IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Prediction of Drug-Target Interactions by Ensemble Learning Method from Protein Sequence and Drug Fingerprint

Xinke Zhan[1], Zhu-Hong You[1], Jinfan Cai[1], Liping Li[1],Changqing Yu[1],Jie Pan[1],Jiangkun Kong[1]

[1]College of Information Engineering, Xijing University, Xi'an 710123, China

Corresponding author: Zhu-Hong You (e-mail: zhuhongyou@gmail.com).

This research is funded by the National Natural Science Foundation of China, grant number 61722212 and 61873212.

ABSTRACT Predicting the target-drug interactions (DITs) is of great important for screening new drug candidate and understanding biological processes. However, identifying the drug-target interactions through traditional experiments is still costly, laborious and complicated. Thus, there is a great need for developing reliable computational methods to effectively predict DTIs. In this study, we report a novel computational method combining local optimal oriented pattern (LOOP), Position Specific Scoring Matrix (PSSM) and Rotation Forest (RF) for predicting DTI. Specifically, the target protein sequence is firstly transformed as the PSSM, in which the evolutionary information of protein is retained. Then, the LOOP is used to extract the feature vectors from PSSM, and the sub-structure information of drug molecule is represented as fingerprint features. Finally, RF classifier is adopted to infer the potential drug-target interactions. When the experiment is carried out on four benchmark datasets including *enzyme*, *ion channel*, *G protein-coupled receptors (GPCRs)*, and *nuclear receptor*, we achieved the high average prediction accuracies of 89.09%, 87.53%, 82.05%, and 73.33% respectively. For further evaluating the proposed method, we compare the prediction performance of the proposed method with the state-of-the-art support vector machine (SVM) and K-Nearest Neighbor (KNN). The comprehensive experimental results illustrate that the proposed method is reliable and efficiency for predicting DTIs. It is anticipated that the proposed method can become a useful tool for predicting a large-scale potential DTIs.

INDEX TERMS Drug-target interaction, Local optimal oriented pattern, Position specific scoring matrix, Rotation forest.

I. INTRODUCTION

The prediction of interactions between drugs and target proteins is a critical part of drug discovery pipeline as it can help find a novel drug candidate [1, 2] and understand side effects. With the rapid development of human genome project and molecular medicine, it provides favorable conditions for identifying drug-target interactions (DITs). The Food and Drug Administration (FDA) permitted a limited number of drugs to reach the market, because of a number of drug candidates are rejected due to its adverse side effects and inefficiency [3-5]. Moreover, developing a new drug and approving procedure cost more than 1.8 billion dollars and almost nearly 10 years [6]. Detecting potential drug-target interactions is always an important area and a hot topic of research, which can result in finding new protein targeted drug. In the past decades, much effort has been devoted to identifying drug-target interactions through many biological experiments. However, some traditional experiments are still high-cost, time-consuming, and a high false rate is also inevitable drawbacks at the same time. Thus, in order to reduce the cost and time of the experiment, it is increasingly important and necessary to develop novel computational methods which are stable and reliable for verifying drug-target interactions.

With the great explosion number of publicly-available data in biology and chemistry, several different types of related databases, such as Therapeutic Target Database (TTD) [7, 8], SuperTarget and Matador [9], Kyoto Encyclopedia of Genes and Genomes (KEGG) [10], and DrugBank [11, 12], have been established. These public databases store a number of known drug-target interactions which validate through experimental. This also provides a good basis for researchers to develop novel computational methods to predict DTIs. In recently years, traditional computational methods are mainly contain three parts, namely, docking-based [13, 14], ligand-based methods [15], and chemogenomic approaches [16]. The docking-based method which is a useful molecular modeling method can

predict the interaction between the compound and the target protein accurately. However, the three-dimensional (3D) structure of some target protein is complicated to obtain, such as *ion channels* and *GPCRs* [17, 18]. More importantly, the 3D structures of the proteins are difficult to obtain which only can perform by the methods such as NMR and x-ray crystallography. Therefore, the docking-based method is subject to certain restrictions. The ligand-based methods predicting drug-target interactions are most based on a QSAR framework [19] which basic assumption is that ligand with chemical similarity also have similar biochemical activities [20]. At present, most QSAR models are built for a specific target, so that it can only predict the molecular interaction of a target. The performance of the constructed QSAR model would not be excellent if the number of known active molecules for a specific target is insufficient. Generally speaking, it is highly imperative to develop efficient and robust computational methods for predicting the identification of drug-target interactions [21, 22].

Until now, there have been proposed many computational methods based on machine learning in order to solve the limitations of traditional computational method. For example, Nidhi et al. [23] proposed a multiple-category Laplacian-modified naïve Bayesian model to train 964 target categories in the (World of Molecular BioAcTivity) WOMBAT database and predicted the top three most potential compound targets in the MDDR database. Liu et al. [24] proposed a novel prediction algorithm, namely neighborhood regularized logistic matrix factorization (NRLMF), which focus on predicting the probability whether a drug would interact with a target and also study local structure of drug-target pairs for further improving the accuracy of DTIs. Wang et al. [25] developed a computational method which combines auto covariance (AC) and rotation forest for predicting potential drug-target interactions. Mei et al. [26] developed a novel approach namely BLM-NII, which integrated neighbor-based interaction profile inferring (NII) into bipartite local model (BLM), the method achieved excellent improvement in inferring unknown drug-target interaction. Huang et al. [27] proposed a computational model using extremely randomized trees for predicting drug-target interactions, the improvement of this work mainly come from the protein sequence is converted into pseudo substitution matrix representation (Pseudo-SMR) descriptor that can retain evolutionary information. Chen et al. [28] proposed NetCBP, a semi-supervised learning based model for identifying DTIs by using labeled and unlabeled interaction information. You et al. [29] designed a new computational method, named DTIRF, which fully utilized drug molecular structure and protein sequence and employ feature weighted rotation forest (FwRF) for predicting DITs. Besides, the development of drug-target interactions is also conducive to the development of drug-drug interactions (DDIs) [30-34]. For instance, Zhang et al. [35] reported a label propagation method with linear neighborhood information called LPLNI, the method considered drug-drug linear neighborhood similarity as the manifold of drugs, and combing the known drug-target interactions and the drug-drug linear neighborhood similarity to predict unobserved DTIs. Sridhar et al. [36] report a probabilistic approach which fully utilizes multiple drug-based similarities and known interactions, this method obtained excellent performance and find five novel interactions validated by external sources.

In this article, we propose a novel computational method based on target protein sequence and drug substructure fingerprints. The method combines local optimal oriented pattern (LOOP), position specific scoring matrix (PSSM) and rotation forest (RF) for predicting DTIs. Specially, we first transform the target protein sequence into PSSM in order to retain biological evolutionary information, and consider molecular substructure fingerprints are considered as the feature of drugs. We then applied local optimal oriented pattern (LOOP) to extract the 256 dimension feature vectors from PSSM. Finally, we utilize rotation forest to predict the DTIs. The proposed method would evaluate the performance of proposed method by using five-fold cross validation on four benchmark datasets: *enzyme*, *ion channel*, *GPCRs* and *nuclear receptor*. We also compared the proposed method with the state-of-the-art support vector machine (SVM) and KNN on four benchmark datasets. The comprehensive experimental results demonstrate that the proposed method is feasible and effectively for identifying drug-target interactions on a large scale.

II. Materials and Methodology

A. Golden Standard Datasets

In this study, we execute the experiment for predicting DTIs on four golden standard datasets, namely *enzyme*, *ion channels*, *GPCRs*, and *nuclear receptor*, respectively. These datasets are collected from DrugBank [11], KEGG BRITE [10], SuperTarget & Matador [9], and BRENDA [37] which were considered as high-reliability databases and were widely used in various experimental methods. Table 1 summarized the statistical information of drug target interaction. The number of drugs known to target *enzyme*, *GPCRs*, *ion channels*, *nuclear receptor* are 445, 233, 210, and 54, respectively. The number of target proteins of these benchmark datasets is 664, 95, 204, and 26, respectively. And the number of each dataset of known drug-target interacting is 2926, 635, 1476, and 90, respectively. In total, 5127 drug-target interaction pairs are collected after screening of these drugs and targets which were chosen as positive sample sets in the experiments.

TABLE 1. The statistical information of four drug-target data.

Dataset	Drugs	Target Proteins	Interactions
Enzyme	445	664	2926
GPCRs	223	95	635
Ion Channels	210	204	1476
Nuclear Receptor	54	26	90

The DTIs network can be denoted as a bipartite graph in which nodes represent targets or drugs, and edges represent the interaction between nodes. Specially, if the relationship is existed between the nodes, connect the nodes with edges, otherwise Otherwise there is no correlation between the nodes. It is worth noting that the edges in the initial bipartite graph represent the real drug-target interactions have been validated by biological experiments, however, the known initial edges in whole connected bipartite graph only account for few portion. Taking the *ion channel* dataset as an example, the ion channel dataset contains 204 target proteins, 210 drugs, and experimentally verified 1476 pairs of DTIs. These corresponding connections have up to 42,840 (210×204) edges. However, 1476 initial connections which is less than the number of possible negative samples ($42840-1476 = 41{,}364$) would cause a bias problem due to unbalance samples. For correcting this problem, we randomly select the number of the negative samples which is same as the number of the positive samples. It is obviously that the number of real interaction negative samples we selected is very small on a large bipartite graph. Finally, the negative samples of *enzyme*, *ion channels*, *GPCRs*, and *nuclear receptor* datasets were 2926, 1476, 635, and 90, respectively.

B. Molecular Substructure Fingerprint of Drug

There have been proposed different kinds of drug compounds descriptors, such as constitutional, quantum chemical properties, topological and geometrical. Recently, some researches [38-40] illustrate that using a variety of molecular substructure fingerprints to represent drug compounds is effective. It records the existence of substructure after separating the drug molecular into fragments. Moreover, the structure properties of the drug molecules were encoded in binary bits, which can directly know whether specific substructure fragments in the drug molecules exist or not. It can not only avoids the error transfer and accumulation in the process of molecular descriptor calculation but also reduce the workload of molecular descriptor calculation and screening. Given a specific drug molecular, the vector is set to be 1 if the substructure is presence or the vector is set to be 0. By doing this, the complex structure of drug molecules can be described as molecular substructure fingerprint. In this work, the molecular substructure fingerprints are represented as the drug molecules features, which can be collected from the PubChem System and its website is (https://pubchem.ncbi.nlm.nih.gov/). There store 881 substructures information in drug fingerprint. As a result, the molecular feature of the drug is 881 binary vectors.

C. Position-Specific Scoring Matrix

Up to now, there have been existed many feature descriptors for protein sequence, due to effective descriptors can boost the performance of identifying DTIs. Position-specific scoring matrix (PSSM) [41] is one of the descriptor which carries the evolutionary information [42] of sequence and gives probability scores of any given amino acid for a specific position, and widely used in previously work such as protein binding site prediction, protein secondary structural prediction and protein subcellular localization. In order to convert the target protein sequence, the Position-Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST) [43] which can search and compare the homologous sequence of each target protein sequence is adopted to create PSSM of each target protein sequence. Here, the PSSM can be expressed as follows:

$$W = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,j} & \cdots & p_{1,20} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,j} & \cdots & p_{2,20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{i,1} & p_{i,2} & \cdots & p_{i,j} & \cdots & p_{i,20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{L,1} & p_{L,2} & \cdots & P_{L,\mathrm{j}} & \cdots & P_{L,20} \end{bmatrix} \quad (1)$$

where W is a matrix which construct is $L \times 20$; L represents the length of target protein sequence and 20 is the number of amino acid, and denotes the mutation score which represents the probability of amino acid i residue change into amino acid j in the process of biological evolution. In this experiment, we employed PSI-BLAST tool to transform each protein sequence into a PSSM. The parameter of *e-value* is set to 0.001 and maximum number of iterations is 3, other parameters were set to default values. Here, the meaning of *e-value* is describing the random background noise that exists for matches between target protein sequence and iteration values allows the PSI-BLAST to find and tune to the specific properties of the query and its homologs until no new sequences are detected. The example of convert target protein sequence into PSSM is displayed in Figure 1.

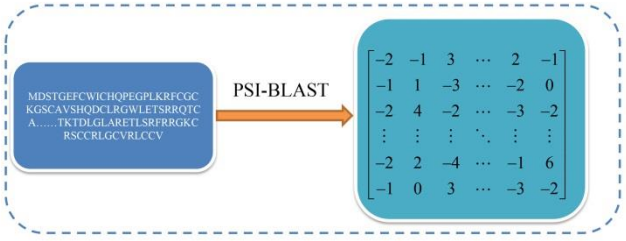


FIGURE 1. Given an example of convert target protein sequence into PSSM by using PSI-BLAST tool.

D. Local Optimal Oriented Pattern

Local optimal oriented pattern (LOOP) was proposed by Charkraborti et al. [44, 45]. It is texture descriptors which encode repeated local patterns in images as binary codes, and it is a popular type of feature used for classification in computer vision. Because of the disadvantage of local binary pattern (LBP) [46] and local derivative pattern (LDP) [47] is the arbitrary sequence of binarization weights that adds

dependency to orientation. Thus, LOOP presents a nonlinear amalgamation of LBP and LDP that overcomes these drawbacks while preserving these strengths. It integrates the strength of two texture descriptors LDP and LBP for assigning weights and finding the intensity differences. In LOOP algorithm, The LOOP feature is obtained by calculating for each image pixel using a 3×3 neighborhood around each pixel.

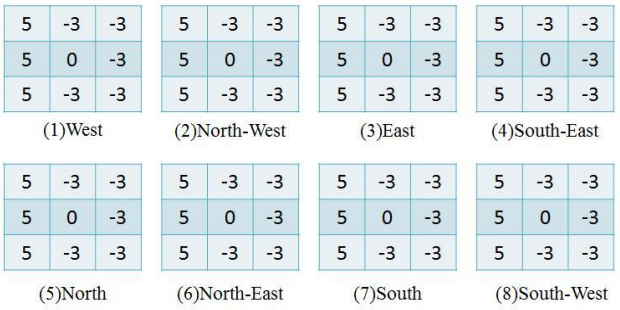


FIGURE 2. Kirsch masks for each direction.

Firstly, it computed the eight responses of the Kirsch masks which can be seen from Figure 2, $(l_n, n = 0,1,\cdots,7)$ corresponding to pixels with intensities $(i_n, n = 0,1,\cdots,7)$ to obtain the intensity variation in the eight directions. Secondly, based on the rank of the magnitude value l_n, each pixel obtained different weight n. Finally, computing the LOOP code for the center pixel (i, j) as follows:

$$LOOP(i,j) = \sum_{n=0}^{7} s(i_n - i_c) \times 2^n \tag{2}$$

$$s(x) = \begin{cases} 1 & if \quad x \geq 0 \\ 0 & otherwise \end{cases} \tag{3}$$

where i_c is the intensity of the center pixel. Here, the input signal PSSM is a N×20. In this work, the each target protein sequence would be represented by 256 feature vectors after using LOOP feature descriptor.

E. Rotation Forest (RF) Classifier

Rotation forest (RF) first proposed by Rodriguez et al. [48] is widely used for classification. The RF algorithm focuses on improving the difference and accuracy of the base classifier. In this work, we adopt RF as a classification model for predicting DTIs. Specifically, the RF randomly divides entire sample set into K subsets, and principal component analysis (PCA) method is adopted to transform the subsets which make the difference between each subset. Finally, the prediction score is obtained after training different base classifiers. Let Q be the training sample set which size is $N \times n$, where N denotes the number of samples. Let R be the feature set, and the corresponding label be the $Y = [\mathrm{y}_1, \mathrm{y}_2, \cdots, \mathrm{y}_\mathrm{n}]^T$. The feature set is randomly divided into K equal subsets. Suppose the number of decision trees is T, which can be denoted as $H_1, H_2, \cdots H_L$, respectively. The rotation forest classifier can be described as follows:

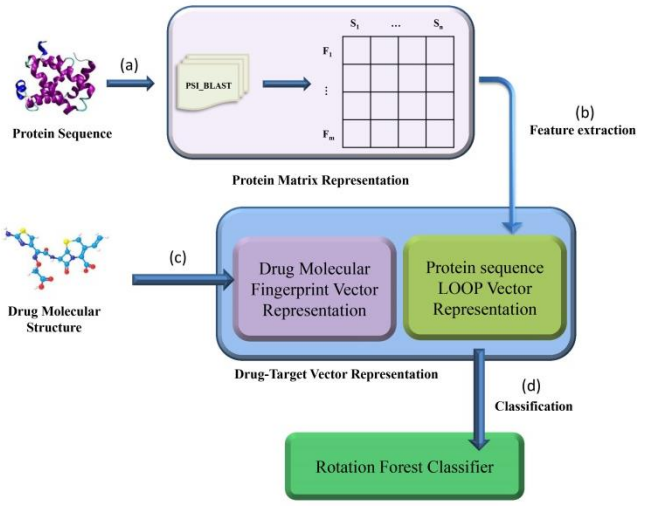


FIGURE 3. The workflow of the proposed approach for predicting drug-target interaction. (a) transforming the target protein sequence into the PSSM. (b) extract LOOP descriptors from PSSM. (c) fingerprint information representation. (d) adopting rotation forest to classify the potential drug-target interactions.

(1) Select the suitable parameter K, The feature set R is randomly divided into K subsets, each subset contains n/K features.

(2) Let R_{ij} denote the jth sub-feature set of the training set, which used to train the ith classifier T_i. For each subset, a new training set Q'_{ij} is generated after a bootstrap resampling with 75 percent of training set Q.

(3) Apply principal component analysis (PCA) on Q'_{ij} to produce the coefficients in matrix F_{ij}, which is a matrix of $M \times 1$. F_{ij} can be represented as $\alpha_{ij}^{(1)}, \cdots, \alpha_{ij}^{(\mathrm{M}_j)}$.

(4) The coefficients obtained in the matrix F_{ij} are constructed a sparse rotation matrix P_i, which is shown as follows:

$$P_i = \begin{bmatrix} \alpha_{i1}^{(1)}, \cdots, \alpha_{i1}^{(M_1)} & 0 & \cdots & 0 \\ 0 & \alpha_{i2}^{(1)}, \cdots, \alpha_{i2}^{(M_2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \alpha_{iK}^{(1)}, \cdots, \alpha_{iK}^{(M_K)} \end{bmatrix} \tag{4}$$

Given a sample w, let $d_{ij}(wP_i^{\alpha})$ be the probability which is predicted whether w belongs to y_i by the classifier H_i. Then, calculate the confidence of the class by means of the average combination, and the formula is shown below:

$$\eta_j(\mathrm{w}) = \frac{1}{T} \sum_{i=1}^{L} d_{ij}(wP_i^{\alpha}) \tag{5}$$

The test sample w will be assigned the category with the greatest possible. The workflow of the proposed method for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3026479, IEEE Access

predicting potential drug-target interactions is shown in Figure 3.

III. Experimental and Results

A. Evaluation Criteria

In this work, in order to evaluate the performance of the propose method, we use the evaluation measures such as the overall prediction accuracy (Accu.), sensitivity (Sens.), precision (Prec.), and Matthews correlation coefficient (MCC). They are defined as follows:

$$Acc. = \frac{TN + TP}{TN + TP + FN + FP} \tag{6}$$

$$Prec. = \frac{TP}{TP + FP} \tag{7}$$

$$Sen. = \frac{TP}{TP + FN} \tag{8}$$

$$Spec. = \frac{TN}{TN + FP} \tag{9}$$

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TN + FN) \times (TP + FP) \times (TN + FP) \times (FN + TP)}} \tag{10}$$

where false positive (FP) is the number of drug-target pairs which are predicted as interacting pairs incorrectly; true negative (TN) denotes the count of samples which are classified as non-interacting correctly; true positive (TP) represents the number of samples which are predicted as interacting pairs correctly; and false negative (FN) is the number of true samples which are predicted as non-interacting pairs incorrectly. Moreover, the receiver operating characteristic (ROC) curves were also computed to evaluate the performance of the proposed method. To Summarize the ROC curve in a numerical way, the area under an ROC curve (AUC) was also computed for better analyze the propose method.

B. Choosing the parameters

Before executing a series of experiment, the parameters need to be optimized in our proposed model. Especially in rotation forest algorithm, optimizing two corresponding parameters of K and L for capturing best performance of parameters is necessary in the prediction model. Where K is the number of feature subsets and L represents the number of decision tree. We employ the grid research method to select the optimized parameters K and L. Under different parameters, the accuracy of RF generation is shown in Figure 4. We can see from Figure 4, the optimal parameters K=26 and L=25 have better performance than other parameter. As a result, we set K=26 and L=25 in this paper.

C. Performance of the proposed method

For the fairness of this study, when predicting the DTIs datasets of *enzyme*, *ion channels*, *GPCRs* and *nuclear receptor*, five-fold cross-validation would be adopted in this work in order to avoid the over-fitting of the prediction model. Specifically, the entire dataset was evenly divided into five parts which four parts used for training and one part for testing. By doing this, we constructed five training model,

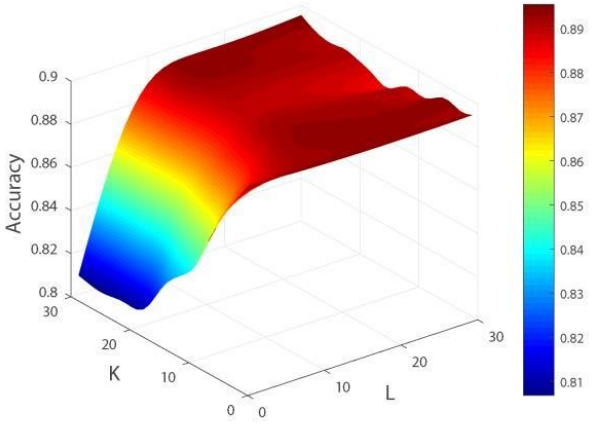


FIGURE 4. Accuracy surface obtained for optimizing *K* and *L*.

and obtained prediction score for infer the drug-target pair whether interact or not. The prediction results of *enzyme*, *ion channel*, *GPCRs*, and *nuclear receptor* datasets are shown in Table 2-5.

TABLE 2. five-fold cross-validation results performed on *enzyme* dataset by using the proposed method.

Testing set	Acc.(%)	Prec.(%)	Sen.(%)	Spec.(%)	MCC(%)
1	88.72	89.96	86.85	90.54	79.96
2	88.12	89.81	86.32	89.97	79.05
3	89.32	90.38	88.07	90.57	80.91
4	90.34	90.72	89.62	91.05	82.54
5	88.97	90.80	86.95	91.03	80.37
Average	**89.09 ±0.82**	**90.33 ±0.44**	**87.56 ±1.32**	**90.63 ±0.44**	**80.57 ±1.30**

TABLE 3. five-fold cross-validation results performed on *ion channel* dataset by using the proposed method.

Testing set	Acc.(%)	Prec.(%)	Sen.(%)	Spec.(%)	MCC(%)
1	85.25	87.19	82.77	87.76	74.83
2	90.34	89.97	91.99	88.49	82.46
3	89.15	88.03	89.29	89.03	80.62
4	86.61	86.14	87.58	85.62	76.80
5	86.27	85.08	87.15	85.43	76.31
Average	**87.53 ±2.13**	**87.28 ±1.87**	**87.76 ±3.37**	**87.26 ±1.65**	**78.20 ±3.20**

TABLE 4. five-fold cross-validation results performed on *GPCRs* dataset by using the proposed method.

Testing set	Acc.(%)	Prec.(%)	Sen.(%)	Spec.(%)	MCC(%)
1	81.89	81.62	84.09	79.51	70.24
2	82.68	87.40	79.86	86.09	71.23
3	82.68	80.95	83.61	81.82	71.33
4	80.31	76.34	84.06	77.04	68.30
5	82.68	81.10	83.74	81.68	71.34
Average	**82.05 ±1.03**	**81.48 ±3.94**	**83.07 ±1.81**	**81.23 ±3.34**	**70.49 ±1.31**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3026479, IEEE Access

TABLE 5. five-fold cross-validation results performed on *nuclear receptor* dataset by using the proposed method.

Testing set	Acc.(%)	Prec.(%)	Sen.(%)	Spec.(%)	MCC(%)
1	72.22	68.42	76.47	68.42	59.75
2	66.67	68.18	75.00	56.25	54.15
3	80.56	77.78	82.35	78.95	68.62
4	63.89	75.00	57.14	73.33	52.92
5	83.33	90.91	66.67	95.24	69.65
Average	**73.33 ±8.47**	**76.06 ±9.29**	**71.53 ±9.80**	**74.44 ±14.33**	**61.02 ±7.86**

It can be observed from Table 2 that when predicting *enzyme* dataset by using the proposed method. It yielded the good result with high average accuracy, precision, sensitivity, specificity, and MCC of 89.09%, 90.33%, 87.56%, 90.63%, and 80.57%, respectively. The standard deviations of the results were 0.82%, 0.44%, 1.32%, 0.44% and 1.30%, respectively. When using the proposed method to predict *ion channel* dataset, we obtained the result of average accuracy, precision, sensitivity, specificity, and MCC of 87.53%, 87.28%, 87.76%, 87.26% and 78.20%, respectively. Their standard deviations were 2.13%, 1.87%, 3.37%, 1.65% and 3.20%, respectively. When applying our method on *GPCRs* dataset, the average of accuracy, precision, sensitivity, specificity, and MCC of 82.05%, 81.48%, 83.07%, 81.23% and 70.49%, respectively, and corresponding standard deviation were 1.03%, 3.94%, 1.81%, 3.34% and 1.31%, respectively. When performing DTI prediction on the *nuclear receptor* dataset, the average accuracy, precision, sensitivity, specificity, and MCC come to be 73.33%, 76.06%, 71.53%, 74.44%, and 61.02% with corresponding standard deviations of 8.47%, 9.29%, 9.80%, 14.33%, and 7.86%, respectively. It is noteworthy that the results of *unclear receptor* dataset have higher standard deviations, which was mainly caused by the number of whole data samples. Meanwhile, the average AUC of *enzyme*, *ion channel*, *GPCRs*, and *nuclear receptor* datasets were 0.9532, 0.9349, 0.8882, and 0.8199, respectively. The ROC curves performed are shown in Figures 5-8.

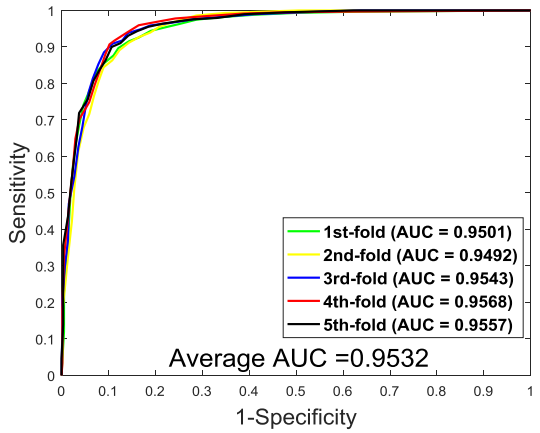


FIGURE 5. The ROC curves of the proposed method on *enzyme* dataset.

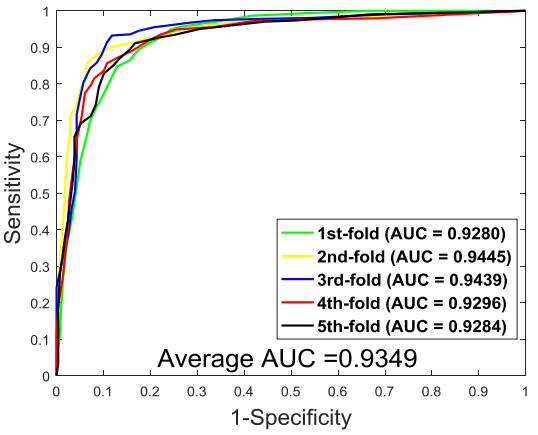


FIGURE 6. The ROC curves of the proposed method on *ion channel* dataset.

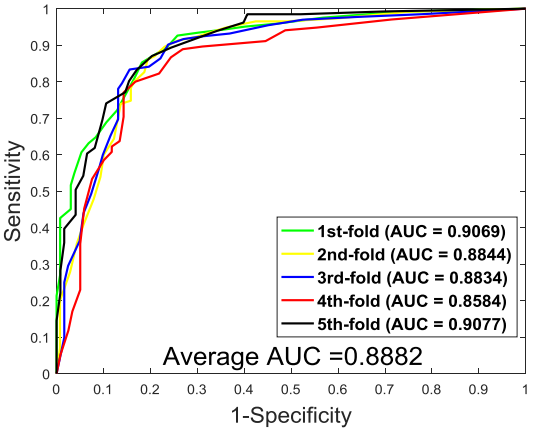


FIGURE 7. The ROC curves of the proposed method on *GPCRs* dataset.

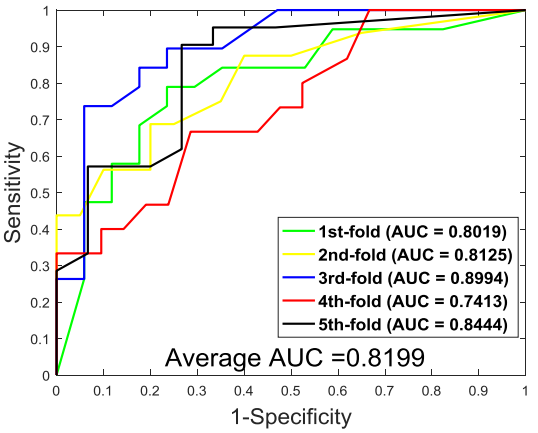


FIGURE 8. The ROC curves of the proposed method on nuclear receptor dataset.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3026479, IEEE Access

D. Comparison between LPQ descriptor model and the proposed model

For evaluating the impact of LOOP describer on the proposed mode effectively, we compare with different extraction method in computational model. In this section, we employ local phase quantization (LPQ), which is first proposed by Ojansivu and heikkila [49], to evaluate the performance in predicting DTIs. The LPQ is an effective operator in texture feature extraction that remain the blur-invariant property. The cross-validation results of LPQ descriptor combined with RF classifier on four benchmark datasets are summarized in Table 6. It can be seen that the results of the proposed method are improved than LPQ model including overall accuracy, precision, sensitivity, MCC, respectively, expect specificity of GPCRs dataset. In order to further disscuss the effectiveness of the propsoed method, we compare the computation complexity between LPQ-based and LOOP-based which summarized in Table 7. The five-fold cross-validation method was still used in this comparison. Briefly, LOOP-based method can be computed from 1.13s to 46.67s faster than LPQ-based method on *enzyme*, *ion channel*, and *nuclear receptor*, expect the dataset of *GPCRs*, the LOOP-based method can be computed from 7.31s slower than LPQ-based method.The ROC curves were also computed that were summarized in Figure 9. Each AUC value of the propsoed method on benchmark dataset is a bit higher than LPQ-based model.As a result, our proposed computational method is efficient to predict poteintial drug-target interactions.

TABLE 6. Experimental results comparison on LPQ and LOOP with the same RF classifier of four benchmark dataset.

Dataset	Method	Acc .(%)	Prec. (%)	Sen. (%)	Spec. (%)	MCC (%)
Enzyme	LPQ+RF	88.58 ±0.81	89.95 ±1.28	86.88 ±1.19	90.30 ±1.16	79.76 ±1.26
	LOOP+RF	**89.09 ±0.82**	**90.33 ±0.44**	**87.56 ±1.32**	**90.63 ±0.44**	**80.57 ±1.30**
Ion Channel	LPQ+RF	85.66 ±0.90	85.76 ±1.10	85.52 ±0.98	85.81 ±0.90	75.41 ±1.27
	LOOP+RF	**87.53 ±2.13**	**87.28 ±1.87**	**87.76 ±3.37**	**87.26 ±1.65**	**78.20 ±3.20**
GPCRs	LPQ+RF	80.71 ±2.38	81.27 ±2.88	80.08 ±5.45	**81.37 ±4.13**	68.80 ±2.79
	LOOP+RF	**82.05 ±1.03**	**81.48 ±3.94**	**83.07 ±1.81**	81.23 ±3.34	**70.49 ±1.31**
Nuclear Receptor	LPQ+RF	71.11 ±7.51	68.79 ±17.29	78.44 ±14.43	66.65 ±13.30	57.76 ±7.19
	LOOP+RF	**73.33 ±8.47**	**76.06 ±9.29**	**71.53 ±9.80**	**74.44 ±14.33**	**61.02 ±7.86**

TABLE 7. Computational complexity comparison for RF classifier based on feature LPQ and LOOP.

Dataset	LPQ-Average computation time(s)	LOOP-Average computation time(s)
Enzyme	534.72	**516.71**
Ion Channel	275.99	**229.32**
GPCRs	**93.59**	100.90
Nuclear Receptor	14.44	**13.31**

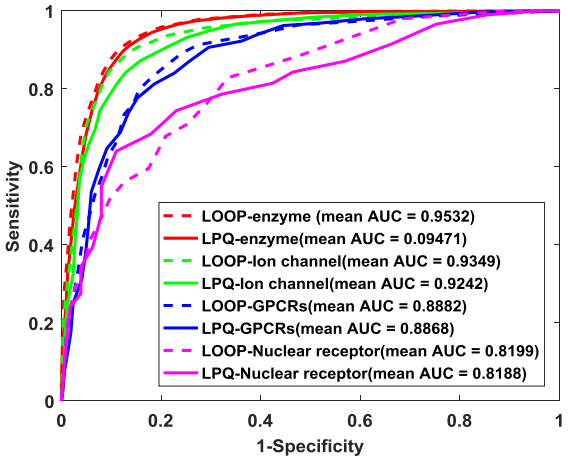


FIGURE 9. Performance comparison on LPQ and LOOP of four benchmark datasets.

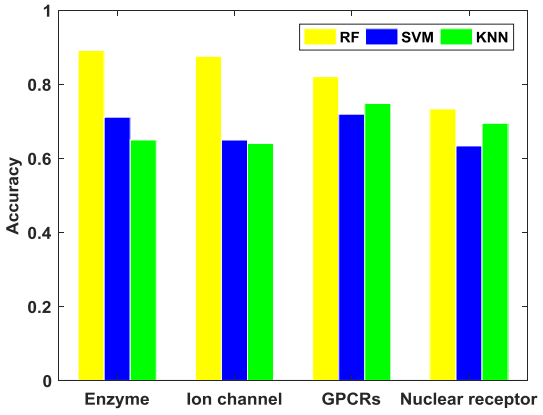
E. Comparison of RF with other models

Many machine learning models have been used to predict DTIs and most of them are based on traditional classifiers. In order to further evaluate the effectiveness of the proposed method for predicting drug-taeget interactions, we compare the proposed method by adopting the same extraction method with the state-of-the-art support vector machine (SVM) and K-nearest neighbor algorithm. For the RF model, we adopt the parameters K=26 and L=25 which has been experimented in this paper. The SVM classifier, as one of widely used machine learning algorithm, have an excellent performance in solving classification problem. K-Nearest Neighbor (KNN) belongs to supervised learning, and KNN is widely used because of its simplicity and effeciency. When training the SVM models, the LIBSVM tool [50] of SVM is selected to prediction DTIs and adopt the Gaussian kernel in SVM. Meanwhile, there have two parameters c and g need to be optimized appropriately. As mentioned above, we still adopt grid research method to choose the best parameters, we choose the best optimized parameters by employing in total of 400 times to optimizing pair of (c, g) which using 20 different values of g and 20 different values of c. The KNN alogorthm need to optimize the best parameter the number of neighbors k and distance measuring function. Here, the k and distance measuring function are selected as 2 and L1.

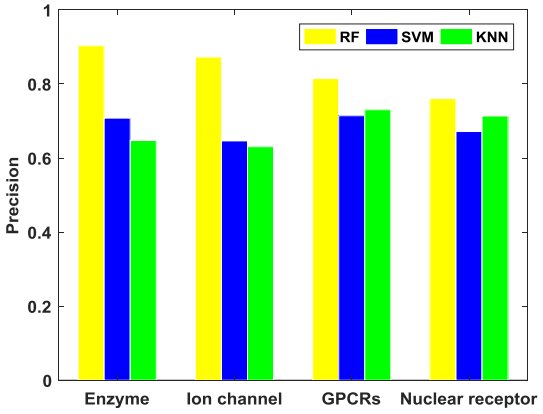
Figure 10 reports the experiment results of RF, SVM, and KNN classifiers in four benchmark datasets of *enzyme*, *ion channel*, *GPCRs* and *nuclear receptor*. In Figure 10 (a)-(f), it can be seen that the results of RF is significantly better than SVM and KNN in terms of overall accuracy, precision, sensitivity, specitivity, MCC, and AUC, respectively. For example, the accuracy gaps between RF and SVM for the foure datasets are 18.01%, 22.61%, 10.16%, and 10.00%, respectively. Similarly, the accuracy gaps between RF and KNN are 24.15%, 23.53%, 7.25%, and 3.89%, respectively. As a result, we can make a conclusion that RF is more accurate than support vector machine and KNN.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3026479, IEEE Access

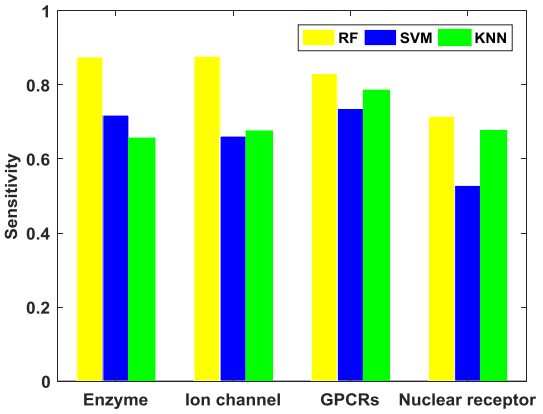
For further comparing their performance, we report the ROC curve in Figure 11 and Figure 12 which plots the false positive rate (1-sepcificity) against the true positive rate (sensitivity). The AUC values represent the performance of each classifier, the higher values of AUC, The stronger performance in predicting potential drug-target interactions. Among them, the AUC values gaps between RF and SVM for the four datasets come to be 0.1661, 0.2214, 0.1080, and 0.1441, respectively. Similarly, the AUC values gaps between RF and KNN for the four datasets come to be 0.3037, 0.2947, 0.1398, and 0.1200, respectively. Hence, we conclude that the RF have much better performance in prdicting DTIs.



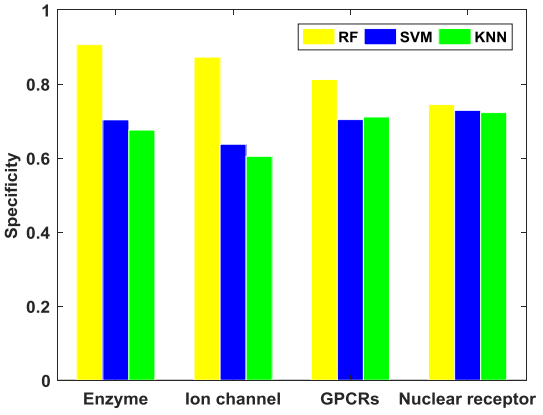
(a)



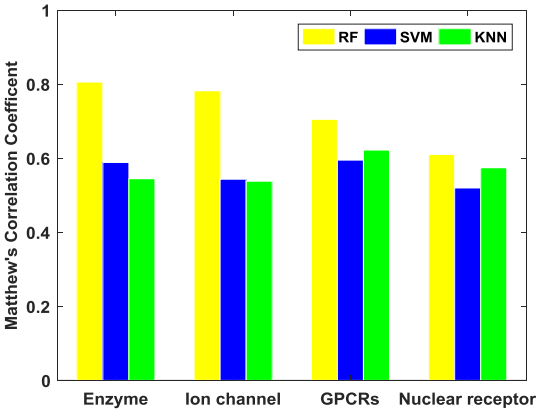
(b)



(c)

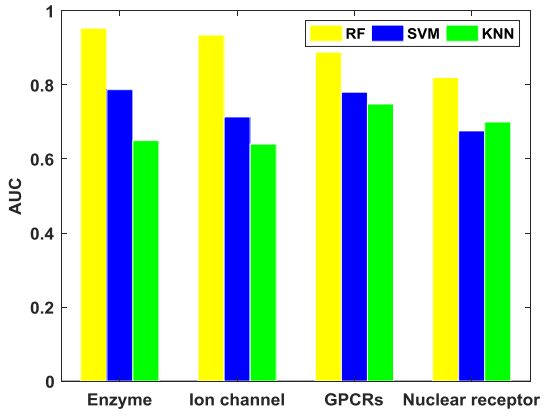


(d)



(e)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3026479, IEEE Access



(f)

FIGURE 10. Performance comparison with six validation metrics on three classifiers: Rotation forest (yellow bar), support vector machine (blue bar) and K-Nearest neighbor (green bar). (a) Accuracy. (b) Precision. (c) Sensitivity. (d) Specificity. (e) MCC. (f) AUC.

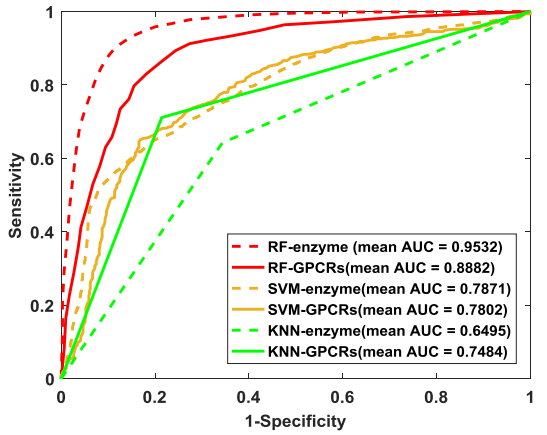


FIGURE 11. ROC curve performed on *enzyme* and *GPCRs* datasets using different classifiers of RF, SVM, and KNN.

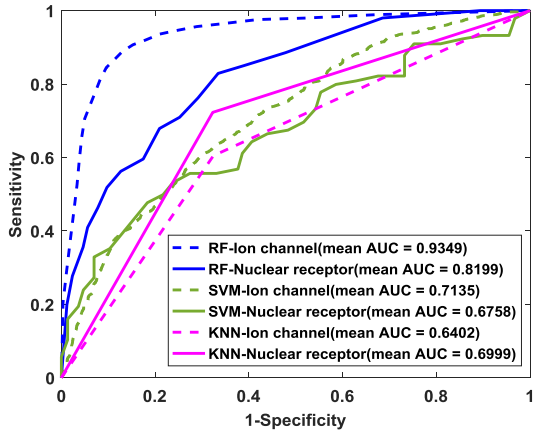


FIGURE 12. ROC curve performed on *Ion channel* and *Nuclear receptor* datasets using different classifiers of RF, SVM, and KNN.

F. Comparison with the previous works

Nowadays, a lot of computational methods have been proposed for predicting protein-protein interactions. In this work, we compared the prediction performance between our proposed method and the existing methods including SIMCOMP [51], KBMF2K [52], MLCLE [53], AM-PSSM [54]. These methods also adopted the same five-fold cross-validation on four benchmark datasets, and the differences of these existing methods were computational framework. Here, the average AUC values of the previous works were listed in Table 8. It can be observed that the method we proposed has a significant improvement in prediction performance. The growth of average AUC on the datasets of *enzyme*, *GPCRs*, and *ion channel* were 0.0902, 0.1359, and 0.0212, respectively. As for *nuclear receptor* dataset, it can be ignored that the result of average AUC is lower than SIMCOMP which the gap is 0.0361. Generally speaking, the comparison results illustrate that the LOOP descriptors combined with rotation forest can effectively improve the prediction performance for drug-target interaction.

TABLE 8. Performance comparison of the proposed model and other excellent models on four benchmark datasets.

Dataset	Our Method	SIMCOMP	KBMF2K	MLCLE	AM -PSSM
Enzyme	**0.9532**	0.863	0.832	0.842	0.843
Ion Channels	**0.9349**	0.776	0.799	0.795	0.722
GPCRs	**0.8882**	0.867	0.857	0.850	0.839
Nuclear Receptor	0.8199	**0.856**	0.824	0.790	0.767

IV. Discussion

In this paper, a novel computational approach is proposed by combining local optimal oriented pattern (LOOP), position-specific scoring matrix (PSSM), and rotation forest (RF) classifier for predicting potential drug-target interactions. Specifically, the target protein sequence is transformed into PSSM, then employing LOOP to extract local feature vectors, and combine drug fingerprint information to form a new drug-target pair which helps improve predicting performance. Our method mainly focuses on fully utilizing evolutionary information and drug fingerprint information to improve the potential drug-target interactions.

We carried out a plenty of experiments on four benchmark datasets, including *enzyme*, *ion channel*, *GPCRs*, and *nuclear receptor*, which were provided by Yamanishi et al. [55]. The results of proposed method we obtained are reliable. Meanwhile, we compared the proposed method with the state-of-the-art SVM, KNN by employing the same feature extraction method. Our proposed method has the smallest improvement in specificity by 1.56% of *nuclear receptor* dataset, and the biggest improvement in MCC by 24.41% of *ion channel* dataset. The results illustrate that the LOOP descriptor can capture the feature vectors and improve the predicting performance effectively. When comparing with the LPQ descriptor, the performance results

of proposed method still shows good prediction ability. Even compared with the previously works in Table 7, our method has been improved in forecasting, even if the range of improvement is not significant.

Although the predictive performance is improving, our method still depends on manual feature extraction, which exist some noise information and could not capture all effective features. These drawbacks lead to certain limitations in predictive ability. From a technical viewpoint, only using local information can hardly guarantee the improvement of prediction performance for predicting DTIs. Combing local feature and global feature have more conducive to the integrity of feature information. Hence, the performance of our method could be further improved by integrating more effective information which could detect more evolutionary information for improving the accuracy of DTIs.

V. Limitations and Future work

Although a plenty of experiments have been carried out, there still have inevitable limitations in this paper. As we know, feature extraction and classification are mainly two steps in predicting DTIs. We would discussion the limitations from two aspects in this study. On one hand, the feature information extracted by LOOP descriptor is limited to local information, and the global information can hardly capture completely that result in incomplete feature information. On the other hand, when the feature vectors are input to the rotation forest, the data transformation would cause the loss of feature information. Hence, we obtained the better result by optimizing the key parameter that can maximize the retention of effective feature information. From another perspective, it also shows that the difference of the feature vector obtained by the proposed feature extraction method needs to be improved.

In future work, the feature extraction and classification are still plays a crucial role in identifying drug-target interactions. We would still focus on researching more effective feature extraction method for extracting feature vectors which contain more evolutionary information and less noise information. Meanwhile, the importance of classifier is still cannot be ignored. Moreover, with the rapidly development of high-throughput data, developing more robust intelligent machine learning methods to improve the prediction accuracy on a large scale would brought greater challenges.

VI. Conclusions

In this paper, we proposed a novel computational approach combines local optimal oriented pattern (LOOP), position-specific scoring matrix (PSSM), and rotation forest (RF) classifier. The method we proposed for predicting DTIs by fusing molecular fingerprint information and protein sequence information. In the experiment, we adopted five-fold cross-validation method for further evaluating the potential drug-target interaction on four benchmark datasets, including *enzyme*, *ion channel*, *GPCRs*, *nuclear receptor*. The proposed method achieved results of average accuracies 89.09%, 87.53%, 82.05%, and 73.33%, respectively. According to comprehensive experimentation, the predictive performance of proposed method was significantly well in predicting DTIs, especially when comparing different classifier, the proposed method shows the robust and stable in predicting DTIs. The main improvement come from the feature extraction of LOOP which integrates the strength of two texture descriptors local binary pattern (LBP) and local derivative pattern (LDP) and find the texture intensity differences in drug-target pairs. However, due to the limitation of the feature extraction method, we still need to make great efforts to improve the effectiveness of feature extraction methods. In future work, we would foucus more on imporving the performance of computational approach and reduce the computational cmplexity. We expect this study could accelerate related biomedical research and it can be a useful tool when predicting DTIs.

REFERENCES

[1] Y.-C. Wang, Z.-X. Yang, Y. Wang, N.-Y Deng, "Computationally probing drug-protein interactions via support vector machine," *Letters in Drug Design & Discovery*, vol. 7, no. 5, pp. 370-378, Jun. 2010.

[2] Z. Xia, L.-Y. Wu, X. Zhou, S.-T. Wong, "Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces," *BMC systems biology*, vol. 4, no. S2, pp. S6, Sep. 2010.

[3] Y. Yamanishi, E. Pauwels, M. Kotera, "Drug side-effect prediction based on the integration of chemical and biological spaces," *Journal of chemical information and modeling*, vol. 52, no.12, pp. 3284-3292, Nov. 2012.

[4] M. Kuhn, M. AI. Banchaabouchi, M. Campillos, L. J. Jensen, C. Gross, A. C. Gavin, P. Bork, "Systematic identification of proteins that elicit drug side effects," *Molecular systems biology*, vol. 9, no.1, pp. 663, Jan. 2013.

[5] A. Ezzat, M. Wu, X.-L, Li, C.-K, Kwoh, "Drug-target interaction prediction via class imbalance-aware ensemble learning," *BMC bioinformatics*, vol. 17, no. 19, pp. 267-276, Dec. 2016.

[6] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, A. L. Schacht, "How to improve R&D productivity: the pharmaceutical industry's grand challenge," *Nature reviews Drug discovery*, vol. 9, no. 3, pp. 203-214, Feb. 2010.

[7] F. Zhu, B. Han, P. Kumar, X. Liu, X. Ma, X. Wei, L. Huang, Y. Guo, L. Han, C. Zheng, "Update of TTD: therapeutic target database," *Nucleic acids research*, vol. 38, no. suppl_1, pp. D787-D791, Jan. 2010.

[8] Y. Wang, S. Zhang, F. Li, Y. Zhou, Y. Zhang, Z. Wang, R. Zhang, J. Zhu, Y. Ren, Y. Tan, "Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics," *Nucleic acids research*, vol. 48, no. D1, pp. D1031-D1041, Jan. 2020.

[9] S. Günther, M. Kuhn, M. Dunkel, M. Campillos, C. Senger, E. Petsalaki, J. Ahmed, E. G. Urdiales, A. Gewiess, L. J. Jensen, "SuperTarget and Matador: resources for exploring drug-target relationships," *Nucleic acids research*, vol. 36, no. supple_1, pp. D919-D922, Jan. 2008.

[10] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Hirakawa, "From genomics to chemical genomics: new developments in KEGG," *Nucleic acids research*, vol. 34, no. supple_1, pp. D354-357, Jan. 2006.

[11] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, M. Hassanali, "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic acids research*, vol.

36, no. suppl_1, pp. D901-D906, Jan. 2008.

[12] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, M. Hassanali, P, Stothard, Z. Chang, J. Woolsey, "DrugBank: a comprehensive resource for *in silico* drug discovery and exploration," *Nucleic acids research*, vol. 34, no. suppl_1, pp. D668-D672, Jan. 2006.

[13] D.–L. Ma, D. S.-H. Chan, C.-H. Leung, "Drug repositioning by structure-based virtual screening," *Chemical Society Reviews*, vol. 42, no. 5, pp. 2130-2141, 2013.

[14] B. Waszkowycz, D. E. Clark, E. Gancia, "Outstanding challenges in protein-ligand docking and structure-based virtual screening," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 1, no. 2, pp. 229-259, Mar. 2011.

[15] Y.-F. Dai, X.-M. Zhao, "A survey on the computational approaches to identify drug targets in the postgenomic era," *BioMed research international*, vol. 2015, Apr. 2015.

[16] P. R. Caron, M. D. Mullican, R. D. Mashal, K. P. Wilson, M. S. Su, M. A. Murcko, "Chemogenomic approaches to drug discovery," *Current opinion in chemical biology*, vol. 5, no. 4, pp. 464-470, Aug. 2001.

[17] X. Periole, A. M. Knepp, T. P. Sakmar, S. J. Marrink, T. Huber, "Structural determinants of the supramolecular organization of G protein-coupled receptors in bilayers," *Journal of the American Chemical Society*, vol. 134, no. 26, pp. 10959-10965, Jun. 2012.

[18] A. L. Hopkins, "Predicting promiscuity," *Nature*, vol. 462, no. 7270, pp. 167-168, Nov. 2009.

[19] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, B. K. Shoichet, "Relating protein pharmacology by ligand chemistry," *Nature biotechnology*, vol. 25, no. 2, pp. 197-206, Feb. 2007.

[20] E. van der Horst, J. E. Peironcely, A. P. IJzerman, M. W. Beukers, J. R. Lane, H. W. Vlijmen, M. T. Emmerich, Y. Okuno, A. Bender, "A novel chemogenomics analysis of G protein-coupled receptors (GPCRs) and their ligands: a potential strategy for receptor de-orphanization," *Bmc Bioinformatics*, vol. 11, no. 1, pp. 1-12, Jun. 2010.

[21] F.-R. Meng, Z.-H. You, X. Chen, Y. Zhou, J.-Y. An, "Prediction of drug–target interaction networks from the integration of protein sequences and drug chemical structures," *Molecules*, vol. 22, no. 7, pp. 1119, Jun. 2017.

[22] L. Wang, Z.-H. You, L.-P. Li, X. Yan, W. Zhang, "incorporating chemical sub-structures and protein evolutionary information for inferring drug-target interactions," *Scientific reports*, vol. 10, no. 1, pp. 1-11, Apr. 2020.

[23] Nidhi, M. Glick, J. W. Davies, J. L. Jenkins, "Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases," *Journal of chemical information and modeling*, vol. 46, no. 3, pp. 1124-1133, Mar. 2006.

[24] Y. Liu, M. Wu, C. Miao, P. Zhao, X.-L. Li, "Neighborhood regularized logistic matrix factorization for drug-target interaction prediction," *PLoS computational biology*, vol. 12, no. 2, pp. e1004760, Feb. 2016.

[25] L. Wang, Z.-H. You, X. Chen, X. Yan, G. Liu, W. Zhang, "Rfdt: A rotation forest-based predictor for predicting drug-target interactions using drug structure and protein sequence information," *Current Protein and Peptide Science*, vol. 19, no. 5, pp. 445-454, May. 2018.

[26] J.-P. Mei, C.-K. Kwoh, P. Yang, X.-L. Li, J. Zheng, "Drug–target interaction prediction by learning from local information and neighbors," *Bioinformatics*, vol. 29, no. 2, pp. 238-245, Jan. 2013.

[27] Y.-A. Huang, Z.-H. You, X. Chen, "A systematic prediction of drug-target interactions using molecular fingerprints and protein sequences," *Current Protein and Peptide Science*, vol. 19, no. 5, pp. 468-478, May. 2018.

[28] H. Chen, Z. Zhang, "A semi-supervised method for drug-target interaction prediction with consistency in networks," *PloS one*, vol. 8, no. 5, pp. e62975, May. 2013.

[29] Z.-H. You, L.-P. Li, X. Yan, W. Zhang, H.-F. Wang, "DTIRF: Predicting Drug-Target Interactions Based on Improved Rotation Forest from Drug Molecular Structure and Protein Sequence," Oct. 2019.

[30] W. Zhang, Y. Chen, D. Li, X. Yue, "Manifold regularized matrix factorization for drug-drug interaction prediction," *Journal of biomedical informatics*, vol. 88, pp. 90-97. Dec. 2018.

[31] S. Dere, S. Ayvaz, "Prediction of Drug–Drug Interactions by Using Profile Fingerprint Vectors and Protein Similarities," *Healthcare Informatics Research*, vol. 26, no. 1, pp. 42-49. Jan. 2020.

[32] R. Ferdousi, R. Safdari, Y. Omidi, "Computational prediction of drug-drug interactions based on drugs functional similarities," *Journal of biomedical informatics*, vol. 70, pp. 54-64. Jun. 2017.

[33] N. Rohani, C. Eslahchi, "Drug-Drug interaction predicting by neural network Using integrated Similarity," *Scientific reports*, no. 9, no.1, pp. 1-11. Sep. 2019.

[34] L. Liu, L. Chen, Y. H. Zhang, L. Wei, S. Cheng, X. Kong, Y. D. Cai, "Analysis and prediction of drug–drug interaction by minimum redundancy maximum relevance and incremental feature selection," *Journal of Biomolecular Structure and Dynamics*, vol. 35, no. 2, pp. 312-329. Apr. 2016.

[35] W. Zhang, Y. Chen, D. Li, "Drug-target interaction prediction through label propagation with linear neighborhood information," *Molecules*, vol. 22, no. 12, pp. 2056. Nov. 2017.

[36] Sridhar, D., Fakhraei, S., & Getoor, L. (2016). A probabilistic approach for collective similarity-based drug–drug interaction prediction. Bioinformatics, 32(20), 3175-3182.

[37] I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, D. Schomburg, "BRENDA, the enzyme database: updates and major new developments," *Nucleic acids research*, vol. 32 no. suppl_1, pp. D431-D433, Jan. 2004.

[38] Z. Wu, F. Cheng, J. Li, W. Li, G. Liu, Y. Tang, "SDTNBI: an integrated network and chemoinformatics tool for systematic prediction of drug–target interactions and drug repositioning," *Briefings in bioinformatics*, vol. 18, no. 2, pp. 333-347, Mar. 2017.

[39] A. Ezzat, M. Wu, X.-L. Li, C.-K. Kwoh, "Drug-target interaction prediction using ensemble learning and dimensionality reduction," *Methods*, vol. 129, no. 1, pp. 81-88, Oct. 2017.

[40] Y. Ding, J. Tang, F. Guo, "Identification of drug-target interactions via multiple information integration," *Information Sciences*, vol. 418, pp. 546-560, Dec. 2017.

[41] M. Gribskov, A. D. McLachlan, D. Eisenberg, "Profile analysis: detection of distantly related proteins," *Proceedings of the National Academy of Sciences*, vol. 84, no. 13, pp. 4355-4358, Jul. 1987.

[42] G. Raicar, H. Saini, A. Dehzangi, S. Lal, A. Sharma, "Improving protein fold recognition and structural class prediction accuracies using physicochemical properties of amino acids," *Journal of theoretical biology*, vol.402, no. 7, pp. 117-128, Aug. 2016.

[43] S. F. Altschul, E. V. Koonin, "Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases," *Trends in biochemical sciences*, vol. 23, no. 11, pp. 444-447, Nov. 1998.

[44] T. Chakraborti, B. McCane, S. Mills, U. Pal, "LOOP descriptor: encoding repeated local patterns for fine-grained visual identification of lepidoptera," *Comput Vis Pattern Recogn*, 2017.

[45] T. Chakraborti, B. McCane, S. Mills, U. Pal, "LOOP Descriptor: Local Optimal-Oriented Pattern," *IEEE Signal Processing Letters*, vol. 25, no. 5, pp. 635-639, May. 2018.

[46] T. Ojala, M. Pietikainen, D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions," *Proceedings of 12th International Conference on Pattern Recognition*, *IEEE*, vol. 1, pp. 582-585, Oct. 1994.

[47] T. Jabid, M. H. Kabir, O. Chae, "Gender classification using local directional pattern (LDP)," *In: 2010 20th International Conference on Pattern Recognition*, *IEEE*, pp. 2162-2165, Oct. 2010.

[48] J. J. Rodriguez, L. I. Kuncheva, C. J. Alonso, "Rotation forest: A new classifier ensemble method," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 10, pp. 1619-1630, Aug. 2006.

[49] V. Ojansivu, J. Heikkilä, "Blur insensitive texture classification using local phase quantization," *In International conference on image and signal processing. Springer, Berlin, Heidelberg.* pp. 236-243. July. 2008.

[50] C.-C. Chang, C.-J. Lin, "LIBSVM: A library for support vector

machines," *ACM transactions on intelligent systems technology*, vol. 2, no. 3, pp. 1-27, May. 2011.

[51] H. Öztürk, E. Ozkirimli, A. Özgür, "A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction," *BMC bioinformatics*, vol. 17, no. 1, pp. 128. Mar. 2016.

[52] M. Gönen, "Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization," *Bioinformatics*, vol. 28, no. 18, pp. 2304-2310, Sep. 2012.

[53] K. Pliakos, C. Vens, G. Tsoumakas, "Predicting drug-target interactions with multi-label classification and label partitioning," *IEEE/ACM transactions on computational biology and bioinformatics*, pp.1-1. Nov. 2019.

[54] Z. Mousavian, S. Khakabimamaghani, K. Kavousi, A. Masoudi-Nejad, "Drug–target interaction prediction from PSSM based evolutionary information," *Journal of pharmacological and toxicological methods*, vol. 78, pp. 42-51. Apr. 2016.

[55] Y. Yamanishi, M. Kotera, M. Kanehisa, S. Goto, "Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework," *Bioinformatics*, vol. 26, no. 12, pp. i246-i254. Jun. 2010.