A large-scale prediction of protein-protein interactions based on random forest and matrix of sequence

Kenan Wang [1], Xiaoman Zhao [2] and Xue Wang [2, *]

[1] University College London, Institute of Child Health, London, WC1E 6AE, The United Kingdom
[2]Institute of Intelligent Machine, Hefei Institutes of Physical Science, Chinese Academy of Sciences, HeFei 230031, China

Abstract. Protein-protein interaction (PPIs) is an important part of many life activities in organisms, and the prediction of protein-protein interactions is closely related to protein function, disease occurrence, and disease treatment. In order to optimize the prediction performance of protein interactions, here a RT-MOS model was constructed based on Random Forest (RF) and Matrix of Sequence (MOS) to predict protein-protein interactions. Firstly, MOS is used to encode the protein sequences into a 29-dimensional feature vector; Then, a prediction model RT-MOS is build based on random forest, and the RT-MOS model is optimized and evaluated using the test set; Finally, the optimized model RT-MOS is used for prediction. The experimental results show that the accuracy rates of the RT-MOS model on the benchmark dataset and the non-redundant dataset are 97.18% and 91.34%, respectively, and the accuracies on four external datasets of C.elegans, Drosophila, E.coli and H.sapiens are 96.21%, 97.86%, 97.54% and 97.75%, respectively. Compared with the existing methods, it is found that it is superior to the existing methods. The experimental results show that the model RT-MOS has the advantages of saving time, preventing overfitting and high accuracy, and is suitable for large-scale PPIs prediction.

Keywords: Random forest; Matrix of Sequence; Protein-protein interaction.

1. Introduction

Protein-protein interaction (PPIs) is an important part of many life activities in organisms. Almost all life processes are related to protein interactions, such as metabolism, signal transduction, cell cycle regulation, metabolism, apoptosis and immune response. The

study of protein interaction can help people to fundamentally understand the mechanism of disease, so as to prolong the life of patients with genetic diseases and improve their quality of life. In terms of PPIs prediction, many high-throughput experimental methods have emerged in recent years [1-3]. However, those methods are chemical experimental methods, which are time-consuming and laborious, and large-scale protein interaction prediction is difficult to achieve.

Machine learning makes it possible to predict large-scale protein interactions. So far, a large number of machine learning models have emerged, including support vector machines (SVM), neural networks (NN), naive Bayes, K-nearest neighbor and decision tree, which have been used to predict protein interactions [4-6]. Although the calculation methods of protein interaction prediction have been developed to some extent, there are still some limitations. For example, general machine learning models may not be able to deal with the noise value of protein sequences well [7]. Random Forest (RF) has high prediction accuracy, has a good tolerance for outliers and noises in data, and is not prone to over fitting problems. It can well solve the shortcomings of traditional machine learning such as decision trees [7]. RF has also been applied in protein interaction prediction. For example, Qi et al. [8] proposed a new method to predict protein interaction prediction by calculating similarity based on random forest, and achieved a positive prediction rate of 70.45%. In 2014, Bhowmick et al. [9] built a protein interaction prediction model based on random forest, and obtained 89% accuracy, which confirmed the effectiveness of RF algorithm applied to protein interaction prediction. PPIs need to convert heterogeneous amino acid sequences into homogeneous vector features (i.e. protein coding). In 2019, Gui et al. [10] proposed a MOS protein coding method based on deep learning. This method considers the frequency information of the entire amino acid sequence, and has the advantages of simple coding and time-saving. In view of the advantages of random forest in processing noise and over fitting, as well as the advantages of simple and time-saving sequence matrix coding, we build protein interaction prediction model based on random forest and sequence matrix to optimize the prediction performance of protein interaction prediction model.

* Corresponding author: 181543681@qq.com

2. Materials and methods

2.1 Matrix of Sequence（MOS）

2.1.1 Amino acid classification

First, 20 kinds of conventional amino acids are divided into 7 groups according to the dipole and volume of side chain (see Table 1). Then, referring to the classification in Table 1, replace the amino acid sequence with the corresponding category of amino acid in the amino acid classification table, and the dimension of the sequence matrix will be 20×20 down to 7×7.

Table 1. Amino acid classification based on side chain dipole and volume.

Number	Amino acid
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)

2.1.2 Definition of MOS

Vector of protein sequence (VOS): Hypothetical non-empty finite set:$\Omega = \{w_1,...,w_N\}$, where w_i is amino acid classification. Given sequence: $S=S_1,S_2,...,\ S_L$, where L represents the length of sequence S, $S_i \in \Omega$, $1 \le i \le L$. The sequence vector of a given sequence S can be expressed as:VOS=($Cw_1,...,Cw_N$) , where is the number of occurrences of the w_i in the sequence S. Based on the definition of the sequence vector, the sum of all elements in the sequence matrix is equal to L.

Matrix of Sequence(MOS) , Hypothetical non-empty finite set: $\Omega = \{w_1,...,w_N\}$, where N is the number of categories of the sequence. Given sequence: $S=S_1,S_2,...,\ S_L$, where L represents the length of sequence S, $S_i \in \Omega$, $1 \le i \le L$. The sequence matrix of a given sequence S can be expressed as: $MOS=[m_{ij}]_{N\times N}$.

2.1.3 Algorithm of MOS

Hypothetical non-empty finite set: $\Omega = \{w_1,...,w_N\}$, where N is the number of categories of the sequence. Given sequence: $S=S_1,S_2,...,\ S_L$, where L represents the length of sequence S, $S_i \in \Omega$, $1 \le i \le L$. The sequence matrix of a given sequence S can be expressed as:

Input sequence: $S=S_1,S_2,...,\ S_L$;

Output sequence matrix: $MOS=[m_{ij}]_{N\times N}$.

The sequence matrix algorithm is calculated as follows:

Step 1. Initial value is set up: $i \leftarrow L$, $VOS \leftarrow VOS_0=0$, $MOS \leftarrow MOS_0=0$.

Step 2. $VOS[s_i] \leftarrow VOS[s_i]+1$.

Step 3. $MOS[s_i] \leftarrow MOS[s_i]+VOS$..

Step 4. $i \leftarrow i-1$..

Step 5. If $i \ge 1$, *go to step 2.*

To reduce the computational vector, we first classify 20 amino acids into 7 classes according to the amino acid classification method in Table 1. Thus, a protein sequence can be represented by a matrix of 7×7. The next step is to standardize m_{ij} of each matrix element ranging from 0 to 1.To distinguish the lengths of the protein. Finally, a total 29-dimensional vector has been built to represent each protein sequence.

2.2 Random Forest(RF)

Random Forest (RF) is an algorithm that Breiman et al. [11] combined random feature selection method and Bagging idea to integrate multiple decision trees. In most cases, Bagging method is used for training in random forests, samples are selected randomly, and samples are trained by playback sampling.

RF is an integrated classifier constructed by several decision tree models $\{h(X,\theta_K),k=1,...,K\}$ in Bagging integration mode, where θ_K is an independent random vector with the same distribution, and K is the number of decision trees in the forest. Input x, output $f(x)=majority\{h(X,\theta_K),k=1,...,K\}$.

Given training data set $D=\{(x_i,y_i),x_i \in X,y_i \in Y,i=1,2,...,n\}$, Where n represents the sample size of dataset D, X represents the set of p-dimensional feature vectors, and Y represents the category vector. The RF margin function can be expressed as:

$$mr\ (X,Y) = av_k I(h(X,\theta_K)=Y) - max\ av_k I(h(X,\theta_K)=j) \quad (1)$$

Where, $j \ne Y$, $mr\ (X,Y)$ represents the margin function, $I\ (.)$ represents the indicative function, and $h(X,\theta_K)$ represents the classification model sequence.

3. Data Set Construction

3.1 Benchmark data set and Non-redundant data set

Benchmark data set and Non-redundant data set were provided by Pan et al. [12]. Benchmark data set includes positive correlation data set and negative correlation data set. The positive data set is from the Human Protein Reference Database (HPRD, 2007), and the negative datasets is constructed by subcellular localization information. Most protein sequences range in length from 100 to 1000. Protein pairs containing less than 50 residues and uncommon amino acid sequences (containing B, J, O, U, X and Z) are deleted. The data set obtained includes 36591 pairs of positive correlation samples and 36324 pairs of negative correlation samples. 30000 positive correlation samples and 30000 negative correlation samples are randomly selected each time to form a training set, and the rest was used as a test set. On the basis of Benchmark datasets, delete protein sequence pairs with sequence identity ≥ 25%, and the resulting data set is non redundant. The datasets contains 3899 positive correlation protein pairs and 4262 negative correlation protein pairs.

3.2 External Datasets

In order to verify the prediction performance of RT-MOS model, in addition to the benchmark data set and non redundant data set, four different species of datasets are also selected as external datasets. See Table 1 for details of the data set. It can be seen from Table 1 that data volume of positive correlation samples and negative correlation samples of four species are divided according to 1:1 ratio. The training set accounts for about 60% of the total sample volume, and the remaining 40% is used as the test set.

Table 2. External data set.

Data set	Positive correlation data set (pair)	Negative correlation data set (pair)	Training set (pair)	Test set (pair)
C. elegans	4030	4030	4836	3224
Drosophila	21975	21975	26370	17560
E. coli	6594	6594	7912	5276
H. sapiens	37027	37027	44432	29622

4. Results

4.1 Model construction

4.1.1 Experimental Design

The RT-MOS model is designed and implemented based on the Keras framework. It is written in Python language and supports both CPU and GPU. The flow chart of experimental design is shown in Fig.1, mainly including data acquisition, data processing and model building. Data acquisition refers to obtaining protein interaction data sets from HPRD, Swiss Port, PIR and UniProt databases; Data processing refers to the use of MOS coding to extract features from protein interaction data sets and convert letter sequence data into computer recognized feature vectors; The model construction is to input the coded feature vectors into machine learning, train the protein interaction prediction model by adjusting and optimizing parameters, test the model using test sets, and finally evaluate and compare the model.

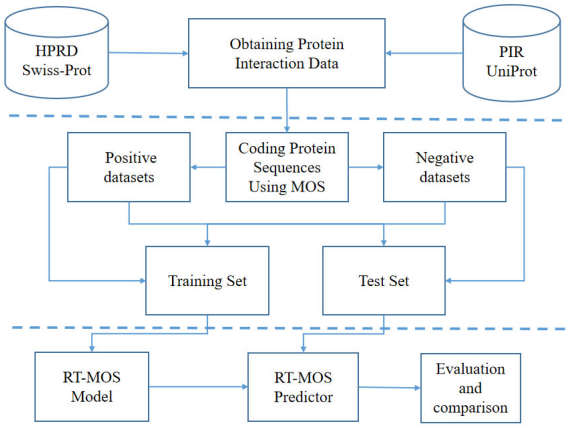

Fig. 1. The flow chart of experimental design.

4.2 Prediction performance of PPIs

4.2.1 Prediction performance on benchmark data set

In order to study the prediction performance of the model RT-MOS, three prediction models, KNN-MOS, DT-MOS and AB-MOS, were constructed by combining K-Nearest Neighbor (KNN), Decision Tree (DT), Adaptive Boosting (AB) and MOS feature extraction methods. Through experiments, the average prediction performance of the four models is shown in Table 3. It can be seen from Table 3 that the accuracy range of the four models is 67.99-97.18%, with a variation of 29.19%. The accuracy of KNN-MOS, DT-MOS, AB-MOS and RT-MOS models are 86.25%, 94.19%, 67.99% and 97.18% respectively. Among them, the accuracy, AUC, recall, precision and loss of RT-MOS are significantly higher than those of KNN-MOS, DT-MOS and AB-MOS.

Table 3. Prediction performance on benchmark data set.

Method	Accuracy(%)	AUC(%)	Recall(%)	Precision(%)	Loss
KNN-MOS	86.25	86.24	82.05	82.41	4.75
DT-MOS	94.19	94.20	93.56	91.94	2.01
AB-MOS	67.99	68.00	69.54	61.73	11.06
RT-MOS	97.18	97.19	95.96	96.44	0.97

The training time and prediction time of different prediction models vary greatly, as shown in Table 4. It can be seen from Table 4 that the model KNN-MOS has the shortest training time (0.1938 s), the longest test time(49.0567 s) and prediction time(49.4834 s). The reason for the short training time of KNN-MOS may be that the KNN training stage only includes feature vectors and category labels of stored training samples. In the testing phase, the test point needs to be classified by assigning the most frequently used label among the k training samples closest to the query point, which requires

higher computation, so the testing time and prediction time are long. Comparing the four prediction models of KNN-MOS, DT-MOS, AB-MOS and RT-MOS, it was found that the training time, testing time and prediction time of the model DT-MOS are optimal. Although the training time (4.6204), testing time (0.04) and prediction time (0.0392) of the model RT-MOS are higher than those of the model DT-MOS (4.0988), testing time (0.0063) and prediction time (0.0026). However, the accuracy of DT-MOS is 94.19%, which is significantly lower than that of RT-MOS (97.18%).

Table 4. Training time, testing time and prediction time on the benchmark data set.

Method	Train time (s)	Test time (s)	Prediction time (s)
KNN-MOS	0.1938	49.0567	49.4834
DT-MOS	4.0988	0.0063	0.0026
AB-MOS	12.5092	0.0853	0.0842
RT-MOS	4.6204	0.0400	0.0392
DNN-MOS[10]	N	N	0.1261

4.2.2 Prediction performance on non-redundant data set

In order to evaluate the generalization performance of the RT-MOS model, we tested the performance of the RT-MOS model on non redundant data set, and obtained 90.34% accuracy. The specific prediction results are shown in Table 5. Table 5 shows that the accuracy, recall and AUC of RT-MOS on non redundant data set are 91.34%, 95.52% and 93.76% respectively. However, Gui et al.[10] and Shen et al.[13] obtained 88.29% and 85.84% accuracy on non redundant data set, respectively. It can be seen that the accuracy of RT-MOS on non redundant data set is better than that of DNN-MOS and SAE-AC. The protein interaction of RT-MOS model on low similarity data set is still effective, and it can be used to predict protein interaction on low similarity data set.

Table 5. Prediction performance on the non-redundant dataset.

Method	Accuracy(%)	Recall(%)	AUC(%)
RT-MOS	91.34	95.52	93.76
DNN-MOS[10]	88.29	93.63	92.23
SAE-AC[13]	85.84	N/A	N/A

It can be seen from Table 3 that the accuracy of the model RT-MOS on the non redundant data set is 97.18%, while the accuracy of DNN-MOS [10] and SAE-AC [13] on the non redundant data set is 94.34% and 97.19% respectively.Table 5 shows that the accuracy of RT-MOS, DNN-MOS and SAE-AC models on non redundant datasets is 91.34%, 88.29% and 85.84% respectively. To sum up, the prediction performance of RT-MOS, DNN-MOS and SAE-AC models on non redundant datasets is better than that on non redundant datasets. This shows that in protein interaction prediction, the sequence identity of data set has a great impact on the performance of models. Reducing the sequence identity of data set will lead to a decline in prediction performance.

4.2.3 Prediction performance on external datasets

To further verify the generalization performance of the model RT-MOS, four external data sets (C. elegans, Drosophila, E. coli and H. sapiens) are applied to the model RT-MOS (see Table 6). It can be seen from Table 6 that the accuracy of the model RT-MOS on the four external data sets of C.elegans, Drosophila, E.coli and H.sapiens is 96.21%, 97.86%, 97.54% and 97.75% respectively. Among them, the prediction performance of Drosophila, E.coli and H.sapiens is better than that on the benchmark data set (the accuracy rate is 97.18%). The experimental results in Table 6 show that the model RT-MOS has also achieved good prediction performance in protein interaction prediction of other species, and the model RT-MOS has good generalization ability.

Table 6. Prediction performance on external datasets.

Dataset	Accuracy(%)	AUC(%)	Recall(%)	Precision(%)	Loss
C. elegans	96.21	96.21	95.54	95.02	0.9442
Drosophila	97.86	97.86	97.23	96.79	0.7849
E. coli	97.54	97.54	97.18	96.63	0.7668
H. sapiens	97.75	97.75	97.12	96.45	0.7775

4.3 Comparison with existing methods

In order to verify the effectiveness of the model RF-MOS in protein interaction prediction, the model RF-MOS is compared with existing methods, and the comparison results are shown in Table 7. The datasets used by all methods in Table 7 are human datasets downloaded from the Human Protein Reference Database (HPRD). It can be seen from Table 7 that the accuracy of existing methods is between 83.90% and 94.10%, and the best result is the model CS-SVM. The model RT-MOS and DNN-MOS adopt the same coding method, but RT-MOS achieves 97.18% accuracy, which is significantly better than DNN-MOS (93.35%), with an increase of 3.83 percentage points. Through comparison, we found that the model RT-MOS can improve the prediction performance of existing methods, and have the advantages of saving time, preventing over fitting and high accuracy, and is suitable for large-scale protein interaction prediction.

Table 7. Comparison with existing prediction methods.

References	Method	Accuracy
Shen's work [13]	SVM-CT	0.8390
You's work[14]	ELM-LD	0.8480
Zhou's work[15]	SVM-LD	0.8876
Guo's work[16]	SVM-AC	0.9067
Zhang's work[17]	CS-SVM	0.9410
Gui's work [10]	DNN-MOS	0.9335
Our work	RF-MOS	0.9718

5. Conclusion

The random forest algorithm has been applied in many fields. Although it has also been applied in protein interaction prediction, the prediction performance still needs to be improved. Therefore, we used the random forest algorithm, combined with MOS protein sequence coding method, to build the RF-MOS protein interaction prediction model, and obtained 97.18% accuracy, 97.19% AUC, 95.96% recall, and 96.44 prediction. Compared with the models constructed by other machine learning algorithms, the prediction performance of RF-MOS model is better than that of KNN-MOS, DT-MOS, AdaBoost MOS and other models. The reason for the good prediction performance of RF-MOS model may be that random forest is an integrated learning method, integrating multiple decision trees can avoid the defect of a single decision tree. Since bagging is equivalent to sampling samples and features, it can avoid over fitting. The model RF-MOS still achieves good accuracy when using low dimensional feature vectors, avoiding problems such as large error, low accuracy and over fitting. Moreover, the RF training is fast, the accuracy of prediction results is high, and it can carry a large number of inputs. In addition, the MOS coding dimension is low and time-saving, so the model RF-MOS is suitable for processing large-scale sample data. In view of the above advantages, RF-MOS model can be a useful complement to protein interaction prediction.

References

1. P. Uetz, L. Giot L, G. Cagney, A. Traci, T. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S.Fields, J. M. Rothberg. Nature, 403, 623-627(2000).
2. D. J. LaCount, M. Vignali, R. Chettier, A. Phansalkar, R. Bell, J. R. Hesselberth, L. W. Schoenfeld, I. Ota, S. Sahasrabudhe, C. Kurschner, S. Fields, R. E. Hughes, Nature, 438, 103-107(2005).
3. J. R. Parrish, J. Yu, G. Liu, J. A. Hines, J. E. Chan, B. A. Mangiola, H. Zhang, S. Pacifico, F. Fotouhi, V. J. DiRita, T. Ideker, P. Andrews, R. L. F. Jr, Genome Biol., 8, R130(2007).
4. P. Chatterjee, S. Basu, M. Kundu, M. Nasipuri, D. Plewczynski, Cell Mol. Biol. Lett., 16, 264-278(2011)
5. S. Dohkan, A. Koike, T. Takagi, Silico Biol., 6, 515-529(2006)
6. X. W. Chen, M. Liu, Bioinformatics, 21, 4394-4400(2005)
7. G. Biau, Journal of Machine Learning Research, 13, 1063-1095(2012)
8. Y. Qi, J. Klein-seetharaman, Z. Bar-joseph, Pac. Symp. Biocomput, 10, 531-542(2015)
9. S. S. Bhowmick, I. Saha, G. Mazzocco, U. Maulik, L. Rato, D. Bhattacharjee, D. Plewczynski, Molecular Biosystems, 10, 820-830(2014)
10. X. Wang, Y. J. Wu, R. J. Wang, Y.Y. Wei, Y. M. Gui, Plos one. 14, e0217312(2019)
11. L. Breiman, Machine Learning, 45, 5-32(2001)
12. X. Y. Pan, Y. N. Zhang, H. B. Shen, Journal of Proteome Research, 9, 4992-5001(2010)
13. J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, H. Jiang, Proc. Natl Acad. Sci. 104, 4337-4341(2007)
14. Z. H. You, Z. Ji, X. Luo, X. Gao, S. L, Biomed Res Int., 2014, 598129(2014)
15. Y. Z. Zhou, Y. GAO, Y. Y. Zheng, Adv. Comput. Sci. Edu. Appl., 202, 254-262(2011)
16. Y. Z. Guo, M. L. Li, X. M. Pu, G. B. Li, X.M. Guang, W.J. Xiong, J. Li, Bmc Research Notes, 3, 145-152(2010)
17. Y. N. Zhang, X. Y. Pan, Y. Huang, H. B. Shen, Journal of Theoretical Biology, 283, 44-52(2011)