High Accuracy Drug-Target Protein Interaction Prediction Method based on DBN

Wanrong Gu
School of mathematics and information
South China Agricultural University
Guangzhou, China
guwanrong@scau.edu.cn

Guohua Wang
School of mathematics and information
South China Agricultural University
Guangzhou, China
wangguohua@scau.edu.cn

Ziye Zhang
School of Mathematical
South China University of Technology
Guangzhou, China
ma_zere_x1@mail.scut.edu.cn

Yijun Mao
School of mathematics and information
South China Agricultural University
Guangzhou, China
yijunmao@163.com

Xianfen Xie
School of Economy
Jinan University
Guangzhou, China
txiexianfen2009@jnu.edu.cn

Yichen He
School of mathematics and information
South China Agricultural University
Guangzhou, China
heyichen666@hotmail.com

***Abstract*—Drugs may have multiple drug targets, and the most of targets are composed of different proteins. Therefore, the study of drug-target interaction (DTI) prediction has important meaning in drug repositioning, drug development time shortening and the cost of drug research and development reducing. Most of the existing methods are based on shallow learning model. The prediction accuracy is not high. In this paper, we proposed a deep belief network-based DTI prediction algorithm: we extracted extended connected fingerprint of the drug from the molecular structure. And then, we extracted the structure characteristics of the three peptide of the protein from the amino acid sequence of the protein. At last, we train the deep belief network by the characteristic vector extracted from drugs and proteins. In our proposed method, we fully use of the characteristics in the deep learning network and integrate the empirical feature selection into the deep belief network. Base on the public data set and compared with the state-of-the-art approaches, the experimental results show that our method outperforms the other algorithms in massive data sets.**

***Index Terms*—Matrix decomposition, Recommendation system, Distributed processing, Score prediction**

I. Introduction

Prediction of drug-protein interactions is an important research in recent drug effects discovery. Drug discovery research is an inefficient and costly research and development process [1], with very expensive costs about $\$1.8$ billion for each new molecular entity discovery. Besides, scientist would take about 10 years to make a new drug to reach medical market; for example, only about 20 new drugs have been approved by FDA as new molecular entity every year [2]. For the past decade, the drug development has decreased year by year [3]. For the above reasons, the abandoned or existing drugs are very important and urgent [4] for their new use. Such a new use among drug and protein research is called drug repurposing or repositioning [5].

In the past few decades, many scholars have studied drug-protein interaction systematically [3]. Newly-published studies on drug discovery follow the model of *one molecule, one goal, one disease*. This model identifies that some molecules interact with special proteins. However, the main limitation is that the drug is map to a target protein based on prior experience in mathematics model. In fact, many diseases are very complex with multi-factors, for instances, they would contact with different genes or different pathways. The accuracy of drug discovery would be very easily affected in this multifactorial environment. Traditional studies ignore the complex diseases associated with the complex factors. For the above issues, this research pattern would not good for drug discovery as expected [6], [7]. Most drug targets are cell proteins, which are designed to treat or diagnose disease through selective interactions with the compound. Current papers shows that classical drug targets has about 130 protein families [8], [9]. It is estimated that there are about $6,000 \sim 8,000$ pharmacological targets in the human genome. So far, however, only a small number of these targets have validated with approved drugs, and a lot of presumed drug targets still to be verified in future work [8], [10], [11].

Corresponding author: Xianfen Xie. This work was financially supported by Guangdong Natural Science Foundation Project (2018A030313437), Ministry of Education Humanities and Social Sciences Research Youth Fund Project (18YJCZH037), Guangdong Science and Technology Program Project (2018A070712021), 2019 Guangzhou philosophy and Social Science Planning Project (2019GZGJ31) and The 13th five-year Plan Project of philosophy and social science in Guangdong Province (GD18CXW01).

II. Related Works

Traditional DTI prediction methods can be roughly divided into two categories: docking simulation method and ligand-based method [12]. Docking simulation method deduces the relationship between drug and protein from the structural information of the drug, the amino acid sequence and structural information of the protein. Ligand-based method mine the potential relationship through the interactions between drug ligands and of target protein ligands. Usually, the accuracy of the former is higher than that of the latter, but the material

structure needs to be calculated in docking simulation method. Therefore, it cost a lot of time. In addition, some protein structure information is not suitable for the use of docking simulation, such as G protein coupling receptor. The ligand-based method does not require too much computation on the material structure and has well scalable. However, when the number of drugs and protein ligands is insufficient, the effectiveness of this method is not satisfactory. Similarity-based DTI prediction approach is also a common method. Yamanishi [13] used statistics and digraphs to make DTI predictions, which are classified according to the target protein type of the Kegg DRUG data. Four standard data sets, Enzyme, GPCRs, Icon channel and Nuclear receptors were constructed by Yamanishi [14] to record the drug-protein relationship, drug similarity and protein similarity. These standard data sets are widely used by subsequent DTI prediction research studies. DTI Predictions are similar to recommended technology, allowing for the use of interconnectedness of objects and information carried by objects to predict potential associations or user ratings. Therefore, there are a lot of recent researches using recommendation technology theory to study DTI prediction. DTI prediction studies usually used the well-known databases such as Drugbank, Chembl and Kegg DRUG [15]–[17], which record the characteristics of drugs, target protein characteristics and reaction conditions.

III. DTI Prediction Method Based on Deep Confidence Network

A. Parametric Training for Deep Trust Networks

In the state of a given visible layer node, the probability distribution of the state of the hidden layer node can be obtained. Similarly, the probability distribution of the visible layer can be obtained by giving the node state of the hidden layer:

$$p(h_j = 1|V) = \sigma(a_j + \sum_{i=1}^{n_V} v_i w_{ij}) \tag{1}$$

$$p(v_i = 1|H) = \sigma(b_i + \sum_{j=1}^{n_H} h_j w_{ij}) \tag{2}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$. The above two formulas get the probability that each node state in H or V is an open state (when $h_j = 1$ or $v_i = 1$). In order to train the RBM model, the node state needs to be sampled according to the probability distribution of the node state.

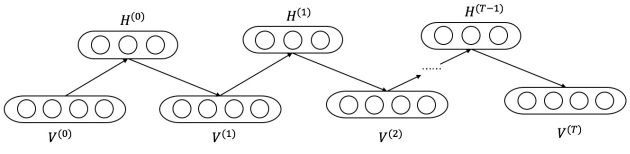


Fig. 1. T-step Gibbs Sampling IN RBM.

After Gibbs sampling, the state of H and $V^{'}$ can be obtained, and the row vector of the i row of W representing the value matrix by $W_{i,*}$, and The parameters update formula for RBM is as follows:

$$W_{i,*} = W_{i,*} + \alpha[p(h_i = 1|V^{(0)})V^{(0)} - p(h_i = 1|V^{'})V^{'}]) \tag{3}$$

$$a = a + \alpha[V^{(0)} - V^{'}]) \tag{4}$$

$$b = b + \alpha[p(H = 1|V^{(0)}) - p(H = 1|V^{'})]) \tag{5}$$

$$p(H = 1|V^{(t)}) = [p(h_1 = 1|V^{(t)}), p(h_2 = 1|V^{(t)}), ..., p(h_H = 1|V^{(t)})]^T \tag{6}$$

The learning rate is α, and integrating the Gibbs sampling and parameter updating formulas, and the entire RBM training algorithm is as follows:

Algorithm 1 Training algorithm flow

(1) Initialize the RBM parameters, in general, initialize the Value Matrix $W \sim N(0, 0.01)$, visible layer and hidden layer set a=b=0;
(2) Input individual or batch training samples into RBM for Gibbs sampling;
(3) Update the RBM parameters according to Eq.(2) to (5);
(4) Calculation of RBM model energy,If the energy function is not convergent or does not meet the specified training steps, Otherwise,out of the training;

B. The Back Propagation Method Is Used to Fine-tune The Network

In this equation,θ is DBN network parameters, $x^{(i)}$ is the input data,$y^{(i)}$ is the classification label for the corresponding input, $h_\theta(x^{(i)}$ is the output of the network. The DBN training process is as follows:

C. Using Deep Confidence Network in DTI Prediction

The interactions between drugs and proteins depend to a large extent on their chemical structure.Carrier proteins have specificity. They can only transport one kind of substance or similar nature when they transmit materials inside and outside the cell, and different affinity for different substances is partly determined by the structure of drugs and protein [18].

Set the network output to $output$ and the predicted threshold t, and the DTI implementation flow chart, as shown in Fig. 2, requires a training set and validation set to train and validate the DBN. Once the training is complete, it can be used in a predictive framework. Prior to the prediction phase, drug and protein information is collected from publicly available drug databases such as Drugbank, Chembl and Kegg, that means need to build the SMILES of medicines and protein amino acid sequences. In the prediction phase, you also need to enter the corresponding drug and protein feature sequence,and enter them to the DBN model. Finally, the network output is a real

Algorithm 2 Sampling algorithm flow

Require: $(X, Y) = [(x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_m, y_m)]$,$x_i$ is the network input,y_i is the actual classification of the corresponding input.

(1) In the pre-training stage, let DBN model have K-layer RBM, and $[x_1, x_2, x_3, ..., x_m]$ as the input of the first layer of RBM, and in accordance with the training methods in section 2.1, training the first layer of RBM until the energy value of the energy function converges, and makes $[x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, ..., x_m^{(1)}]$ the output of the first layer of RBM.

(2) Let $k = 2$,using $[x_1^{(k-1)}, x_2^{(k-1)}, x_3^{(k-1)}, ..., x_m^{(k-1)}]$ as the input training RBM of the k-layer RBM until the energy function converges, and get $[x_1^{(k)}, x_2^{(k)}, x_3^{(k)}, ..., x_m^{(k)}]$. So on, training the RBM layer by layer in hierarchy order.

(3) In the fine-tuning stage, X is input to the DBN to get the prediction $\widetilde{Y}$ and the crossover entropy is used to calculate the error of classification, and using Back Propagation to adjust the DBN network parameters.

number on domain $[0, 1]$, and compares the real value with the threshold t. When the real number is greater than t, the prediction result is that a given drug interacts with the protein and, conversely, no interaction occurs.

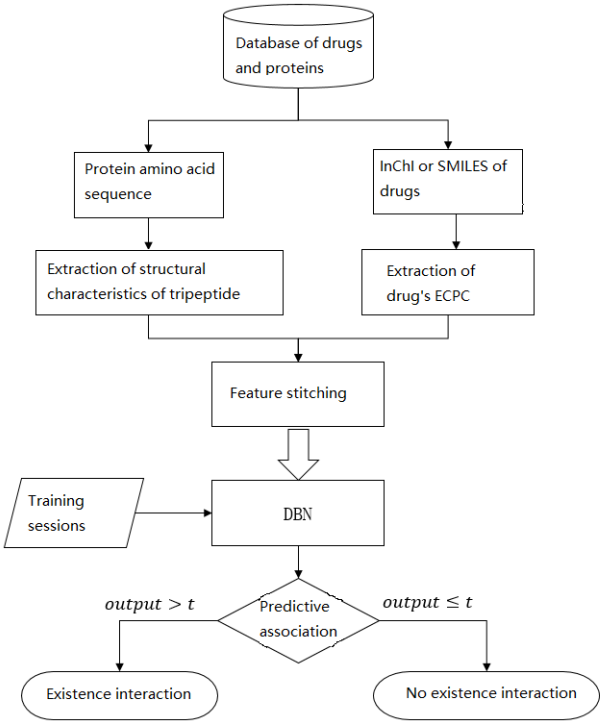


Fig. 2. Process framework for DTI prediction using DBN.

IV. Analysis of Experiments And Results

Based on four standard data sets, this paper simulates DTI prediction algorithm based on implicit semantic model, network and deep learning, and AUPR and AUROC evaluation indicator are used to compare these advantages and disadvantages and make specific analysis. Then, the self-built large data set based on $Drugbank$ database is used to analyze the predictive performance of DBN based DTI prediction for large-scale data.

A. Experimental Data Sets

In this paper, four data sets of $atkinson$, $GPCRs$, $Ionchannel$ and $Nuclearreceptor$ proposed by Yamanishi were used for experiments, as shown in TABLE I:

TABLE I
THE BASIC DATASETS OF THE EXPERIMENTAL COMPARISON IN THIS PAPER.

Data sets	Drugs	Proteins	Incidence number
Enzyme	445	664	2926
GPCRs	223	95	635
Ion channel	210	204	1476
Nuclear receptor	54	26	90

In the experiment, five fold cross validation was used to divide four data sets into five equal parts, and the experimental results were the average value of five experiments. In addition, two larger datasets have been constructed based on the $approved$ and $experiment$ datasets in the $Drugbank$ database. The $approved$ datasets have been approved by the FDA. Experiment datasets contain drugs and proteins that have interacted with each other in past experiments but have not been approved by the FDA. As shown in TABLE II, $approved$ represents an FDA-approved dataset. In this dataset, cold associations are not processed. $approved+experimentis$ a combination of FDA certification and experimental certification, and the dataset excludes cold associations data, which drugs and proteins with a correlation of less than 2. The two datasets randomly selected 20% of the association as the test set and the remaining 80% as a training set.

TABLE II
THE DATASETS BASED ON DRUGBANK.

Data sets	Drugs	Proteins	Incidence number
approve	1802	1705	6922
approve+experiment	742	478	5211

The negative sample selection in the dataset is sampled at random and the drug-protein pair with the same number of positive samples is selected as a negative sample, the total association size of $Enzyme$, $GPCRs$, $Ionchannel$, $Nuclearreceptor$, $approved$ and $approved + experiment$ is 5852,1270,2952,180,13844 and 10422.

Feature extraction of drugs and proteins using open source software rdkit and propy.Rdkit is a software for processing chemical informatics, which can quickly obtain chemical structure information through the compounds' InChI or SMILES, and has a large number of built-in functions to calculate certain properties and fingerprint characteristics of the compound. Propy is used to calculate the feature of amino

acid sequence of proteins, and it has a lot of built-in protein descriptors [20]. Both of these tools have corresponding libraries on python, allowing quickly get the feature of drug and protein, and then stitching together the two features to form the input vector of the DBN.

B. Baseline Methods

In this paper, we use the following baseline methods, which belong to the implicit semantic model, and the combination of network deep learning and SVM. The baseline methods: 1) Convex hull non-negative matrix factorization(CHNMF); 2) Weighted distribution method(WM); 3) Multiscale feature depth representation interaction prediction(MFDR)

C. Experimental Results And Analysis

1) Experiment 1: Comparison of The Prediction Experiments of Each New Method in TABLE I Data Set: The AUPR values and AUROC values of the algorithms under different data sets are given in TABLE III and IV respectively. As can be seen from the table, the DBN method used in this paper has better AUPR values in $Enzyme$, $Ionchannel$ and $Nuclearreceptor$ than the other three methods, and AUROC values are superior to other methods in the $Enzyme$ and $Ionchannel$.

TABLE III
COMPARISON OF AUPR VALUES OF EACH ALGORITHM.

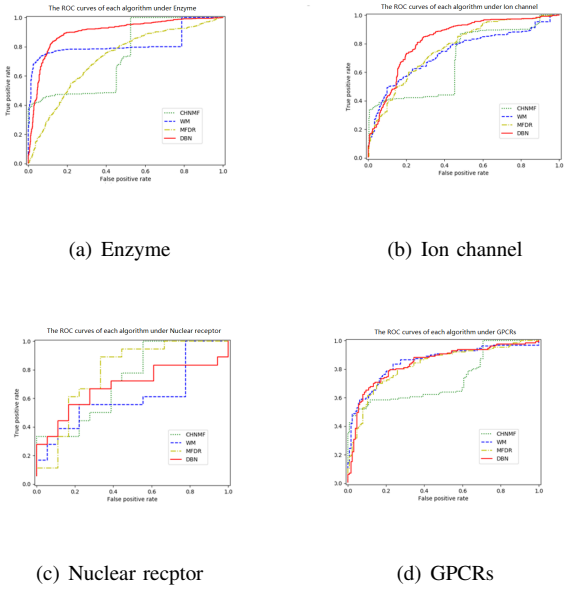
	Enzyme	GPCRs	Ion channel	Nuclear receptor
CHNMF	0.7769	0.7998	0.7360	0.7528
WM	0.8802	0.8681	0.7628	0.6868
MFDR	0.6818	0.8307	0.7569	0.7400
DBN	0.8902	0.8374	0.8082	0.7558

TABLE IV
COMPARISON OF AUROC VALUES OF EACH ALGORITHM.

	Enzyme	GPCRs	Ion channel	Nuclear receptor
CHNMF	0.7368	0.7296	0.6839	0.7253
WM	0.8207	0.8511	0.7372	0.6111
MFDR	0.7203	0.8263	0.7676	0.7809
DBN	0.9006	0.8420	0.8270	0.6821

From the above table analysis, it is found that DBN has poor prediction performance under training of small data. As the size of data set increases, DTI prediction performance based on DBN also improves. This indicates that DTI prediction method based on DBN is more suitable for larger drug and protein samples, and not for the insufficient data.

DBN, by means of the RBM structure, trains the network parameters of each layer in order of hierarchy, makes the network fit the distribution of training set as best as possible, and finally trains the whole classification by the method of fine-tuning. Finally, constructing a classifier which can well express the input features, and the classifier is deeper level than the SVM. This allows DBN to perform better in big data sets, while under small data sets, there is a poor performance due to too little data fitting. Fig. 5 compares the ROC curves of each algorithm in different data sets.



(a) Enzyme (b) Ion channel

(c) Nuclear recptor (d) GPCRs

Fig. 3. The ROC curve of each algorithm in each data set.

From Fig. 3 , we can see that:

(1) Because of the small amount of $Nuclear\ receptors$ datasets, the ROC curve of the four algorithms is in a ladder shape. In the other three data sets, the ROC curve becomes smooth with the increase of data.

(2) In Fig. 3(a), the non-deep-learning CHNMF and WM methods, the ROC curve under this data set is not smooth.

(3) Besides $Nuclear\ receptors$, the ROC curve of DBN is better than the other three algorithm, which shows that the DBN algorithm can distinguish the positive and negative samples well under the other three data sets.

2) Experiment 2: DBN Method in The Prediction Experiment of The Data Sets in TABLE II: The network structure of this experiment is the same as that in the previous paper (10048-500-500-500-1). The truncated normal distribution which average value of the initialization network parameters is 0 and the standard deviation is 0.01 is used as the output random number. In the pre-training stage, the learning rate is 0.001, and the first three layers of RBM are pre-trained. Gibbs sampling iteration times are 1, batch size is 32, and the pre-training times of each layer are 800,1000,1000 respectively. In the fine-tuning stage, the initial learning rate is 0.0001, and the learning rate is automatically adjusted according to the change of cross entropy. The $batch$ size is 16, and a total of 10,000 iterations are performed.

Fig. 4 shows the value of AUPR and AUROC recorded for DBN in each iteration 200 times in the fine-tuning stage after pre-training under two data sets. Since there is no pre-training in the last layer, both AUPR and AUROC are around 0.5 at the beginning of fine-tuning.

Fig. 4. The AUPR and AUROC curves of DBN under different iterations.

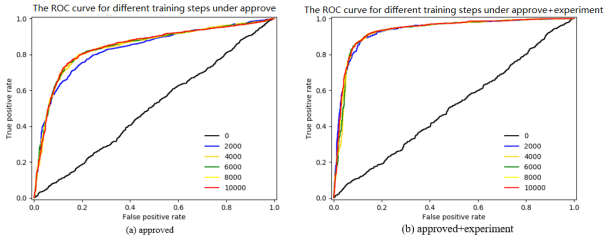


Fig. 5. The ROC curve of different iterations of DBN under the $approved + experiment$ data set.

From Fig. 4 and 5, we can see that:

(1) After pre-trained Network, and after about 400 times during the fine-tuning phase,AUPR and AUROC in $approved$ increased to 0.83762 and 0.8203 respectively,and and AUROC in the $approved + experiment$ increased to 0.9120 and 0.91368 respectively. That means after pre-training, DBN can already get a better result after about 400 fine-tuning iterations.By the training of 4000 steps, AUROC was basically stable, while AUPR was still fluctuating, and it started to be stable about the iteration of 5000 steps.After 10,000 fine-tuning iterations, AUPR and AUROC in $approved$ are stable at 0.8403 and 0.8461 respectively, and AUPR and AUROC in $approved + experiment$ are stable at 0.9174 and 0.9347 respectively.

(2) The experimental results of $approved + experiment$ were all better than $approved$, with a difference of nearly 0.1.The ROC curve at different iterations can also be found that the latter is relatively similar between ROC curves after 2000 iterations. However, there is a gap between the ROC curve of the former in 2000 iterations and the comparison of the latter. This suggests that the convergence rate in the $approved + experiment$ dataset is faster because all the correlations in the dataset are greater than 2.

V. Conclusion

This paper presents the prediction of DTI problem with deep confidence network, which is a deep learning model, and has the ability to construct implicit features autonomously, which can excavate the intrinsic association factor effectively. In order to integrate deep confidence network and DTI problem framework, the characteristics of drug fingerprint and amino acid sequence of protein were transformed and extracted, and as input of the model, the classification results were obtained. In order to verify the effectiveness of the proposed approach in DTI prediction, based on the measured data set and compared with the new method, the experimental results show that the deep confidence network method with the increase of data sets, DTI prediction accuracy improved, which helps solve the problem of data set prediction in a larger space.

References

[1] Dickson M, Gagnon JP. Key factors in the rising cost of new drug discovery and development. Nature Reviews Drug Discovery. 2004;3(5):417–29.

[2] Chen H, Zhang Z. A semi-supervised method for drug-target interaction prediction with consistency in networks. Plos One. 2013;8(5):e62975.

[3] Chen X, Yan CC, Zhang X, Zhang X, Dai F, Yin J, et al. Drugtarget interaction prediction: databases, web servers and computational models. Briefings in Bioinformatics. 2016;17(4):696.

[4] Booth B, Zemmel R. Prospects for productivity. Nature Reviews Drug Discovery. 2004;3(5):451–6.

[5] Wang Y, Zeng J. Predicting drug-target interactions using restricted Boltzmann machines. Bioinformatics. 2013;29(13):126–34.

[6] Chen X, Liu MX, Yan GY. Drugtarget interaction prediction by random walk on the heterogeneous network. Molecular Biosystems. 2012;8(7):1970.

[7] Iskar M, Zeller G, Zhao XM, Noort VV, Bork P. Drug discovery in the age of systems biology: the rise of computational approaches for data integration. Curr Opin Biotechnol. 2012;23(4):609–616.

[8] Drews J. Drug discovery: a historical perspective. Science. 2000;287(5460):1960–1964.

[9] Hopkins AL, Groom CR. The druggable genome. Nature Reviews Drug Discovery. 2002;1(9):727–730.

[10] Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? Nature reviews Drug discovery. 2006;5(12):993.

[11] Landry Y, Gies JP. Drugs and their molecular targets: an updated overview. Fundamental & clinical pharmacology. 2008;22(1):1–18.

[12] Ding H, Takigawa I, Mamitsuka H, Zhu S. Similarity-based machine learning methods for predicting drug–target interactions: a brief review. Briefings in bioinformatics. 2013;15(5):734–747.

[13] Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. Bioinformatics. 2008;24(13):i232–i240.

[14] Yamanishi Y, Kotera M, Kanehisa M, Goto S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics. 2010;26(12):i246–i254.

[15] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic acids research. 2011;40(D1):D1100–D1107.

[16] Kanehisa M. The KEGG database. In: In SilicoSimulation of Biological Processes: Novartis Foundation Symposium 247. vol. 247. Wiley Online Library; 2002. p. 91–103.

[17] Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic acids research. 2013;42(D1):D1091–D1097.

[18] Cobanoglu MC, Liu C, Hu F, Oltvai ZN, Bahar I. Predicting drug–target interactions using probabilistic matrix factorization. Journal of chemical information and modeling. 2013;53(12):3399–3409.

[19] Rogers D, Hahn M. Extended-connectivity fingerprints. Journal of chemical information and modeling. 2010;50(5):742–754.

[20] Cao DS, Xu QS, Liang YZ. propy: a tool to generate various modes of Chous PseAAC. Bioinformatics. 2013;29(7):960–962.