stochastic network for protein

secondary structure prediction. Presented at the 31st Int. Conf. Mach. Learn., Beijing, China, 2014.

[22] S. Wang, J. Peng, J. Ma, and J. Xu. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields, Scientific Reports. 2016. DOI: 10.1038/srep18962

[23] D.H. Hubel, and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, J. Physiol. 1962, vol.160, pp. 106-154.

[24] J. Xu，X. Luo，G. Wang，H. Gilmore，and A. Madabhushi. A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images. Neurocomputing. 2016, vol. 191, pp. 214-223.

[25] M. Fu，P. Xu，X. Li，Q. Liu，M. Ye，and C. Zhu. Fast crowd density estimation with convolutional neural networks. Engineering Applications of Artificial Intelligence. 2015, vol. 43, pp. 81-88.

[26] P. Qin, W. Xu, and J. Guo. An empirical convolutional neural network approach for semantic relation classification. Neurocomputing. 2016, vol. 190, p.1–9

[27] B.Rost,C.Sander. Prediction of protein secondary structure at better than 70% accuracy.J.Mol.Biol.1993,vol. 232(2), 583-599.

[28] K Chen, LA Kurgan and J Ruan. Prediction of protein structural class using novel evolutionary collocation based sequence representation. J Comput Chem. 2008, vol.29,p.1596–1604.

[29] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer and Y. Ng. Andrew. Self-taught learning: Transfer learning from unlabeled data. ICML. 2007

[30] Andrew Ng, Jiquan Ngiam, Chuan Yu Foo, Yifan Mai, Caroline Suen. UFLDL Tutorial. http://deeplearning.stanford.edu/wiki/index.php/ UFLDL_Tutorial, 7 April 2013.