DTIRF: Predicting Drug-Target Interactions Based on Improved Rotation Forest from Drug Molecular Structure and Protein Sequence

lei wang (leiwang@ms.xjb.ac.cn)
Chinese Academy of Sciences https://orcid.org/0000-0003-0184-307X

Zhu-Hong You
Chinese Academy of Sciences

Li-Ping Li
Chinese Academy of Sciences

Xin Yan
Zaozhuang University

Wei Zhang
Zaozhuang University

Hai-Feng Wang
Zaozhuang University

Research article

Keywords: drug-target interaction; rotation forest; pseudo position-specific score matrix

Posted Date: October 31st, 2019

DOI: https://doi.org/10.21203/rs.2.15799/v2

Abstract

Background: The identification and prediction of Drug-Target Interaction (DTI) is the basis for screening drug candidates, which plays a vital role in the development of innovative drugs. However, due to the time-consuming and high cost constraints of biological experimental methods, traditional drug target identification technologies are often difficult to develop on a large scale. Therefore, in silico methods are urgently needed to predict drug-target interactions in a genome-wide manner. Results: In this article, we design a new in silico approach, named DTIRF, to predict the DTI combine feature weighted Rotation Forest (FwRF) classifier with protein amino acids information. More specifically, we first use Position-Specific Score Matrix (PSSM) to numerically convert protein sequences and utilize Pseudo Position-Specific Score Matrix (PsePSSM) to extract their features. Then a unified digital descriptor is formed by combining molecular fingerprints representing drug information. Finally, the feature weighted rotation forest is applied to implement on Enzyme, Ion Channel, GPCR, and Nuclear Receptor data sets. The results of the five-fold cross-validation experiment show that the prediction accuracy of this approach reaches 91.68%, 88.11%, 84.72% and 78.33% on four benchmark data sets, respectively. To further validate the performance of the DTIRF, we compare it with other excellent methods and Support Vector Machine (SVM) model. In addition, 7 of the 10 highest predictive scores in predicting novel DTIs were validated by relevant databases. Conclusions: The experimental results of cross-validation indicated that DTIRF is feasible in predicting the relationship among drugs and target, and can provide help for the discovery of new candidate drugs.

Background

Identifying the interactionship among drugs and targets is a crux area in genomic drug discovery, which not only helps to understand various biological processes, but also contributes to the development of new drugs [1, 2]. The emergence of molecular medicine and the completion of the Human Genome Project provide better conditions for the identification of new drug target proteins. Although the researchers have made a lot of efforts, only a small number of candidate drugs can be approved by the Food and Drug Administration (FDA) to enter the market so far [3-5]. An important reason for this situation is due to the inherent defects of the experimental methods. As is known to all, biological laboratory methods to identify DTI are usually limited to small-scale studies, and are expensive and time-consuming. Computational methods can narrow the scope of candidate targets and provide supporting evidence for the drug target experiments, thus speeding up drug discovery. Therefore, computation-based methods are urgently required to improve efficiency and reduce time in identifying potential DTIs across the genome. [2, 4, 6-8].

In recent years, researchers have developed a variety of computation-based methods to analyze and predict DTI, which can be broadly divided into two categories: network-based methods and machine learning-based methods [9]. Network-based methods usually describe the relationship between drugs and target proteins as a heterogeneous network, and predict DTI by analyzing the correlation and similarity between nodes. For example, Wu *et al.* [10] proposed the SDTBNI model in 2016, which searches for

unknown DTIs through new chemical entity-substructure linkages, drug-substructure linkages and known DTI networks. Zhang *et al.* [11] proposed a novel DTI prediction model based on LPLNI. The model uses data points reconstructed from neighborhood to calculate the linear neighborhood similarity of drug-drug. Based on biomedical related data and Linked Tripartite Network (LTN), Zong *et al.* [12] used the target-target and drug-drug similarities calculated by DeepWalk to predict DTI.

Machine learning-based methods usually extract the features of drug and target data by related algorithms, and effectively predict potential DTIs by supervised or semi-supervised methods. For example, Peng *et al.* [13] combines the biological information of targets and drugs with PCA-based convex optimization algorithms to predict new DTIs using semi-supervised inference method. Ezzat *et al.* [14] used ensemble learning algorithm to predict DTI by decrease features with subinterval features through three dimensionality reduction models. Generally speaking, drugs with chemical similarity also have similar biochemical activity, that is, they can bind to similar target proteins. Based on the above assumptions, the use of medicinal chemical molecular structure information and protein sequence information to predict the DTI model has achieved good results. For example, Wen *et al.* [15] extracted drug and target features from their chemical substructure and sequence information, and used deep belief network (DBN) to predict potential DTI. The model proposed in this paper belongs to the machine learning-based method based on this assumption.

In this article, according to the assumption that the interaction among drugs and target proteins largely depend on the information of target protein sequences and drug molecular sub-structure fingerprints, a novel machine learning-based model is proposed to infer potential DTI. Our feature combines the fingerprint of the drug molecule structure and the protein sequence encoded by a feature extraction method called Pseudo Position-Specific Score Matrix (PsePSSM). In the experiment, we adopt the feature weighted rotation forest classifier (FwRF) to predict the results on the four DTI benchmark data sets, including *Enzyme, Ion Channel, GPCR* and *Nuclear Receptor*. In order to verify the performance of the proposed model, we also compare with SVM classifier model, different feature extraction models and existing excellent methods. The promoting experimental results have shown that the feature extraction is very effective, and the designed classifier has high recognition performance.

Results And Discussion

Evaluation Criteria

In this paper, accuracy (Accu.), sensitivity (Sen.), precision (Prec.), and Matthews correlation coefficient (MCC) are used to estimate the performance of DTIRF. Their formulas are as follows: (see Formulas 1 - 4 in the Supplementary Files)

where TP is the number of drug-target pairs that are related to each other to be correctly identified; FP is the number of drug-target pairs that are related to each other to be incorrectly identified; TN is the number of drug-target pairs that are not related to each other to be correctly identified; FN is the number of drug-target pairs that are not related to each other to be incorrectly identified. Moreover, the receiver operating

characteristic (ROC) curve [16, 17] and area under the ROC curve (AUC) are used to visually display the performance of the classifier.

Model Construction

To optimize the performance of the DTIRF, the grid search method is applied to explore the parameters of PsePSSM and FwRF. When extracting feature by PsePSSM, the parameters in the formula 5 can be adjusted to increase the amount of information. In the experiment we explored the effects of different PsePSSM parameters on the performance of classifiers on *Enzyme* data set. After optimization, we set the parameter of PsePSSM to 34, and the parameters the feature selection ratio , the feature subset and the decision tree number of FwRF classifier to 0.8, 16 and 21, respectively. Figure 1 display the prediction results of different FwRF parameters, where an optimal choice of K=16 and L=21 are finally selected.

Evaluation of Model Prediction Ability

After finding the optimal parameters of the DTIRF, we put them in benchmark data sets, including *Enzyme*, *Ion Channel*, *GPCR* and *Nuclear Receptor.* In order to avoid over-fitting of the model, we use five-flod cross-validation method to evaluate the performance of the model. More specifically, we split the data set into five subsets, one of which is taken as the test set, and the remaining four are used as the training set. Then, the cross-validation process will be repeated five rounds. The results from the 5 times are then averaged to produce the final result.

Table 1-4 list the predicted results by the proposed approach on four benchmark data sets. In *Enzyme* data set, we gained the average of accuracy, sensitivity, precision, MCC, and AUC were 91.68%, 90.84%, 92.39%, 83.39%, and 91.72%. Their standard deviations were 0.84%, 1.68%, 1.37%, 1.68%, and 1.06%. In *Icon Channel* data set, we achieved these evaluation criteria were 88.11%, 90.30%, 86.57%, 79.02%, and 88.27%. Their standard deviations were 1.01%, 1.61%, 2.29%, 1.55%, and 1.36%. In *GPCR* data set, we yielded the average of these evaluation criteria were 84.72%, 84.73%, 84.73%, 74.06%, and 85.57%. Their standard deviations were 1.94%, 3.45%, 4.21%, 2.68%, and 2.28%. In *Nuclear Receptor* data set, we gained the average of these evaluation criteria were 78.33%, 81.97%, 78.08%, 65.56%, and 75.31%. Their standard deviations were 5.34%, 7.85%, 12.56%, 6.05%, and 5.87%. Figure 2-5 draws the ROC curve generated from DTIRF on the four benchmark data sets.

Table 1. Experimental results of cross-validation of the proposed model on *Enzyme* data set.

Test set	Accu.(%)	Sen.(%)	Prec.(%)	MCC(%)	AUC(%)
1	90.51	89.20	91.27	81.03	90.04
2	92.82	93.22	92.59	85.64	92.96
3	91.62	90.19	92.74	83.28	92.09
4	91.97	89.68	94.40	84.05	91.79
5	91.47	91.90	90.96	82.94	91.73
Average	**91.68±0.84**	**90.84±1.68**	**92.39±1.37**	**83.39±1.68**	**91.72±1.06**

Table 2. Experimental results of cross-validation of the proposed model on *Icon Channel* data set.

Test set	Accu.(%)	Sen.(%)	Prec.(%)	MCC(%)	AUC(%)
1	86.61	90.38	83.76	76.76	86.16
2	87.63	91.92	84.78	78.22	87.83
3	88.98	91.61	87.22	80.36	89.68
4	88.31	89.67	87.62	79.33	89.07
5	89.02	87.93	89.47	80.43	88.59
Average	**88.11±1.01**	**90.30±1.61**	**86.57±2.29**	**79.02±1.55**	**88.27±1.36**

Table 3. Experimental results of cross-validation of the proposed model on *GPCR* data set.

Test set	Accu.(%)	Sen.(%)	Prec.(%)	MCC(%)	AUC(%)
1	82.28	86.21	77.52	70.73	82.86
2	87.01	88.62	85.16	77.38	88.82
3	86.22	86.52	88.41	76.00	86.72
4	84.63	80.33	86.73	73.83	84.37
5	83.46	81.95	85.83	72.37	85.11
Average	**84.72±1.94**	**84.73±3.45**	**84.73±4.21**	**74.06±2.68**	**85.57±2.28**

Table 4. Experimental results of cross-validation of the proposed model on *Nuclear Receptor* data set.

Test set	Accu.(%)	Sen.(%)	Prec.(%)	MCC(%)	AUC(%)
1	69.44	83.33	65.22	55.90	72.22
2	77.78	85.00	77.27	64.34	69.69
3	80.56	92.31	66.67	67.47	74.25
4	83.33	77.78	87.50	72.05	75.31
5	80.56	71.43	93.75	68.03	85.08
Average	**78.33±5.34**	**81.97±7.85**	**78.08±12.56**	**65.56±6.05**	**75.31±5.87**

Due to the high redundancy of DTI network, in order to fully evaluate the performance of the proposed model, we also use leave-one-cluster-out cross-validation method to experiment on four benchmark data sets. Leave-one-cluster-out cross-validation uses a cluster-based (i.e., similarity-driven) approach to separating datasets into training sets and test sets that systematically explore the effects of non-uniform training data. Specifically, we implemented a series of leave-one-cluster-out cross-validation at the protein and drug levels on the benchmark data sets. Taking protein as an example, our implementation process is as follows: firstly, we perform a standard normalization of the input data; secondly, we use Yamanishi's "Protein sequence similarity matrix" to cluster proteins at 0.4 thresholds; thirdly, each time one cluster is used as the validation set and the remaining clusters are used as the training set to perform leave-one-cluster-out cross-validation; finally, we summarize the results of each cluster to get the final results. Similarly, for drugs we use Yamanishi's "Compound structure similarity matrix" for clustering and use the

same method. The results of leave-one-cluster-out cross-validation on the four benchmark data sets are shown in table 5. It can be seen from the table that the proposed model can also achieve good results under the leave-one-cluster-out cross-validation method. These results indicate that the proposed model has good robustness and can effectively predict the relationship between drugs and targets.

Table 5. Experimental results of leave-one-cluster-out cross-validation of the proposed model on four benchmark data sets

Data set	level	Accu.(%)	Sen.(%)	Prec.(%)	MCC(%)	AUC(%)
Enzyme	drug	90.67	91.18	90.26	81.34	90.44
	protein	91.32	91.08	91.52	82.64	91.46
	average	90.99	91.13	90.89	81.99	90.95
Icon Channel	drug	88.52	90.79	86.84	77.11	88.68
	protein	87.77	90.31	85.94	75.64	88.22
	average	**88.14**	**90.55**	**86.39**	**76.38**	**88.45**
GPCR	drug	84.72	86.46	83.56	69.49	82.81
	protein	81.97	82.99	81.33	63.95	79.71
	average	**83.35**	**84.72**	**82.44**	**66.72**	**81.26**
Nuclear Receptor	drug	77.78	74.44	79.76	55.68	76.76
	protein	76.67	80.00	75.00	53.45	77.65
	average	**77.22**	**77.22**	**77.38**	**54.57**	**77.21**

Comparison between the proposed model and LPQ descriptor models

To evaluate the impact of PsePSSM algorithm on the proposed model, we compare it with Local Phase Quantization (LPQ) on four benchmark data sets in this section. The LPQ feature extraction algorithm is based on the blur invariance property of the Fourier phase spectrum [18-20] and originally described in the article for texture description by Ojansivu and Heikkila [21]. Table 6 summarizes the cross-validation results generated by LPQ algorithm combined with FwRF classifier on four benchmark data sets. From the table we can see that DTIRF has achieved the best results in all the evaluation indicators including accuracy, sensitivity, precision, MCC and AUC. Detailed five-fold cross-validation results on four benchmark data sets are presented in Supplementary Materials Table S1-S4. In the comparison experiment, we set the same parameters for the FWRF classifier. We can see from the comparison results that PsePSSM algorithm combined with FwRF classifier does helps to improve the performance of the model.

Table 6. Experimental results of the FwRF classifier combined with LPQ algorithm on four benchmark data sets.

Data set	Method	Accu.(%)	Sen.(%)	Prec.(%)	MCC(%)	AUC(%)
Enzyme	FwRF+LPQ	89.63±0.39	89.69±1.82	89.64±2.16	79.32±0.79	89.40±0.98
	DTIRF	**91.68±0.84**	**90.84±1.68**	**92.39±1.37**	**83.39±1.68**	**91.72±1.06**
Icon Channel	FwRF+LPQ	83.97±2.32	86.93±3.03	81.89±3.66	68.13±4.54	84.66±2.01
	DTIRF	**88.11±1.01**	**90.30±1.61**	**86.57±2.29**	**79.02±1.55**	**88.27±1.36**
GPCR	FwRF+LPQ	82.52±2.17	83.87±3.58	81.79±3.78	65.19±4.15	83.19±1.79
	DTIRF	**84.72±1.94**	**84.73±3.45**	**84.73±4.21**	**74.06±2.68**	**85.57±2.28**
Nuclear Receptor	FwRF+LPQ	66.67±7.35	67.64±16.23	67.97±9.98	35.46±10.89	69.56±6.85
	DTIRF	**78.33±5.34**	**81.97±7.85**	**78.08±12.56**	**65.56±6.05**	**75.31±5.87**

Comparison between FwRF and SVM classifier models

As the most versatile Support Vector Machine (SVM) classifier has been widely used by various problems. In order to estimate DTIRF clearly, we compare the results of DTIRF and SVM classifier model on the same data set. The SVM parameters are determined by grid search, and finally set the value of c to 0.5 and the value of g to 0.6. The results of the SVM classifier optimization can be viewed in the supplementary materials table S9. From the table we can see that DTIRF has achieved excellent results on the four benchmark balance data sets. Among the evaluation parameters Accuracy, sensitivity, MCC and AUC, DTIRF have achieved the highest results, and DTIRF on precision is only slightly lower than that of SVM model in *Enzyme* and *Icon Channel* data sets. Detailed five-fold cross-validation results on four benchmark data sets are presented in Supplementary Materials Table S5-S8. This result indicates that the FwRF classifier is suitable for the proposed model and can effectively improve the performance of the model.

Table 7. Experimental results of the SVM classifier model on four benchmark data sets

Data set	Method	Accu.(%)	Sen.(%)	Prec.(%)	MCC(%)	AUC(%)
Enzyme	PsePSSM +SVM	84.20±0.60	69.90±1.70	**98.00±0.50**	71.50±1.00	84.30±1.20
	DTIRF	**91.68±0.84**	**90.84±1.68**	92.39±1.37	**83.39±1.68**	**91.72±1.06**
Icon Channel	PsePSSM +SVM	81.90±1.20	69.70±3.70	**92.40±2.20**	66.00±1.90	81.70±1.20
	DTIRF	**88.11±1.01**	**90.30±1.61**	86.57±2.29	**79.02±1.55**	**88.27±1.36**
GPCR	PsePSSM +SVM	70.00±2.10	50.40±7.80	82.30±3.30	42.80±4.90	70.10±2.70
	DTIRF	**84.72±1.94**	**84.73±3.45**	**84.73±4.21**	**74.06±2.68**	**85.57±2.28**
Nuclear Receptor	PsePSSM +SVM	63.30±3.60	57.60±7.90	67.50±14.60	29.60±7.40	61.80±5.80
	DTIRF	**78.33±5.34**	**81.97±7.85**	**78.08±12.56**	**65.56±6.05**	**75.31±5.87**

Comparison with existing methods

The prediction of the relationship between drugs and targets has drawn increasing interest of researchers. So far, a lot of excellent computational approaches have been designed. To better verify the proposed approach, we compare it with other existing methods using five-fold cross-validation on the same benchmark data sets. Table 8 lists the details of other excellent methods and DTIRF on four benchmark data sets in terms of the AUC. It is seen that the results obtained by DTIRF on *Enzyme* data set are significantly higher than those of other existing methods, and the results achieved on *Icon Channel* and *GPCR* data sets by DTIRF only lower than the highest result 0.73% and 0.13%. The performance of DTIRF on *Nuclear Receptor* data set is not very good, it may be because the sample number of the *Nuclear Receptor* data set is too small, and the training of the classifier is not sufficient

Table 8. Performances of other excellent methods and DTIRF on four benchmark data sets in terms of the AUC.

Data set	DTIRF	KBMF2K [22]	NetCBP [23]	SIMCOMP[24]	RFDT[25]
Enzyme	0.9172	0.832	0.8251	0.863	0.915
Icon Channel	0.8827	0.799	0.8034	0.776	0.890
GPCR	0.8557	0.857	0.8235	0.867	0.845
Nuclear Receptor	0.7531	0.824	0.8394	0.856	0.723

Case study

To further validate DTIRF's ability to predict potential DTI, we use all known interactions to train the model and then predict unknown interactions. We selected 10 drug-target pairs with the highest predictive score to validate in SuperTarget [26]. SuperTarget is a database that collects drug-target relations and currently stores 332,828 DTIs. As shown in Table 9, 7 of the top 10 predicted highest scores were confirmed. This result indicates that DTIRF can effectively predict the potential DTIs. It is worth noting that although we have not found evidence of the interaction of the remaining 3 drug-target pairs, we can not completely deny the possibility of their interactions.

Table 9.The top 10 new predicted interactions by DTIRF

Drug ID	Drug name	Taregt protein ID	Target protein name	Validation source
D00691	Dihydroxypropyltheophylline	hsa5150	PDE7A_HUMAN	SuperTarget
D00348	Isotretinoino	hsa6256	RXRA_HUMAN	SuperTarget
D00139	Xanthotoxine	hsa1543	CP1A1_HUMAN	SuperTarget
D02340	Loxapinsuccinate	hsa1812	DRD1_HUMAN	SuperTarget
D00493	Prochlorpermazine	hsa3356	5HT2A_HUMAN	unconfirmed
D00542	Bromochlorotrifluoroethane	hsa1571	CP2E1_HUMAN	SuperTarget
D00585	Mifepristone	hsa2099	ESR1_HUMAN	SuperTarget
D00454	Olanzapine	hsa1813	DRD2_HUMAN	unconfirmed
D03365	Transdermal Nicotine	hsa1137	ACHA4_HUMAN	SuperTarget
D00106	Epoprostenol	hsa5733	PE2R3_HUMAN	unconfirmed

Materials And Methodology

Benchmark data sets

In this article, we applied four protein targeting data sets, including *Enzyme, Ion Channel, GPCR* and *Nuclear Receptor*. These data sets are applied as the benchmark data sets by Yamanishi et al. [27] and collected from the BRENDA [28], DrugBank [29], SuperTarget & Matador [26] and KEGG BRITE [30]. They can be downloaded at http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/ [27]. The number of drugs was 445, 210, 233 and 54, and the number of target proteins was 664, 204, 95 and 26 in these benchmark data sets, respectively. Among these data, 5127 pairs of drug-target were confirmed to interact with each other, corresponding to 2926, 1476, 635 and 90 pairs in four data sets, respectively [25].

The DTI network can be expressed by a bipartite graph in which nodes represent drugs or targets, and edges represent their interactions. If there is a relationship between the nodes, connect them with edges, otherwise they do not connect. The edges in the initial bipartite graph represent the real DTI have been detected by biological experiments. The number of initial edges is relatively small compared to a completely connected bipartite graph. [31-33]. For example, there are totally 54×26=1404 connections in the bipartite graph of the *Nuclear Receptor* data set. However, the experiment detected representatives known drug-target interactions only 90. Therefore, the number of positive drug-target pairs (e.g., 90) accounted for only 6.41% of the total number of drug-target pairs (e.g., 1404), much less than the number of negative drug-target pairs (e.g., 1404-90=1314). The same problem also appears in the other three data sets. In order to solve the problem of data imbalance, we randomly select negative drug-target pairs with the same number of positive drug-target pairs. In fact, such negative samples may contain drug-target pairs with interactions. To reduce this possibility, we randomly selected three times negative sample sets in the experiment. From a statistical point of view, the number of real interactions on a large bipartite graph is chosen as the negative sample set is very small. Eventually, the negative sample numbers of *Enzyme, Ion Channel, GPCR* and *Nuclear Receptor* data sets are 2926, 1476, 635 and 90, respectively.

Molecules description

In recent years, different types of descriptors have been proposed to represent drug compounds, such as quantum chemical properties, topological, constitutional and geometrical. Since the molecular substructure fingerprint does not require the three-dimensional structural information of the molecule and has the advantage of directly reflecting the relationship between molecular properties and structure, more and more researchers use it as a descriptor to predict the relationship between the drug and the target protein. Specifically, we first store all the molecular substructures in the form of a dictionary, and then split a given drug molecule. When it contains a certain substructure, the corresponding bit of the descriptor is assigned to 1; otherwise it is assigned to 0. Finally, we get the drug molecule in the form of Boolean vectors. In the experiment, we use the chemical structure fingerprint set from PubChem System, and the fingerprints property is "PUBCHEM_CACTVS_SUBGRAPHKEYS" in PubChem. A drug fingerprint is recorded as 881 substructures, so the drug molecule feature is the 881-dimensional. Since the drug fingerprint is divided into 881 substructures, the dimension of the drug molecular fingerprint descriptor is 881 dimensions.

Numerical characterization of protein sequences

In the experiment, we used position-specific scoring matrix proposed by Gribskov et al. [34] to convert protein sequence numerically. PSSM is widely used in protein binding site prediction [35, 36], protein secondary structure prediction [37], and prediction of disordered regions [38]. PSSM is an $L \times 20$ matrix that can be expressed as , where 20 represents the number of the amino acids and L denotes the length of the protein sequence. PSSM matrix can be expressed as follows: (see Formula 5 in the Supplementary Files)

where denotes the probability that the ith residue being mutated into the jth amino acid during the evolutionary process of protein multiple sequence alignment.

In the experiment, we use Position-Specific Iterated BLAST (PSI-BLAST) tool [39, 40] to generate PSSM based on *SwansProt* data set. PSI-BLAST can generates the 20-dimensional vector indicating the mutation conservation probabilities of 20 different amino acids. To obtain high homologous and broad homologous sequences, we set the parameter iterations to 3 and the parameter e-value to 0.001, and keep other parameters as default values. The *SwissProt* database and PSI-BLAST toolkit can be downloaded at http://blast.ncbi.nlm.nih.gov/Blast.cgi.

Feature extraction algorithm

Effective protein feature descriptors can not only mine useful information, but also improve the performance of the approach. In this study, we introduce the feature extraction algorithm Pseudo Position-Specific Score Matrix (PsePSSM), which concept from Chou *et al.* [41]. The PsePSSM is expressed by formula as follows: (see Formula 6 in the Supplementary Files)

where denotes the raw score generated by PSI-BLAST, which is typically a positive or negative integer. This is not the final score, because if it exceeds 20 amino acids, the score may contain 0; if the same conversion procedure continues, the score may remain unchanged. The positive number signifies that the frequency of corresponding mutations in the alignment is higher than that of accidental expectations. Conversely, the negative number signifies that the frequency of corresponding mutations in the alignment is lower than that of accidental expectations. However, base on the PSSM formula, proteins of different lengths will produce a matrix of different numbers of rows. Therefore, equation 3 is used to convert the PSSM matrix into a uniform pattern. (see Formulas 7 and 8 in the Supplementary Files)

where indicates the average score of *P* protein when amino acid residues evolve into j-type amino acids. However, if only is used to indicate protein *P*, all information about sequence order will be lost during evolution. In order to prevent this from happening, we introduce the idea of pseudo amino acid to improve equation 3. Therefore, according to the formula 5, we can get the features of segmented PsePSSM: (see Formula 9 in the Supplementary Files)

where is a related factor for j-type amino acid, whose contiguous distance is along each segmented protein sequence. The flow chart of the proposed model is shown below.

Feature Weighted Rotation Forest Classifier

In this paper, the feature weighted rotation forest (FwRF) is proposed. Compared with the original rotation forest, it adds the function of weight selection. Through this function, we can remove the noise features with small weights, thus increasing the content of useful information and improving the accuracy of prediction. The weights of the features are calculated by statistical method. The feature *F* for the class can be obtained by the following formula. (see Formula 10 - 12 in the Supplementary Files)

The implementation steps of feature weighted rotation forest are as follows: Firstly, the weights of all features are calculated by equation 6; secondly, the features are sorted according to the weights; finally, the desired features are selected according to the given feature selection rate *r*. After performing these steps, we get a new data set and send it to rotation forest.

Original Rotation forest (RF) [42, 43] is a widely used classifier algorithm. Assuming contains *S* training samples, where in be an n-dimensional feature vector. Let *X* is the training sample set, *Y* is the corresponding labels and *F* is the feature set. Then *X* is *S*×*n* matrix, which is composed of n observation feature vector composition. Assuming that the number of decision trees is *N*, then the decision trees can be expressed as . The algorithm is executed in the following steps.

(1) Using the appropriate parameter K to randomly divide *F* into *K* independent and uncrossed subsets, the number of each subset feature is .

(2) A corresponding column of features in the subset is selected from the training set *X* to form a new matrix . Then, 75% of the data is extracted from *X* in the form of bootstrap to form a new set .

(3) Use matrix as the feature transform to generate coefficients in matrix .

(4) Using the coefficients obtained from the matrix to form a sparse rotation matrix , the expression of which is as follows: (see Formulas 13 and 14 in the Supplementary Files)

Conclusions

Prediction of DTI is a crucial problem for human medical improvement and genomic drug discovery. Under the hypothesis that the drug molecules structures and protein amino acids sequence have a big impact on the relationships among drugs and target proteins, the *in silico* approach is proposed to infer potential drug-target relationships in this article. We implement it on *Enzyme*, *Icon Channel, GPCR* and *Nuclear Receptor* data sets, and obtained excellent results. To further evaluate the performance of the proposed approach, we compared it with PsePSSM algorithm model, the SVM classifier model and other existing methods on the same data sets. Competitive cross-validation experimental results show that the performance of DTIRF has been significantly improved, which demonstrated DTIRF is stable and reliable. In the next study, we plan to try more feature extraction algorithm to better predict DTI.

Abbreviations

DTI: Drug-Target Interaction; FwRF: feature weighted Rotation Forest; PSSM: Position-Specific Score Matrix; PsePSSM: Pseudo Position-Specific Score Matrix; SVM: Support Vector Machine; FDA: Food and Drug Administration; LTN: Linked Tripartite Network; DBN: Deep Belief Network

Declarations

Ethics approval and consent to participate

Not applicable.

Consent to publication

Not applicable.

Availability of data and materials

Source code and dataset can be found: https://github.com/look0012/DTIRF.

Competing interests

The authors declare that they have no competing interests

Funding

This work is supported in part by the National Natural Science Foundation of China, under Grants 61572506, 61702444, and in part by the Pioneer Hundred Talents Program of Chinese Academy of Sciences, and in part by the CCF-Tencent Open Fund, in part by the Chinese Postdoctoral Science Foundation, under Grant 2019M653804, and in part by the West Light Foundation of The Chinese Academy of Sciences, under Grant 2018-XBQNXZ-B-008.

Authors' Contributions

LW, ZHY and WZ considered the algorithm, make analyses, arranged the data sets, carried out experiments, and wrote the manuscript. LPL, XY and HFW designed, performed and analyzed experiments. All authors read and approved the final manuscript.

Acknowledgments

The authors would like to thank all anonymous reviewers for their constructive advices.

Authors' Information

[1]College of Information Science and Engineering, Zaozhuang University, Zaozhuang, 277100, China; [2]Xinjiang Technical Institutes of Physics and Chemistry, Chinese Academy of Sciences, Urumqi, 830011, China; [3]School of Foreign Languages, Zaozhuang University, Zaozhuang, 277100, China.

References

1. Xia Z, Wu L-Y, Zhou X, Wong STC: **Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces**. *Bmc Systems Biology* 2010, **4**.

2. Wang Y-C, Yang Z-X, Wang Y, Deng N-Y: **Computationally Probing Drug-Protein Interactions Via Support Vector Machine**. *Letters in Drug Design & Discovery* 2010, **7**(5):370-378.
3. Landry Y, Gies J-P: **Drugs and their molecular targets: an updated overview**. *Fundamental & Clinical Pharmacology* 2008, **22**(1):1-18.
4. Li Q, Lai L: **Prediction of potential drug targets based on simple sequence properties**. *Bmc Bioinformatics* 2007, **8**.
5. van de Waterbeemd H, Gifford E: **ADMET in silico modelling: Towards prediction paradise?** *Nature Reviews Drug Discovery* 2003, **2**(3):192-204.
6. Kuruvilla FG, Shamji AF, Sternson SM, Hergenrother PJ, Schreiber SL: **Dissecting glucose signalling with diversity-oriented synthesis and small-molecule microarrays**. *Nature* 2002, **416**(6881):653-657.
7. Haggarty SJ, Koeller KM, Wong JC, Butcher RA, Schreiber SL: **Multidimensional chemical genetic analysis of diversity-oriented synthesis-derived deacetylase inhibitors using cell-based assays**. *Chemistry & Biology* 2003, **10**(5):383-396.
8. Wang L, You ZH, Chen X, Li JQ, Yan X, Zhang W, Huang YA: **An ensemble approach for large-scale identification of protein-protein interactions using the alignments of multiple sequences**. *Oncotarget* 2017, **8**(3):5149.
9. Chen X, Yan CC, Zhang X, Zhang X, Dai F, Yin J, Zhang Y: **Drug-target interaction prediction: databases, web servers and computational models**. *Briefings in Bioinformatics* 2016, **17**(4):696.
10. Wu Z, Cheng F, Li J, Li W, Liu G, Tang Y: **SDTNBI: an integrated network and chemoinformatics tool for systematic prediction of drug–target interactions and drug repositioning**. *Briefings in Bioinformatics* 2017, **18**(2):333-347.
11. Zhang W, Chen Y, Li D: **Drug-Target Interaction Prediction through Label Propagation with Linear Neighborhood Information**. *Molecules* 2017, **22**(12):2056.
12. Zong N, Kim H, Ngo V, Harismendy O: **Deep Mining Heterogeneous Networks of Biomedical Linked Data to Predict Novel Drug-Target Associations**. *Bioinformatics* 2017, **33**(15).
13. Peng L, Liao B, Zhu W, Li Z, Li K: **Predicting Drug-Target Interactions With Multi-Information Fusion**. *IEEE Journal of Biomedical & Health Informatics* 2017, **21**(2):561-572.
14. Ezzat A, Wu M, Li XL, Kwoh CK: **Drug-Target Interaction Prediction using Ensemble Learning and Dimensionality Reduction**. *Methods* 2017, **129**:81.
15. Wen M, Zhang Z, Niu S, Sha H, Yang R, Yun Y, Lu H: **Deep-Learning-Based Drug-Target Interaction Prediction**. *Journal of Proteome Research* 2017, **16**(4):1401.
16. Zweig MH, Campbell G: **Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine**. *Clinical chemistry* 1993, **39**(4):561-577.
17. Wang L, Wang H-F, Liu S-R, Yan X, Song K-J: **Predicting Protein-Protein Interactions from Matrix-Based Protein Sequence Using Convolution Neural Network and Feature-Selective Rotation Forest**. *Scientific reports* 2019, **9**(1):9848.

18. Wang H, Song A, Li B, Xu B, Li Y: **Psychophysiological classification and experiment study for spontaneous EEG based on two novel mental tasks**. *Technology and Health Care* 2015, **23**:S249-S262.
19. Li Y, Olson EB: **A General Purpose Feature Extractor for Light Detection and Ranging Data**. *Sensors* 2010, **10**(11):10356-10375.
20. Li Y, Olson EB, Ieee: **Structure Tensors for General Purpose LIDAR Feature Extraction**. In: *IEEE International Conference on Robotics and Automation (ICRA): 2011 May 09-13 2011; Shanghai, PEOPLES R CHINA*. 2011: 1869-1874.
21. Ojansivu V, Heikkila J: **Blur insensitive texture classification using local phase quantization**. *Image and Signal Processing* 2008, **5099**:236-243.
22. Gonen M: **Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization**. *Bioinformatics* 2012, **28**(18):2304-2310.
23. Chen H, Zhang Z: **A Semi-Supervised Method for Drug-Target Interaction Prediction with Consistency in Networks**. *Plos One* 2013, **8**(5).
24. Öztürk H, Ozkirimli E, Özgür A: **A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction**. *BMC Bioinformatics* 2016, **17**(1):1-11.
25. Wang L, You ZH, Chen X, Yan X, Liu G, Zhang W: **RFDT: A Rotation Forest-based Predictor for Predicting Drug-Target Interactions Using Drug Structure and Protein Sequence Information**. *Current Protein & Peptide Science* 2018, **19**(5):445-454.
26. Gunther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Urdiales EG, Gewiess A, Jensen LJ *et al*: **SuperTarget and Matador: resources for exploring drug-target relationships**. *Nucleic Acids Research* 2008, **36**:D919-D922.
27. 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.
28. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D: **BRENDA, the enzyme database: updates and major new developments**. *Nucleic Acids Research* 2004, **32**:D431-D433.
29. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M: **DrugBank: a knowledgebase for drugs, drug actions and drug targets**. *Nucleic Acids Research* 2008, **36**:D901-D906.
30. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG**. *Nucleic Acids Research* 2006, **34**:D354-D357.
31. Li X, Zhu M, Brasier AR, Kudlicki AS: **Inferring Genome-Wide Functional Modulatory Network: A Case Study on NF-kappa B/RelA Transcription Factor**. *Journal of Computational Biology* 2015, **22**(4):300-312.
32. Li XL, Zhao YX, Tian B, Jamaluddin M, Mitra A, Yang J, Rowicka M, Brasier AR, Kudlicki A: **Modulation of Gene Expression Regulated by the Transcription Factor NF-kappa B/RelA**. *Journal of*

Biological Chemistry 2014, **289**(17):11927-11944.

33. Yang J, Zhao Y, Kalita M, Li X, Jamaluddin M, Tian B, Edeh CB, Wiktorowicz JE, Kudlicki A, Brasier AR: **Systematic Determination of Human Cyclin Dependent Kinase (CDK)-9 Interactome Identifies Novel Functions in RNA Splicing Mediated by the DEAD Box (DDX)-5/17 RNA Helicases**. *Molecular & Cellular Proteomics* 2015, **14**(10):2701-2721.
34. Gribskov M, McLachlan AD, Eisenberg D: **Profile analysis: detection of distantly related proteins**. *Proceedings of the National Academy of Sciences of the United States of America* 1987, **84**(13):4355-4358.
35. Chen X-W, Jeong JC: **Sequence-based prediction of protein interaction sites with an integrative method**. *Bioinformatics* 2009, **25**(5):585-591.
36. Wang L, You ZH, Chen X, Xia SX, Liu F, Yan X, Zhou Y, Song KJ: **A Computational-Based Method for Predicting Drug-Target Interactions by Using Stacked Autoencoder Deep Neural Network**. *Journal Of Computational Biology* 2018, **25**(3):361-373.
37. Jones DT: **Protein secondary structure prediction based on position-specific scoring matrices**. *Journal of molecular biology* 1999, **292**(2):195-202.
38. Jones DT, Ward JJ: **Prediction of disordered regions in proteins from position specific score matrices**. *Proteins-Structure Function and Bioinformatics* 2003, **53**:573-578.
39. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic acids research* 1997, **25**(17):3389-3402.
40. Wang L, You Z-H, Xia S-X, Liu F, Chen X, Yan X, Zhou Y: **Advancing the prediction accuracy of protein-protein interactions by utilizing evolutionary information from position-specific scoring matrix and ensemble classifier**. *Journal Of Theoretical Biology* 2017, **418**:105-110.
41. Chou KC: **Prediction of protein cellular attributes using pseudo-amino acid composition**. *Proteins-Structure Function and Genetics* 2001, **43**(3):246-255.
42. Rodriguez JJ, Kuncheva LI: **Rotation forest: A new classifier ensemble method**. *Ieee Transactions on Pattern Analysis and Machine Intelligence* 2006, **28**(10):1619-1630.
43. Wang L, You Z-H, Yan X, Xia S-X, Liu F, Li L-P, Zhang W, Zhou Y: **Using Two-dimensional Principal Component Analysis and Rotation Forest for Prediction of Protein-Protein Interactions**. *Scientific reports* 2018, **8**(1):12874.

Additional File

Additional file 1: Detailed five-fold cross-validation results on *Enzyme*, *Icon Channel, GPCR* and *Nuclear Receptor* benchmark data sets.

Figures

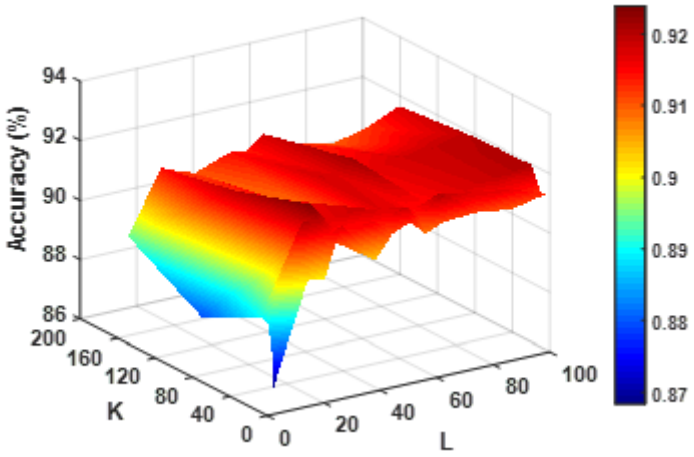

Figure 1

The effect of different rotation forest parameters K and L on classification accuracy

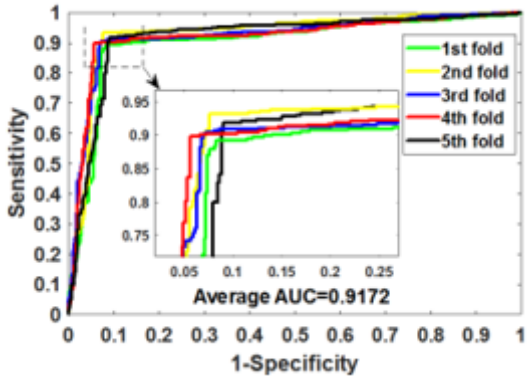

Figure 2

ROC curves obtained by the proposed model on Enzyme data set.

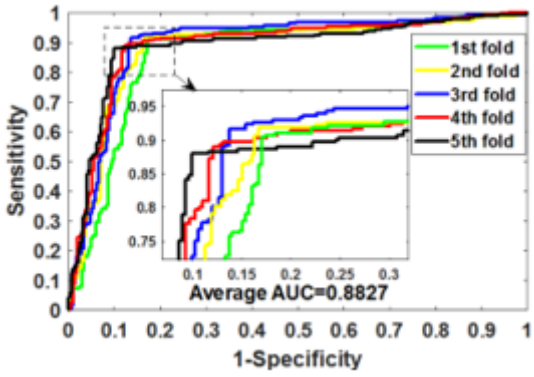

Figure 3

ROC curves obtained by the proposed model on Icon Channel data set.

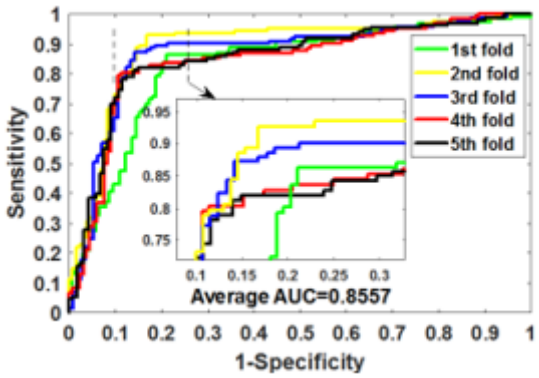

Figure 4

ROC curves obtained by the proposed model on GPCR data set.

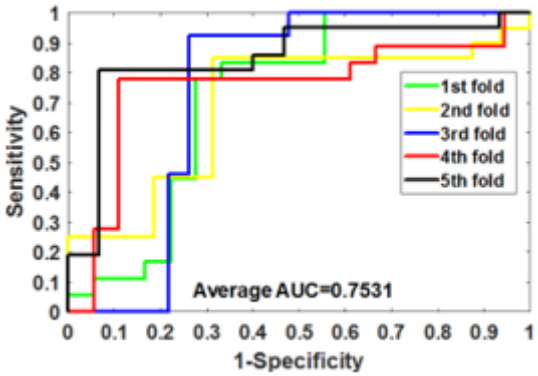

Figure 5

ROC curves obtained by the proposed model on Nuclear Receptor data set.

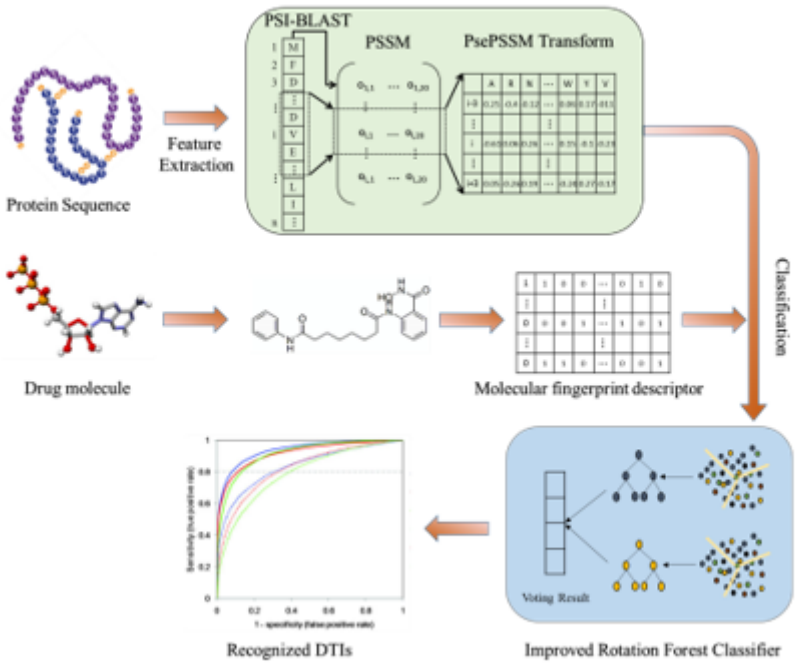

Figure 6

The flow chart of the proposed model

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Formulas1012.jpg
- Formulas13and14.jpg
- Formula7and8.jpg
- Formula9.jpg
- SupplementaryMaterials.docx
- Formulas14.jpg
- Formula5.jpg
- Formula6.jpg