Feature Integration and Feature Augmentation for Predicting GPCR-Drug Interaction

Isabelle Bichindaritz, Guanghui Liu

Intelligent Bio Systems Laboratory, Biomedical and Health Informatics

State University of New York at Oswego, 7060 NY-104, Oswego, NY 13126

ibichind@oswego.edu, guanghui.liu@oswego.edu

Abstract

Accurately predicting the interaction between G-protein-coupled receptors (GPCR) and drugs is of great significance for understanding protein functions and drug discovery and has become a hot spot in current research. To improve the accuracy of GPCR-drug interaction prediction, this paper proposes a new GPCR-Drug interaction prediction method based on multi-feature integration and feature augmentation from deep random forest: First, the sequence features of GPCR from amino acid composition and protein evolution are extracted respectively, and the characteristics of the drug molecule from the molecular fingerprint perspective are formulated; then, the extracted multiple features are combined to obtain the feature representation of the GPCR-Drug pair; finally, based on the proposed GPCR-Drug feature representation method, we use deep random forest to generate augmented features and construct cascaded predictions model. The cross-validation and independent test results on the standard data set verify the effectiveness and greater explainability of the proposed method.

Introduction

G protein-coupled receptors (GPCRs) are one of the largest groups of proteins in vertebrate species (Heuss and Gerber 2000). GPCR-associated proteins could play distinct roles in receptor signaling such as directly mediate receptor signaling, regulate receptor signaling through controlling receptor localization and trafficking, and physically linking the receptor to various effectors as a scaffold, etc. altering receptor pharmacology and/or other aspects of receptor function (Xiao et al. 2013). Involved in many diseases such as cancer, diabetes, neurodegenerative, inflammatory and respiratory disorders, GPCRs are among the most frequent targets of therapeutic drugs (Chou 2005). Current drugs, many of which have excellent therapeutic benefits, are directed towards only a few GPCR members. So, in the drug discovery process, finding the interactions between drugs and target-GPCR is the most important a task of great importance. Therefore, huge efforts are currently underway to develop new GPCR-based drugs, particularly for cancer (Lappano and Maggiolini 2011).

Predicting the interactions between drugs and GPCRs is very useful to reduce the unnecessary waste of time and money in synthesized drug research (Sirois et al. 2005). Some computational methods for the prediction have been presented in recent studies. These methods were developed in this regard based on the knowledge of the 3D (dimensional) structure of protein (Yamanishi et al. 2008). Unfortunately, their usage is quite limited because the 3D structures for most GPCRs are still unknown. To overcome the situation, Xuan Xiao et al. developed a sequence-based classifier, called "iGPCR-drug", to predict the interactions between GPCRs and drugs in cellular networking (Xiao et al. 2013). This model extracts the GPCR pseudo amino acid composition (PseAAC) features and the molecular fingerprint features of the drug generated by the Fourier transform, and combines the two to form a synthetic feature; then it uses the fuzzy K nearest neighbor (Fuzzy KNN) Classification algorithm to make predictions. iGPCR-drug classifier performance is remarkably higher than the rate achieved by the existing peer method developed in 2010. Hu et al. (Hu et al. 2016) proposed a new sequence-based method to predict the GPCR-drug interaction. In this method, the discrete wavelet transform (DWT) was utilized to extract the features of drugs based on their fingerprints. For GPCRs, the pseudo position specific scoring matrix (PsePSSM) features were extracted. Although this advanced model characterized by the combination of progressive feature extraction method (PsePSSM and DWT), ensemble learning method (RF), and post-processing procedure (PPP) was better than the foregoing ones, it seemed that the generalization ability of this advanced model was still limited. Wang et al. (Wang et al. 2020) propose a new powerful sequence-based method for

identifying the GPCR-drug interaction based on WordBook learning from sequences. For GPCRs, a bag-of-words (BoW) model is used to extract sequence features, which can extract more pattern information from low-order to high-order and limit the feature space dimension. For drug molecules, discrete Fourier transform (DFT) is used to extract higher-order pattern information from the original molecular fingerprints. The feature vectors of two kinds of molecules are concatenated and input into a simple prediction engine distance-weighted K-nearest-neighbor (DWKNN). This basic method is easy to be enhanced through ensemble learning. But the results of leave-one-out cross-validation from this method are not satisfactory.

In the current study, we present a novel framework to predict the interactions between GPCRs and drugs. For GPCRs, we extract the PseAAC features and PsePSSM features respectively. For drugs, we extract the characteristics of the drug molecule from the molecular fingerprint perspective. then, we combine the extracted multiple features to obtain the feature representation of the GPCR-Drug pairs. Finally, we use a feature augmentation deep random forest framework to construct predictions model. From the experimental results analysis on the benchmark datasets with both cross-validation and independent validation tests, the feasibility and efficacy of the proposed method are demonstrated.

Materials

The Cross-validation Dataset:

In this study, the main benchmark datasets consist of the interactive GPCR-drug pairs and non-interactive GPCR-drug pairs. The 'interactive' pair means that two counter-parts of the pair are interacted with each other in the KEGG database (http://www.kegg.jp/kegg/); while the 'non-interactive' pair means the pair whose two counter-parts are not interacted with each other in the drug-target networks. This experiment uses a cross-validation dataset containing 1860 GPCR-Drug pairs (620 interactive GPCR-drug pairs and 1,240 non-interactive GPCR-drug pairs) (He et al. 2010). The GPCR-Drug pairs contain 92 unique GPCRs and 217 unique drugs. All the detailed information for the compounds or drugs can be found in the KEGG database.

The independent validation dataset:

To prevent the classifier from over-optimizing on the training set, it is necessary to construct an independent validation dataset to verify the generalization ability of the classifier. We first extract different proteins from the KEGG database, which can interact with 217 drugs in the standard cross-validation set (there are currently 904 such proteins in the database); then delete the proteins that are not GPCRs; In the obtained GPCR-Drug pairs, we then delete the pairs that have already appeared in the 1860 benchmark data set, so that the remaining 130 GPCR-Drug pairs are a positive subset of the independent test set we hope to construct. Similarly, the same method can be used to artificially synthesize 260 GPCR-drug pairs without interaction to construct a negative subset of the independent test set. The positive and negative subsets are combined to form an independent test set. The final independent test set consists of 390 GPCR-drug pairs, including 130 positive samples and 260 negative samples

Methods

Gene Features Extraction

In this study, each sample is composed of a GPCR and a drug pair. Therefore, first the characteristic expression of the GPCR and the drug are obtained separately, and then these two characteristics are combined to represent the characteristics of the sample.

Representing drugs with molecular fingerprints and wavelet

In this study, molecular fingerprints were suggested for the description of drug molecules. Molecular fingerprints are bit-string representations of molecular structure and properties. First, we can obtain a MOL file for each of the drugs concerned from the KEGG database via its code that contains the detailed information of chemical structure. Second, by using a chemical toolbox software called Open Babel (O'Boyle et al. 2011), which can be downloaded from the website at http://www.openbabel.org/, we can convert the MOL file format into its 2D molecular fingerprint file format. Four types of fingerprints: FP2, FP3, FP4 and MACCS, can be generated in the current version of Open Babel. We used the FP2 fingerprint format. It is a path-based fingerprint that identifies small molecule fragments based on all linear and ring substructures and maps them onto a bit-string using a hash function (somewhat similar to the Daylight fingerprints (Dou et al. 2012). It is a 256-bit vector, whose component values are an integer between 0 and 15.

Discrete wavelet transform (DWT) is a crucial tool (Percival and Walden 2006) to reduce the dimensions of drug fingerprint vector in this study. Using DWT, we can reduce the dimensions of drug fingerprint vector from 256-dimensional to 128-dimensional feature vector. A 128-dimensional feature vector is generated by fingerprint wave in the form of:

$$D= (A_1, A_2, \ldots, A_{128})^T \quad (1)$$

Representing GPCR sequences with PseACC

PseACC (Xiao et al. 2013) is a feature representation method proposed by Zhou Guocheng in 2001. The PseACC method encodes the amino acid composition information of a protein and the amino acid sequence information together.

Let $F_{AAC} = (f_1, f_2, \cdots, f_{20})$ be the classic 20-dimensional amino acid composition, and $f_i (i = 1,2, \cdots ,20)$ be the normalized frequency of the 20 original amino acids in the protein. Then, the PseAAC vector is the weighted combination:

$$X_{\text{PseAAC}}^{\lambda} = (x_1, \cdots, x_{20}, x_{20+1}, \cdots, x_u, \cdots, x_{20+6\lambda})^T \quad (2)$$

where

$$x_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{6\lambda} \tau_j}, & (1 \le u \le 20) \\ \frac{\omega \tau_{u-20}}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{6\lambda} \tau_j}, & (20+1 \le u \le 20+6\lambda) \end{cases} \quad (3)$$

where ω is the weighting factor, $\omega = 0.1$. Here, the correlation coefficient λ is set to 20, then the PseAAC feature of GPCR is obtained, and the dimension is 140.

Representing GPCR sequences with PsePSSM

Evolutionary information contained in protein sequence alignments is important for improving the predictive performance. The GPCR sequences involved in this study are given in Online Supporting Information Ω3. Position-specific scoring matrix (PSSM) can partially provide the evolutionary information of protein sequence, which is obtained from multiple sequence alignment. For a GPCR receptor sequence P with L amino acid residues, we generated its PSSM matrix of L×20 using PSI-BLAST (Schäffer et al. 2001) to search the Swiss-Prot database through three iterations with 0.001 as the E value cutoff for multiple sequence alignment against the sequence of the receptor. Let $P_{pssm} = (p_{kj})_{L\times 20}$ be the normalized PSSM of a GPCR protein with L amino acid residues.

Then, the PSSM composition is a 20-dimensional feature vector as defined as:

$$F_{\text{pssm}} = (p_1, p_2, \cdots p_j, \cdots, p_{20})^T \quad (4)$$

In order to remedy the defection of losing sequence-order information, sequence-order information contained in PSSM is then extracted by calculating the correlation factor of each column of a PSSM as follows:

$$\theta^g = (\theta_1^g, \theta_2^g, \cdots, \theta_{20}^g)^T \quad (5)$$

where

$$\theta^{\text{g}} = \frac{1}{L-g} \sum_{t=1}^{L-g} \left(p_{t,j} - p_{t+g,j}\right)^2, 1 \le j \le 20, 0 \le g \le G, G < L \cdot$$

g is the rank of correlation along the receptor sequence. The scalar quantity θ^{g} is the correlation factor by coupling the g-most contiguous PSSM scores along the protein sequence for the amino acid type j. In this study, a compact PsePSSM feature vector is defined as follows:

$$X_{\text{psePSSM}}^{g} = \begin{pmatrix} F_{PSSM} \\ \theta^1 \\ \vdots \\ \theta^G \end{pmatrix} \quad (6)$$

And then, the dimensionality of the redefined PsePSSM feature vector is 20 + G * 20. G is set to be 6 in the current study. Thus, 140-dimensional feature vectors are obtained for GPCR sequence.

Representing GPCR-drug pairs

We combine three different characteristics of GPCR's PseAAC feature, PsePSSM feature, and the drug's Wavelet feature. Then the formulated characteristics of the GPCR-Drug pairs are given by

$$G_{\text{PseAAC+PsePSSM+Wavelet\_FP}} = \begin{pmatrix} X_{\text{PseAAC+psePSSM}} \\ \lambda \cdot D_{\text{wavelet}} \end{pmatrix} = \begin{pmatrix} X_{\text{PseAAC}}^{\eta} \\ X_{\text{psePSSM}}^{g} \\ \lambda \cdot D_{\text{wavelet}} \end{pmatrix} \quad (7)$$

where G represents the combination feature of GPCR-Drug pairs, and λ is the weight coefficient, which is selected as 1/250 after optimization. So, 408-dimensional vectors are obtained to represent features of GPCR-drug pairs.

Deep Random Forest and Feature Augmentation

Random Forest (RF) (Breiman 2001) algorithm is a simple and effective ensemble learning technique proposed by Breiman in 2001. RF have demonstrated to be better or at least comparable to other state-of-the-art methods in both classification, semantic segmentation, and clustering applications (Lepetit et al. 2005).

Recently, deep neural networks (DNNs) (Ba and Caruana 2014; Liu et al. 2019) have become a dominant force in several areas. DNNs are powerful in handling feature relationships. Inspired by the success of deep neural network, Zhou & Feng (Zhou and Feng 2019) proposed deep forest approach which is based on non-differentiable modules and exhibits the possibility of constructing deep models without back-propagation. Essentially, deep forest is a novel decision tree ensemble method with predictive accuracy highly competitive with deep neural networks (DNN) in a broad range of tasks. Deep forest has much fewer hyper-parameters than deep neural networks, but experiments show that excellent performance can be obtained across various domains. Presentation learning in deep neural networks mainly relies on the layer-by-layer processing of raw features. Inspired by this recognition, a cascaded forest structure is designed for deep forest, and the outputs of the preceding layer are used as the input features of the next layer for training. This strategy has more robust feature learning ability than the single-layer model, which promotes the performance of the model. Inspired by this recognition, a procedure of multi-grained scanning is employed to enhance cascade forest. This strategy extracts the sequence or spatial relationships in the original features and ensures the diversity of training models. These strategies promote the success of deep forest in different tasks and inspire us to build deep random forests based on GPCR-drug data.

In deep random forests, each level of cascade receives the characteristic information processed by the upper level and outputs its processing results to the next level. Each level is a collection of decision tree forests. Here, we include different types of forests to encourage diversity. Each forest will

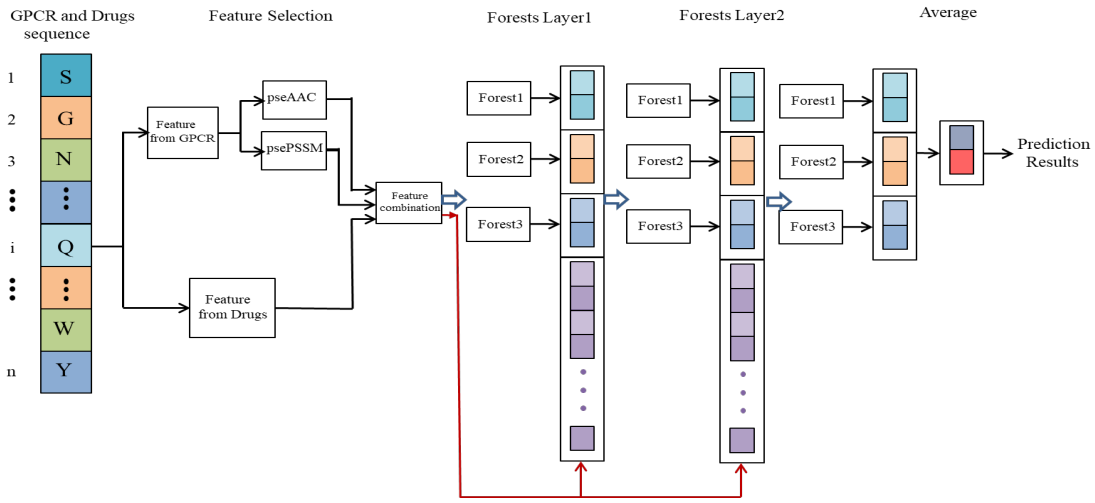

Figure 1. System architecture of the proposed framework for GPCR-drug interaction prediction

generate an estimate of the class distribution by calculating the percentage of training examples of different categories at the leaf node where the relevant instance is located, and then averaging across all trees in the same forest. The estimated class distribution forms a class vector, which is used as augmented feature and then connected with the original feature vector to be input to the next level of the cascade (refer to Figure 1).

Overall Procedure of the Proposed Prediction Framework

In order to provide an intuitive overall picture of how the classifier works, a workflow is provided to show the operation process of GPCR-drug interaction prediction in Figure 1. For a given GPCR-drug pair, firstly, the features of GPCR are extracted by pseAAC and psePSSM methods respectively, and the features of drug are extracted by fingerprint and wavelet method. Then, the three features are combined. We get the PseAAC+PsePSSM+Wavelet integrated features, which dimension is 140+140+128=408. Thirdly, the complex features are classified by a deep random forest prediction method. In deep random forests, two cascaded forest layers are used. In each level of the cascade, we select three different types of forests to add diversity. Here, forest1, forest2, and forest3 are sqrt, log2, and another random splitting nodes rule, respectively. Each forest generates 2-dimensional vectors (2 classes). The next level of cascade will receive 6 (=3x2) augmented features. The combined 414-dimensional features from 408 integrated features and the 6 augmented features are used as the input features of the next layer. The next layer proceeds as like the preceding layer. Finally, we use the average output of three forests and get the final prediction result.

Experimental Result and Discussion

Evaluation of Predictive Ability

The predictive ability of the present approach is evaluated with several measures (Yu et al. 2013), namely, Specificity (Spe), Sensitivity (Sen), the overall Accuracy (Acc), and the Matthews correlation coefficients (MCC). In addition to these four threshold-dependent evaluation indexes, we also exploited another evaluation index area under curve (AUC), which is the area under the Receiver Operating Characteristic (ROC) curve.

Performance Comparison on Different Feature Integration Methods

In this section, we evaluate the performance of the proposed method. As mentioned in section 1.2, there are three different combinations of the features of GPCR-Drug pairing: PseAAC+Wavelet, PsePSSM+Wavelet, and PseAAC+PsePSSM+Wavelet. We can select the best feature encoding system for GPCR-Drug pairing through the comparison experiment of these three combined features. In addition, the quality of the prediction results not only depends on the distinguishing ability of features, but also has a close relationship with the choice of the classifier. Different classifiers have different classification capabilities. Aiming at the training set of 1860 GPCR-Drug pairings, three different synthetic features are compared by using three different classifiers, which are support vector machine (SVM), random forest (RF), and our proposed deep random forest (DRF). Table 1 lists the comparative results of three feature combinations over leave-one-out cross-validation.

From the experimental results in Table 1, it can be found that: (1) Comparison of different features under the same classifier: For the SVM classifier, the evaluation indicators Acc and MCC of the PsePSSM+Wavelet feature are higher than those of the PseAAC+Wavelet feature; and the Acc and MCC of the PseAAC+PsePSSM+Wavelet feature are higher than those of the PsePSSM+Wavelet feature. Such as: PseAAC+PsePSSM+Wavelet feature is 0.684, PsePSSM+Wavelet feature is 0.661, and PseAAC+Wavelet feature is 0.648; an increase of 2.3% and 3.6% respectively. For the RF classifier and the DRF classifier, the experimental results are also similar. (2) On the same feature, the comparison of different classifiers: For the PseAAC+ Wavelet feature, it can be seen from the table that the Acc and *MCC* of the DRF classifier are higher than those of the RF classifier and the SVM classifier, while the Acc and *MCC* of the RF classifier are higher than those of the SVM classifier. For example, the Acc of the DRF classifier is 0.43% and1.45% higher than that of the RF classifier and SVM classifier, respectively. Similarly, it can be found that the PsePSSM+Wavelet feature and the PseAAC+PsePSSM+Wavelet feature also have similar results.

Table 1. Performance comparisons on different feature combinations and different classifiers over leave-one-out cross-validation

Classifiers	GPCR-drug features	*Sen* (%)	*Spe* (%)	*Acc* (%)	*MCC*	*AUC*
SVM	PseAAC+ Wavelet	74.52	89.52	84.52	0.648	0.872
	PsePSSM+ Wavelet	75.32	91.00	85.11	0.661	0.884
	PseAAC+ PsePSSM+Wavelet	80.00	88.79	85.86	0.684	0.905
RF	PseAAC+ Wavelet	78.55	89.03	85.54	0.675	0.898
	PsePSSM+ Wavelet	80.32	89.19	86.24	0.692	0.908
	PseAAC+ PsePSSM+Wavelet	80.48	90.48	87.15	0.710	0.917
DRF	PseAAC+ Wavelet	80.81	89.12	85.97	0.670	0.890
	PsePSSM+ Wavelet	80.16	90.32	86.94	0.706	0.914
	PseAAC+ PsePSSM+Wavelet	80.51	90.83	87.78	0.714	0.921

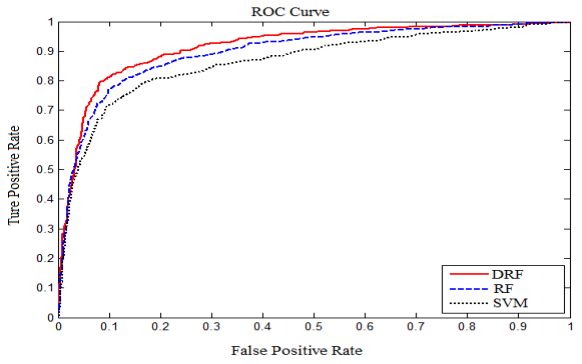

Figure 2. ROC curves of the performance comparison on PseAAC+ PsePSSM+ Wavelet features with DRF, SVM, and RF.

Figure 2 shows the corresponding ROC curves of the DRF, RF, and SVM classifiers based on the experimental results of the PseAAC+PsePSSM+Wavelet feature. It is obvious from the graph that the *AUC* of the solid line is a bit larger than the other lines, while the solid line represents the DRF. This shows that the *AUC* of the deep random forest classifier is significantly higher than that of the RF classifier and the SVM classifier. Similarly, it can be seen from the figure that the *AUC* of the RF classifier is larger than that of the SVM classifier.

Comparison with the Existing GPCR-drug Interaction Prediction Methods

In this section, we will experimentally demonstrate the efficacy of the proposed random forest classifier by comparing it with other existing classifiers/predictors. The comparison with the available methods includes leave-one-out cross validation on the training set and the independent verification test on the independent test set.

1) Comparison on the training set by leave-one-out cross-validation:

We compare the prediction effects of the developed method with three existing GPCR-drug interaction approaches: iGPCR-Drug (Xiao et al. 2013), RF-PPP (Hu et al. 2016), and WordBook (Wang et al. 2020). We conducted a comparative experiment on the 1860 GPCR-Drug pairs data in the cross-validation set. Table 2 lists the performance comparisons between the proposed method, iGPCR-Drug, RF-PPP, and WordBook.

Table 2. Performance comparison between different prediction methods on training set by leave-one-out cross-validation

Predictor	*Sen*(%)	*Spe*(%)	*Acc*(%)	*MCC*	*AUC*
iGPCR-Drug	80.00	88.30	85.50	0.678	N/A
RF-PPP	79.7	92.8	88.3	0.73	N/A
WordBook	81.1	87.1	85.1	0.67	N/A
Proposed method	80.51	90.83	87.78	0.714	0.921

From Table 2, we find that the cross validation of the developed method on the standard training set is better than iGPCR-Drug and WordBook methods. Compared with the iGPCR-Drug and WordBook, the *MCC* of the developed method is improved by 3.6% and 4.4% respectively. As can be seen from Table 2, except *Sen*, the *Spe*, *Acc* and *MCC* values of the proposed methods are lower than RF-PPP. This may be because the proposed method does not perform post-processing after classification.

2) Comparison of independent validation on the independent test set:

To further demonstrate the generalization performance of the proposed method used in this paper, we compare our method with other methods on the independent test set constructed in Section 2.

Table 3. Performance comparisons between the proposed method and other three predictors on the independent validation dataset.

Predictors	*Sen* (%)	*Spe* (%)	*Acc* (%)	*MCC*
iGPCR-Drug	80.8	66.9	71.6	0.45
RF-PPP	83.1	79.6	80.8	0.60
WordBook	83.1	82.7	82.8	0.63
Proposed method	82.5	83.6	83.4	0.64

Table 3 lists the comparative results of iGPCR-Drug, RF-PPP, WordBook predictors, and our method on the independent validation set. It can be found from Table 3 that the proposed method still achieves the best experimental results on the independent set. Compared with iGPCR-Drug, RF-PPP, and WordBook predictors, the *Spe* of the proposed method is increased by 16.7%, 4.0%, and 0.9%, respectively; Acc is increased by 11.8%, 2.6%, and 0.6%, respectively; *MCC* was increased by 0.19, 0.04, and 0.01, respectively. Experiments show that the proposed method is better in terms of *Spe*, *Acc*, and *MCC*. Note that the proposed method is slightly lower than RF-PPP and WordBook methods on the *Sen* index, which means that our method does not have the ability to improve the discrimination of positive subsets. But in terms of overall performance, the proposed method is still better than the other three predictors in terms of generalization ability.

Conclusions

In this study, we have proposed a novel identifying interactive GPCR-drug pair predictor with high accuracy, by integrating multi-view features and applying deep random forest classifier. The drug features are extracted by fingerprint and wavelet, and features of GPCR by the PseAAC and PsePSSM. The features of GPCR-drug pairs can be formulated by combining the features of GPCR and drugs. The good performances of the current model come from the systematic analysis and use of the most discriminative features and the cascaded random forest constructed based on feature augmentation. Compared with other existing predictors on the cross-validation set and independent test set, it is proved that the proposed method is not inferior to the existing GPCR-Drug interaction prediction method in terms of prediction accuracy and generalization ability. Compared with deep neural networks, deep forests are easy to train and have low computational overhead. The generation of each cascade uses cross-validation to avoid overfitting, and the tree structure has better interpretability. As many of the important bioinformatics topics to identify drug-target interaction can be formulated as machine learning problems, the current study provides an effective solution to these topics.

References

Ba, J.; and Caruana, R. 2014. Do deep nets really need to be deep? *Advances in neural information processing systems* 27.

Breiman, L. 2001. Random forests. *Machine learning* 45: 5-32.

Chou, K. C. 2005. Prediction of G-protein-coupled receptor classes. *Journal of proteome research* 4: 1413-1418.

Dou, Y.; Wang, J.; Yang, J.; and Zhang, C. 2012. L1pred: a sequence-based prediction tool for catalytic residues in enzymes with the L1-logreg classifier. *PloS one* 7, e35666.

He, Z.; Zhang, J.; Shi, X.-H.; Hu, L.-L.; Kong, X.; Cai, Y.-D. *et al.* 2010. Predicting drug-target interaction networks based on functional groups and biological features. *PloS one* 5, e9603.

Heuss, C.; Ger, U. 2000. G-protein-independent signaling by G-protein-coupled receptors. *Trends in neurosciences* 23: 469-475.

Hu, J.; Li, Y.; Yang, J. Y.; Shen, H. B.; and Yu, D. J. 2016. GPCR–drug interactions prediction using random forest with drug-association-matrix-based post-processing procedure. *Computational biology and chemistry* 60: 59-71.

Lappano, R.; and Maggiolini, M. 2011. G protein-coupled receptors: novel targets for drug discovery in cancer. *Nature reviews Drug discovery* 10: 47-60.

Lepetit, V.; Lagger, P.; and Fua, P. 2005. Randomized trees for real-time keypoint recognition. *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 2, pp. 775-781. IEEE.

Liu, G. H.; Zhang, B. W.; Qian, G.; Wang, B.; Mao, B.; and Bichindaritz, I. 2019. Bioimage-based prediction of protein subcellular location in human tissue with ensemble features and deep networks. *IEEE/ACM transactions on computational biology and bioinformatics* 17: 1966-1980.

O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; and Hutchison, G. R. 2011. Open Babel: An open chemical toolbox. *Journal of cheminformatics* 3: 1-14.

Percival, D. B.; and Walden, A. T. (2006) *Wavelet methods for time series analysis*. Cambridge University Press.

Schäffer, A. A.; Aravind, L.; Madden, T. L.; Shavirin, S.; Spouge, J. L.; Wolf, Y. I. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic acids research* 29: 2994-3005.

Sirois, S.; Hatzakis, G.; Wei, D.; Du, Q.; and Chou, K.-C. 2005. Assessment of chemical libraries for their druggability. *Computational biology and chemistry* 29: 55-67.

Wang, P.; Huang, X.; Qiu, W.; and Xiao, X. 2020. Identifying GPCR-drug interaction based on wordbook learning from sequences. *BMC bioinformatics* 21: 1-17.

Xiao, X.; Min, J.-L.; Wang, P.; and Chou, K.-C. 2013. iGPCR-Drug: A web server for predicting interaction between GPCRs and drugs in cellular networking. *PloS one* 8, e72234.

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

Zhou, Z.-H.; and Feng, J. 2019. Deep forest. *National Science Review* 6: 74-86.