Current Protein and Peptide Science, 2018, *19*, 468-478

RESEARCH ARTICLE

A Systematic Prediction of Drug-Target Interactions Using Molecular Fingerprints and Protein Sequences

Yu-An Huang[1,#], Zhu-Hong You[2,#,*] and Xing Chen[3,*]

[1]*Department of Computing, Hong Kong Polytechnic University, Hung Hom, Hong Kong;* [2]*Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Ürümqi 830011, China;* [3]*School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, 221116, China*

Abstract: ***Background:*** Drug-Target Interactions (DTI) play a crucial role in discovering new drug candidates and finding new proteins to target for drug development. Although the number of detected DTI obtained by high-throughput techniques has been increasing, the number of known DTI is still limited. On the other hand, the experimental methods for detecting the interactions among drugs and proteins are costly and inefficient.

Objective: Therefore, computational approaches for predicting DTI are drawing increasing attention in recent years. In this paper, we report a novel computational model for predicting the DTI using extremely randomized trees model and protein amino acids information.

Method: More specifically, the protein sequence is represented as a Pseudo Substitution Matrix Representation (Pseudo-SMR) descriptor in which the influence of biological evolutionary information is retained. For the representation of drug molecules, a novel fingerprint feature vector is utilized to describe its substructure information. Then the DTI pair is characterized by concatenating the two vector spaces of protein sequence and drug substructure. Finally, the proposed method is explored for predicting the DTI on four benchmark datasets: Enzyme, Ion Channel, GPCRs and Nuclear Receptor.

Results: The experimental results demonstrate that this method achieves promising prediction accuracies of 89.85%, 87.87%, 82.99% and 81.67%, respectively. For further evaluation, we compared the performance of Extremely Randomized Trees model with that of the state-of-the-art Support Vector Machine classifier. And we also compared the proposed model with existing computational models, and confirmed 15 potential drug-target interactions by looking for existing databases.

Conclusion: The experiment results show that the proposed method is feasible and promising for predicting drug-target interactions for new drug candidate screening based on sizeable features.

ARTICLE HISTORY

Received: October 26, 2017
Revised: November 04, 2017
Accepted: December 05, 2017

DOI:
10.2174/1389203718666161122103057

Keywords: Drug-target interactions, pseudo substitution matrix representation, drug substructure fingerprint, extremely randomized trees, computational model.

1. INTRODUCTION

The identification of drug-target interactions (DTI) has recently emerged as an area of intense research activity due to its crucial role in discovering new drug candidates and finding new proteins to target for drug development. However, the knowledge of drug-target interactions is still deficient and only a small share of them is experimentally tested and is detected as interactive. Much effort has been devoted to use experimental methods to identify drug-protein interactions but the experimental tests are both costly and difficult. It often costs billions of dollars for developing a successful novel chemistry-based drug and nearly a decade for introducing the drug to market. Only few of the drug candidates can be approved to reach the market by Food and Drug Administration (FDA) while most of them fail during clinical trials showing adverse side effects. However, recent researches have definitely showed that the interactions between drugs and some proteins related to specific toxicity greatly influence the side-effects or toxicity of drug compounds. Identifying protein-target interactions help understanding the toxicity of drug candidates. Furthermore, it also contributes to finding new potential targets for an old drug which provides insights into its potential toxicity or new application to treating other disease. Due to the inevitable drawbacks of experimental methods, computational approaches for predicting drug-target interactions have gained increasing attention in recent years. Screening databases of small molecules against certain classes of protein, computational methods can potentially find some drug candidates with better bio-activity from the statistical perspective and therefore accelerates drug discovery.

Due to the development of molecular medicine and the completion of the human genome project, the amount of available knowledge in biology and chemistry rapidly in-

*Address correspondence to these authors at the Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Ürümqi 830011, China; Tel: +86-1816062862; Fax: +86-0991-3838957;
E-mails: zhuhongyou@gmail.com; xingchen@amss.ac.cn
#*The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.*

1875-5550/18 $58.00+.00

creases and enables the researchers to retrieve compound information and properties and study drug-target interaction problems by a systematic integration. A number of databases have been built for storing these data and some of them focus on drug-target relations such as DrugBank [1], Therapeutic Target Database (TTD) [2], Kyoto Encyclopedia of Genes and Genomes (KEGG) [3], SuperTarget and Matador [4], *etc.* These accumulated data of these databases offer significant material for the researches of prediction methods on a genome-wide scale.

A number of computational methods have been proposed for predicting drug-target interactions and most of them belong to two categories: Docking simulation and machine learning. Docking simulation is an effective molecular modeling approach which uses dynamic simulation to predict the positive interactions when drug molecule and protein bound to each other. However, this method usually requires the three-dimensional (3D) structure data of targets (traditional docking) or a large set of drugs (inverse docking). Up to now, the proteins with known 3D structures cover only a small part of all proteins and therefore this requirement is difficult to meet. Compared with the known 3D structure data, the amount of detected protein sequence data are relatively larger and increases exponentially. Hence, it is more practical to build a computational model for predicting DTI based on protein sequence data.

Existing computational models for predicting drug-target interactions usually represent the known drug-target interactions as a bipartite graph. In this bipartite graph, nodes denote drugs and targets and the interactions between them are represented by the edges between these nodes. Therefore, we can treat the DTI prediction problem equivalent to predicting new edges in the bipartite graph. Based on the topology of the graph, some of algorithms have been proposed for predicting the new drug-target interaction. To predict if a drug interacts with a target, these approaches consider the edges involving these two nodes [5]. They can be generally divided into three kinds of supervised inference methods: drug-based similarity inference (DBSI), target-based similarity inference (TBSI) and network-based inference (NBI). Cheng *et al*. [6] has proposed a method derived from the recommendation algorithms of complex network theory for predicting DTIs and this method is based on network-based inference. Fakhraei *et al*. [7] have proposed a prediction model which uses probabilistic soft logic and considers both target-target similarity and drug-drug similarity. However, these algorithms do not work well in the case that the predicted interaction involves a "new drug" or a "new target". Herein, a "new drug" means a drug candidate without any interactions and a "new target" means a target protein without any interactions. In addition, these algorithms do not consider the biological information in the protein domain.

Another popular approach for predicting DTIs is to use machine learning techniques to build a classification model which consider each drug-target pair as one sample. The drug-target pairs which are known to interact are labeled as positive. Each sample is represented by a feature vector which is composed by a drug feature vector and a protein feature vector. The drug features usually extracted from the two-dimensional chemical structures. Francisco *et al*. [8] have proposed a method for predicting DTIs which calculates 2D molecular descriptors for drug feature extraction. Chen *et al*. [9] consider the information of chemical-chemical similarities, chemical-chemical connections and chemical-protein connections and propose an effective classifier for identifying drug-target groups. However, these methods do not take the biological interpretation into account.

In this article, we report a novel computational model for predicting drug-target interactions. We formulate this prediction problem as an extended structure-activity relationship (SAR) classification problem assuming that the interactions between drugs and target proteins greatly depend on the structure information not only from the molecular substructure fingerprints of drug compounds but also from target protein sequences. The positive sample set is constructed by the known interactional drug-target pairs and the negative sample set is randomly connected from the other pairs. We represent drugs by utilizing their molecular substructure fingerprints and encode protein sequence using a novel feature extraction method called Pseudo-SMR. Here, we explore Extremely Randomized Trees (ETs) classifier for building prediction model for four kinds of pharmaceutically useful protein target: *enzymes*, *ion channels*, *GPCRs* and nuclear receptors. ER-trees have inherent advantages to deal with drug-target prediction problem due to its distinctive characters: explicitly randomized cut-points and attribute which can reduce variance, and the usage of the original training set which helps to minimize bias. The goal of our study is to establish an effective prediction model for finding new drug-target interactions and to provide deeper insights into DTIs by seeking the influential factors.

2. MATERIALS AND METHODS

2.1. Golden Standard Datasets

In this study, we explore the proposed method for predicting drug-target interactions on four types of protein targets: *enzymes*, *ion channels*, *GPCRs* and *nuclear receptors*. These data are collected from the KEGG BRITE [3], BRENDA [10], SuperTarget & Matador [4] and DrugBank [1] databases and used as the gold standard datasets by Yamanishi *et al*. [11]. The numbers of drugs known to target enzymes, ion channels, GPCRs and nuclear receptors are 445, 210, 233 and 54 respectively. The numbers of protein known to be targeted by the drugs are 664, 204, 95 and 26 respectively. Among these drug-target pairs, 5127 pairs of them are known to interact with each other. The numbers of known interactions involving enzymes, ion channels, GPCRs and nuclear receptors are 2926, 1476, 635 and 90 respectively. We finally use these known interactions to construct all the four positive sample sets.

Drug-target interaction network is usually modeled as a bipartite graph, where the initial edges describe the real drug-target interactions already detected by experiments. Compared with a completely connected bipartite graph, the number of initial edges is relatively small [12]. Take the *enzymes* dataset for an example, there totally exists to be 445×664=295480 connections in the corresponding bipartite graph. However, there are only 2926 initial edges which represent the known drug-target interactions. Therefore, the

number of positive samples (*e.g.,* 2926) is significantly smaller than the possible number of negative samples (*e.g.,* 295480-2926=292554), which presents a bias problem. To solve this problem we randomly collected the negative samples with the same size of the positive sample datasets. Therefore, the sample numbers of *enzymes*, *ion channels*, *GPCRs* and *nuclear receptors* datasets are 2926, 1476, 635 and 90 respectively. In fact, such negative sample sets may possibly contain drug-target pairs that interact really. However, in view of the large scale of DTI bipartite graph, the number of real interactions pairs which are possibly collected in negative sets is very small.

2.2. Drug Molecules Representation

Different kinds of descriptors for representing drug compounds have been proposed, such as geometrical, topological, constitutional and quantum chemical properties. Recently, some current researches [13] indicate that it is effective to use a variety of molecular substructure fingerprints to represent drug compounds. Substructure fingerprint describes the structure information of a given drug compound using a Boolean substructure vector. It separates the drug molecule into fragments and records the existence of the substructures. Specifically, the pattern of substructure vectors is predefined according to the substructure dictionary and each binary bit in the fingerprint vector denote the presence or absence of a particular substructure. If a substructure exists in a given drug molecule, the corresponding bit in the vector is set to be 1; conversely, it is set to be 0 if the substructure is absent. In this way, substructure fingerprint is capable of describing the complex structure of drug molecules. In this work, the chemical structures fingerprints set we used are collected from the PubChem System. Fingerprints property is "PUBCHEM_CACTVS_ SUBGRAPHKEYS" in PubChem and Base64 encoded which provides a textual description by the binary data. The a drug fingerprint records the information of 881 substuctures and therefore a drug moleclue feature is a 881 binary vector.

2.3. Target Protein Representation

A number of feature extraction methods have been proposed for representing protein sequences. Effective protein feature descriptors can mine the significant information and therefore boost the performance of protein-associated prediction models, such as protein function prediction model and protein-protein interaction prediction model. Most of these feature extraction methods derive from the concept of Chou's pseudo amino acid (PseAA) composition. Extracted from the original protein sequences, this kind of feature descriptors expends the simple amino acid composition by taking the information of sequence order into account. However, there are some novel feature extraction methods utilizing biological kernels. Jaakkola *et al*. [14] have proposed Fisher kernel considering homology information. Leslie *et al*. [15] have put forward another mismatch string kernel for protein sequence representation. Unlike the PseAAC-based methods which extract feature directly from protein sequences, these kernel-based methods remain some kinds of prior biological information and extract more comprehensive feature descriptors.

As a variant of representation method proposed by [16], Substitution Matrix Representation (SMR) is used for the representation of protein feature. Given any protein sequence of length N, this method transform it into an N×20 matrix by using a specific substitution matrix which helps retaining the evolutionary information. In this article, we use the BLOSUM62 matrix (*i.e.,* blocks of amino acid substitution matrix used for comparing the sequences with 62% similarity) for this transformation. Depicting the substitution probabilities of amino acids, this matrix and often used to score alignments between evolutionarily divergent protein sequences. In this transformation, SMR can be depicted as follow:

$$SMR_{N\times 20}=\begin{bmatrix} B_{1,1} & B_{1,2} & \cdots & B_{1,20} \\ B_{2,1} & B_{2,2} & \cdots & B_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ B_{N,1} & B_{N,2} & \cdots & B_{N,20} \end{bmatrix} \quad (1)$$

where N is the length of the given protein sequence; $B_{i,j}$ denotes the rate value of BLOSUM62 and represents the possibility that the *ith* amino acid of the given protein sequence mutates to amino acid j in the evolution process. Since the lengths of protein sequences are different, the SMR matrixes from this transformation cannot be used as feature descriptors directly. To address this problem, we adopt the concept of pseudo amino acid composition in the second step of the feature extraction. Similar to reports in [17], we further considered the neighborhood of each amino acid residues. Specifically, the pseudo-SRM descriptor is obtained from a SMR matrix by the following equations:

$$PseudoSMR(n)=\begin{cases} \frac{1}{N}\sum_{i=1}^{N} M(i,j) & n=1,...,20 \\ \frac{1}{N-\lg}\sum_{i=1}^{N-\lg}[M(i,j)-M(i+\lg,j)]^2 & \begin{matrix} n=20+j+20\cdot(\lg-1) \\ j=1,...,20;\ \lg=1,...,\max lag \end{matrix} \end{cases} \quad (2)$$

and

$$M(i,j)=\frac{SMR(i,j)-\frac{1}{20}\sum_{a=1}^{20}SMR(i,a)}{\sqrt{\frac{1}{20}\sum_{b=1}^{20}\left(SMR(i,b)-\frac{1}{20}\sum_{a=1}^{20}SMR(i,a)\right)^2}} \quad (3)$$

$i = 1..N;\ j = 1..20$

where *lg* represents the distance between one residue and its neighbors; N is the length of a given protein sequence; M is a normalized version of SMR matrix. In this work, we set the value of *maxlag* to be 15. Therefore, every protein sequence is represented by a Pseudo-SMR feature whose length is 320. Considering the features of drug molecules are binary, we normalize the Pseudo-SMR feature into the range from 0 to 1.

2.4. Extremely Randomized Trees

Introduced by Geurts *et al*. [18], Extremely Randomized Tree (ER-Tree) growing algorithm combines the attribute randomization of random subspace with a totally random selection of the cut-point. Extremely randomized tree is a tree-based ensemble method which constructs an ensemble of unpruned decision or regression trees through a top-down procedure. Unlike other tree-based ensemble methods, the ER-Tree algorithm splits nodes by choosing cut-points fully

at random. In addition, it uses all the learning samples (rather than a bootstrap replica) for growing the trees. Specifically, the procedure to build ER-trees works in a recursive fashion by successively splitting the nodes until all the output variable or candidate attributes are constant in training set. In the first step of this recursive procedure, given a training set T with m-dimensional attributes A=$\{a_1,a_2,...,a_m\}$, K attributes, $\{a_1,..., a_k\}$ would be randomly selected without replacement. For each selected attribute, its split value is then randomly generated as $\{v_1,\ldots,v_k\}$. Define the maximal and minimal value of an attribute, a, as $a^s_{\min}$ and $a^s_{\max}$, the split value would be drawn uniformly in [$a^s_{\min}$, $a^s_{\max}$]. All split would be evaluated by a scoring formula as follow:

$$Score_c(s,T) = \frac{2I^s_c(T)}{H_S(T) + H_c(T)} \tag{4}$$

$$H_s(T) = -(\frac{|T_L|}{|T|}\log_2\frac{|T_L|}{|T|} + \frac{|T_R|}{|T|}\log_2\frac{|T_R|}{|T|}) \tag{5}$$

$$H_c(T) = -\sum_{i=1}^{c} p_i \log_2 p_i \tag{6}$$

$$I^s_c(T) = H_c(T) - \frac{|T_R|}{|T|}H_c(T_R) - \frac{|T_L|}{|T|}H_c(T_L) \tag{7}$$

where T_R and T_L are two subtrees divided by the split s from T; $H_C(T)$ denotes the entropy of the classification in T; $H_C(T)$ denotes the split entropy; $I^S_C(T)$ denotes the mutual information of the split outcome and the classification and measures the ability of split s to produce pure successors.

Based on the computed evaluation scores, the split would be chosen with the maximal score:

$$Score(s^*,T) = \max_{i=1,\ldots,K} Score(s_i,T) \tag{8}$$

According to the s*, T is then divided into left and right subtree, T_L and T_R. Both of subtrees are attached to the split node and then separated in the next iteration.

The reason of good accuracy of ER-trees mainly lies in the explicitly randomized cut-points and attribute which can reduce variance, and the usage of the original training set which helps to minimize bias. By averaging over a sufficiently large ensemble of trees, variance caused by randomization can be canceled. In particular, ER-trees can tolerate high levels of bias of class probability estimates when dealing with classification problems.

In this work, we explore this method for predicting the drug-target interactions. We employ scikit-learn to implement ER-trees. Scikit-learn is a python package for machine learning and supports popular and powerful tools for data mining. It is available at http://scikit-learn.org/stable/index.html. In this work, we set the parameters of ER-trees classifier to be the same for all experiments (*i.e., n_estimators*=2000; *max_features*="sqrt").

3. RESULTS AND DISCUSSION

3.1. Evaluation Criteria

To evaluate the performance of the proposed method, we use the following criteria: the overall prediction accuracy (Accu.), sensitivity (Sens.), precision (Prec.) and Matthews correlation coefficient (MCC) were calculated. They are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{10}$$

$$\text{Pr}ecision = \frac{TP}{TP + FP} \tag{11}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)\times(TN + FP)\times(TP + FP)\times(TN + FN)}} \tag{12}$$

where true positive (TP) is the number of drug-target pairs classified as interacting correctly; false positive (FP) is the number of samples classified as interacting incorrectly; true negative (TN) is the number of samples classified as non-interacting correctly; false negative (FN) is the number of samples classified as non-interacting incorrectly. For a deeper evaluation, we also compute the receiver operating characteristic (ROC) curve. To summarize ROC curve in a numerical way, the area under an ROC curve (AUC) was computed.

3.2. Measurement of Structural Diversity of Drug and Protein Molecules

A number of methods have been proposed for measuring the molecular similarity and diversity. In this work, we calculate the average value of the dissimilarity between all the pairwise drug molecules for this evaluation [19]. Specifically, the average dissimilarity of a given drug fingerprint dataset G is calculated as follow:

$$Diversity(G) = \frac{\sum_{i=1}^{N(G)}\sum_{j=1,i\neq j}^{N(G)}[1 - similarityd(i,j)]}{N(G)[N(G)-1]} \tag{13}$$

$$similarityd(f1,f2) = \frac{c}{a+b-c} \tag{14}$$

where $N(G)$ is the number of samples in dataset G; *similarityd(*)* is a function to calculate the similarity between two drug fingerprints by using Tanimoto coefficient; c is the number of mutual substructure existing in both fingerprint f1 and fingerprint f2; a and b are the substructure numbers existing in fingerprint f1 and fingerprint f2 respectively.

For evaluating the diversity of protein sequences, we calculate the similarity between two protein sequences using a normalized version of Needleman–Wunsch score. It is popular to use Needleman–Wunsch algorithm for calculating the alignment score between two protein sequences. Specifically, the average dissimilarity of a given protein sequence dataset P is calculated as follow:

$$Diversity(P) = \frac{\sum_{i=1}^{N(P)} \sum_{j=1, i \neq j}^{N(P)} [1 - similarityp(i, j)]}{N(P)[N(P) - 1]} \quad (15)$$

$$similarityp(p1, p2) = \frac{nw(p1, p2)}{\sqrt{nw(p1, p1)} \times \sqrt{nw(p2, p2)}} \quad (16)$$

where *N(P)* is the number of samples in dataset *P*; *similarityp*(*,*) is a function to calculate the average normalized Needleman-Wunsch between two protein sequences; *nw*(*,*) is the original Needleman-Wunsch function which return the Needleman-Wunsch scores. Herein, we set match award score, mismatch penalty and gap penalty as 5, -1 and -1 respectively. Obviously, the similarity value of 0 for a pair of protein sequences or drug substructure fingerprints denotes that two molecules have no any similarity which the similarity value of 1 means that the information of two molecules are totally the same. Therefore, for any given dataset *A*, the closer Diversity(A) is to 1, the more diverse the dataset *A* is.

Fig. (**1**) shows the analysis results for drug molecules and protein sequences of the four datasets. The average normalized Needleman-Wunsch scores of the datasets of *enzymes*, *ion channels*, *GPCRs* and *nuclear receptors* are 82.30%, 86.26%, 75.02% and 73.97%. The dissimilarity scores of the datasets of *enzyme*, *ion channel*, *GPCRs* and *nuclear receptor* are 65.17%, 61.43%, 59.55% and 53.47%. The dissimilarity score of drug is lower than the protein sequence. These results suggest that the protein sequence datasets are very structurally diverse and complex and that the drug compounds datasets are less diverse and have relatively moderate similarity compared with the protein datasets. It is indicated that the interactions experimented in this work are functionally and structurally diverse.

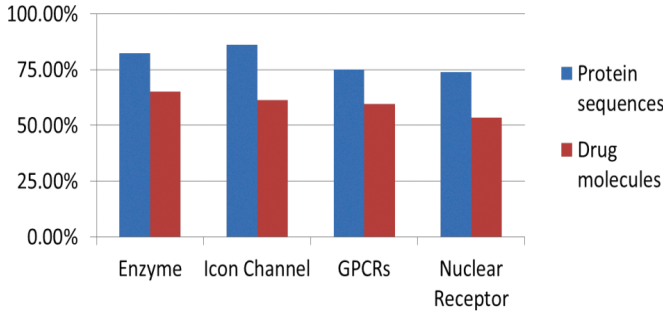

Fig. (1). The diversity analysis of drugs and target proteins for four datasets.

3.3. Assessment of Prediction Ability

For the fairness, we set the same corresponding parameters of ER-Trees classifier when performing on the four different datasets: *enzymes*, *ion channels*, *GPCRs* and *nuclear receptors*. In this work, five-fold cross validation is used for evaluating the prediction performance of the proposed method. Specifically, we evenly divide the dataset into five parts of which four are used for training the ER-Trees and the other part is used for testing. The process is repeated five times and every part can be predicted as a validation set.

Tables **1-4** list the 5-fold cross validation results performed by the proposed model on four datasets (*i.e., enzymes*, *ion channels*, *GPCRs* and *nuclear receptors*). When exploring the *Enzyme* dataset, we obtained the good results of average accuracy, precision, sensitivity and MCC of 89.94%, 90.27%, 89.58% and 81.89% respectively. The standard deviations of these criteria values are 0.83%, 2.09%, 1.04% and 1.34% respectively. When predicting drug-target interactions of *Icon Channel* dataset, the proposed method yielded results of average accuracy, precision, sensitivity and MCC of 87.87%, 87.86%, 87.94% and 78.71% and the standard deviations are 1.34%, 1.95%, 1.13% and 2.03% respectively. When predicting DTIs of *GPCRs* dataset, the averages of accuracy, precision, sensitivity and MCC come to be 82.91%, 82.18%, 83.98% and 71.58% and the standard deviations are 1.51%, 4.54%, 2.32% and 1.97% respectively. When predicting DTIs of *Nuclear Receptor* dataset, the averages of accuracy, precision, sensitivity and MCC come to be 83.33%, 76.39%, 95.29% and 71.56%. However, since the number of samples of *Nuclear Receptor* dataset is only 90, relatively smaller than other datasets, it yields the highest standard deviations which are 5.20%, 8.04%, 2.74% and 7.43% respectively. Figs. (**2-5**) shows the ROC curves performed by the proposed method on *enzymes*, *ion channels*, *GPCRs* and *nuclear receptors*. The average AUC values range from 90.53% to 96.34% (*Enzyme*: 96.01%, *Icon Channel*: 93.82%, GPCRs: 90.53% and Nuclear Receptor: 96.34%), suggesting that a big separation for two classes is indeed obtained from ER-Trees.

These good results collectively suggest that the information including protein sequences and drug substructure fingerprints is sufficient enough for predicting whether a given drug-protein pair interact or not, and that powerful prediction capability for predicting drug-target interactions can be obtained by using a ER-Trees-based model combined Pseudo-SMR protein features and drug substructure fingerprints. This strong prediction performance derives from the feature extraction method for protein sequences and the choice of machine learning classifier. Pseudo-SMR descriptors not only quantitatively describe the differences between amino acids, but also partially incorporate the sequence-order information. ER-Trees classifier performs well due to the ensemble model and its novel random tree splitting strategy.

3.4. Comparison Between Pseudo-SMR Descriptor and Pseudo-AAC Descriptor

In order to evaluate the importance of the transformation of SMR, we compare its performance with that of Pseudo-AAC on *Enzyme* dataset in this section. The Pseudo-AAC here we explored is a variant version which computes the autocovariance and retain the hydrophobicity information of amino acids. Both Pseudo-SMR and Pseudo-AAC descriptors consider the influence of neighbor residues. However, unlike Pseudo-SMR, Pseudo-AAC descriptors are extracted directly from the original protein sequences and therefore do not retain the information from biological substitution matrix. Given a protein sequence *P*, Pseudo-AAC descriptors can be defined in a $20+\lambda$ dimensional space. Herein, we set λ to be 20. It can be formulated as follow:

$$p = [p_1 \quad p_2 \quad \cdots \quad p_{20} \; p_{21} \quad \cdots \quad p_{20+\lambda}]^T \quad (17)$$

Table 1. 5-fold cross validation results performed by proposed model on *Enzyme* dataset.

Test set	Accu.(%)	Prec.(%)	Sen.(%)	MCC(%)	AUC(%)
1	90.34	91.91	88.26	82.53	95.51
2	91.11	92.47	89.64	83.80	96.51
3	89.74	89.34	89.65	81.58	96.41
4	89.57	90.40	89.20	81.31	95.56
5	88.91	87.25	91.13	80.26	96.06
Average	**89.94±0.83**	**90.27±2.09**	**89.58±1.04**	**81.89±1.34**	**96.01±0.46**

Table 2. 5-fold cross validation results performed by proposed model on Icon Channel dataset.

Test set	Accu.(%)	Prec.(%)	Sen.(%)	MCC(%)	AUC(%)
1	87.46	88.14	86.96	78.06	93.36
2	89.83	90.36	88.46	81.70	95.38
3	88.31	88.96	88.08	79.34	93.79
4	87.63	86.27	89.49	78.30	93.89
5	86.15	85.57	86.73	76.13	92.67
Average	**87.87±1.34**	**87.86±1.95**	**87.94±1.13**	**78.71±2.03**	**93.82±1.00**

Table 3. 5-fold cross validation results performed by proposed model on GPCRs dataset.

Test set	Accu.(%)	Prec.(%)	Sen.(%)	MCC(%)	AUC(%)
1	81.89	75.40	86.36	70.07	91.93
2	84.25	86.43	85.21	73.14	90.65
3	84.65	86.29	82.95	74.00	90.88
4	81.10	81.75	80.47	69.35	87.51
5	82.68	81.06	84.92	71.33	91.66
Average	**82.91±1.51**	**82.18±4.54**	**83.98±2.32**	**71.58±1.97**	**90.53±1.77**

Table 4. 5-fold cross validation results performed by proposed model on Nuclear Receptor dataset.

Test set	Accu.(%)	Prec.(%)	Sen.(%)	MCC(%)	AUC(%)
1	83.33	78.26	94.74	71.11	95.67
2	83.33	76.19	94.12	71.79	94.43
3	80.56	75.00	94.74	66.89	95.98
4	77.78	65.00	92.86	64.34	95.78
5	91.67	87.50	100.00	83.67	99.68
Average	**83.33±5.20**	**76.39±8.04**	**95.29±2.74**	**71.56±7.43**	**96.34±1.98**

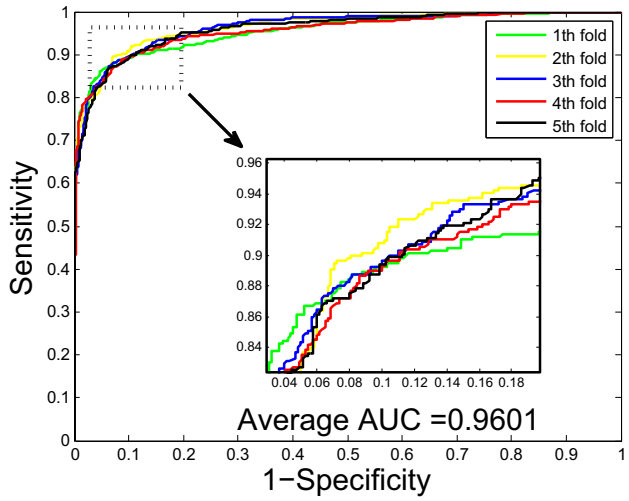

Fig. (2). ROC curves performed by proposed method on *Enzyme* dataset.

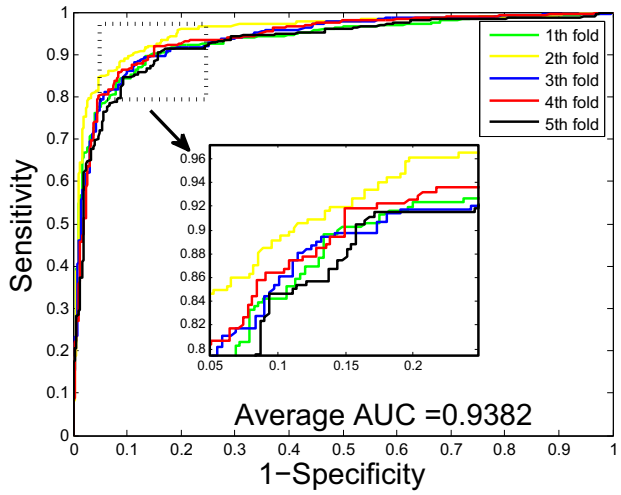

Fig. (3). ROC curves performed by proposed method on *Icon Channel* dataset.

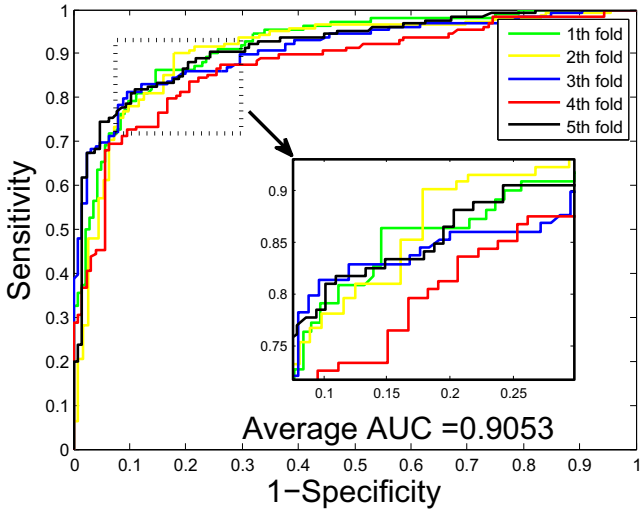

Fig. (4). ROC curves performed by proposed method on *GPCRs* dataset.

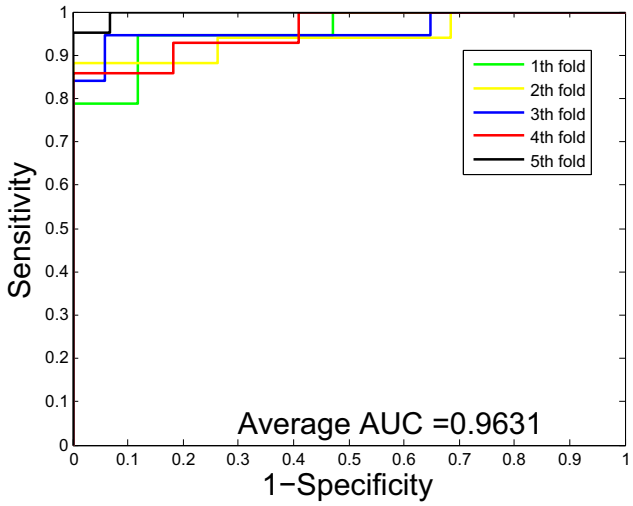

Fig. (5). ROC curves performed by proposed method on *Nuclear Receptor* dataset.

where $p_1,p_2,...p_{20}$ is the values of the conventional amino acid composition, while the rest components are obtained by adopting autocovariance approach (AC) method:

$$AC(i) = \sum_{k=1}^{N-i+20} \frac{(index(p_k)-\mu)\cdot(index(p_{k+i-20})-\mu)}{\sigma\cdot(N-i+20)} \qquad i \in [21,...,20+\lambda] \tag{18}$$

$$\mu = \frac{1}{20}\sum_{i=1}^{20} index(i)\,, \tag{19}$$

$$\sigma = \frac{1}{20}\sum_{i=1}^{20} (index(i)-\mu)^2 \tag{20}$$

where *index(i)* is a function returning the physicochemical property values (hydrophobicity) for the i-th amino acid; μ and σ denote the normalized mean and the variance of hydrophobicity values of the 20 amino acids. By computing AC feature descriptor, we can obtain 20-dimensional feature vectors for retaining the influence of neighbor residues as well as the information of amino acid hydrophobicity.

Table **5** lists list the 5-fold cross validation results performed by the ER-Trees-based model combined with the PseAAC features on the *Enzyme* dataset. The yielded averages of accuracy, precision, sensitivity and MCC come to be 84.96%, 84.37%, 85.82% and 74.44%, significantly lower than those performed by the proposed method, which are 89.85%, 90.31%, 89.33% and 81.76%. In this comparison, the corresponding parameters of ER-Trees and the feature descriptor for drug molecules are the same in these two experiments. From these results, we can see the transformation of SMR can indeed improve the prediction performance of the ER-Trees-based model.

Currently, researches have pointed out that the homology information of protein sequences has significant impact on the prediction accuracy of protein-associated models. The results for highly homologous datasets are usually shown to be statistically significantly higher than those for the datasets with low homology. Homology information can offer useful insights into the relationship between protein and other

Table 5. 5-fold cross validation results performed by ER-Trees model combined with PseAAC descriptors on *Enzyme* dataset.

Test set	Accu.(%)	Prec.(%)	Sen.(%)	MCC(%)
1	84.36	84.42	84.27	73.61
2	84.87	83.59	85.51	74.31
3	84.70	85.08	85.50	74.04
4	86.24	85.07	87.84	76.25
5	84.64	83.67	85.96	73.99
Average	**84.96±0.74**	**84.37±0.73**	**85.82±1.29**	**74.44±1.04**
Proposed method	**89.94±0.83**	**90.27±2.09**	**89.58±1.04**	**81.89±1.34**

molecules and therefore shows great potential in other bioinformatics problems. Using BLOSUM62 matrix, Pseudo-SMR descriptor contains evolutionary information about homologous protein sequence, which allows for estimating the tendency of evolutionary conservation and the various degrees of similarity among the protein sequences. In addition, adapting the concept of Chou's pseudo amino acid composition, Pseudo-SMR descriptor not only remains the protein sequence order but also considers the correlations between neighbor residues and the residues with a moderate distance.

3.5. Comparison between ER-Trees and Support Vector Machine

Support Vector Machine (SVM) is one of the most popular machine learning classifiers for data mining. In this section, we explore it with the same feature descriptors on the *Enzyme* dataset for further evaluation of the performance of ER-Trees. The parameters of SVM are optimized by using a grid search method (c=0.5; g=0.4). Table **6** lists list the 5-fold cross validation results performed by the SVM-based model combined with the proposed descriptors on the *Enzyme* dataset. The preformed averages of accuracy, precision, sensitivity and MCC are 81.90%, 89.97%, 71.87% and 74.44%, significantly lower than those performed by the proposed method, which are 89.85%, 90.31%, 89.33% and 81.76%.

From this comparison, we can see that the ER-Trees classifier obtain a better performance with the proposed feature descriptors than SVM classifier. This is benefited from the ensemble model and its random tree splitting strategy. In addition, the parameters of ER-Trees are more easily optimized than those of Support Vector Machine.

3.6. Comparison with Other Methods

For predicting the drug-target interactions, various kinds of computational model have been proposed. To further evaluate the performance of the proposed method, we here compare it with other previously proposed models which apply the same validation framework of 5-fold cross validation and explore the same datasets (*i.e. Enzymes, Icon Channels, GPCRs and Nuclear Receptors*). The comparison results are listed in Table **7**. It is observed that the model we proposed obtain a significant improvement in the prediction performance for drug-target interaction in term of the yielded AUC values. The growths in average values achieve 0.1281, 0.1348, 0.0483 and 0.124 on the datasets of *Enzymes*, *Icon Channels*, *GPCRs* and *Nuclear Receptors*, respectively. The improvement may come from the effective representation of Pseudo-SMR descriptor as well as the powerful prediction ability of ER-Trees. Unlike these comparison methods which are mainly based on the drug/protein network similarity, the proposed model has a wider application and avoids the information bias in the known drug-target interaction network due to the independence from known drug-target interactions.

3.7. Potential Drug-target Interactions of Top-10 Ranks Verified from Databases

After evaluating the effectiveness of the proposed model by using the 5-fold cross validation method, we here calculate the interaction possibility for all potential drug-target pairs in the datasets of *GPCRs* and *Nuclear Receptors*. Specifically, the whole negative and positive data explored in 5-fold cross validation experiments are used for training and all the unknown drug-target pairs are used as training set. The predicted drug-target pairs with top-10 ranks in the drug's potential target lists are considered as highly potential drug-target interactions and further verified by four public databases (*i.e.* KEGG [3], Supertarget [4], Drugbank [1] and ChEMBL [23]). These databases have been supplemented by some newly detected drug-target interactions since the gold standard data explored in this study was collected in 2008. All the predicted possibilities for all potential drug-target interactions in *GPCRs* and *Nuclear Receptors* can be obtained in Supplementary Table **1** and **2**, respectively. As a result, 15 new drug-target interactions are finally confirmed. Specifically, we confirmed 8 and 7 drug-target interactions based on the predicted results on *GPCRs* and *Nuclear Receptors*, respectively. Note that the high-ranked interactions that are not reported yet may also exist in reality. Based on these results, we anticipate that the proposed model is feasible to predict new drug-target interactions.

Table 6. 5-fold cross validation results performed by SVM model combined with the proposed feature descriptors on *Enzyme* dataset.

Test set	Accu.(%)	Prec.(%)	Sen.(%)	MCC(%)
1	82.91	92.20	71.50	70.89
2	80.43	89.82	68.93	67.70
3	84.10	90.17	75.61	72.81
4	80.60	89.14	70.93	68.20
5	81.48	88.52	72.35	69.32
Average	**81.90±1.57**	**89.97±1.40**	**71.87±2.45**	**69.79±2.09**
Proposed method	**89.94±0.83**	**90.27±2.09**	**89.58±1.04**	**81.89±1.34**

Table 7. Prediction performances of DBSI [6], Yamanishi *et al.*(2010) [20], KBMF2K [21], NetCBP [22] and our method on the four benchmark datasets in terms of average AUC values.

Dataset	The proposed model	DBSI	Yamanishi *et al.*(2010)	KBMF2K	NetCBP
Enzymes	**0.9601±0.0046**	0.8075	0.821	0.832	0.8251
Icon Channels	**0.9382±0.0100**	0.8029	0.692	0.799	0.8034
GPCRs	**0.9053±0.0177**	0.8022	0.811	0.857	0.8235
Nuclear Receptors	**0.9634±0.0198**	0.7578	0.814	0.824	0.8394

Table 8. The newly confirmed drug-target interactions with high ranks in the datasets of *GPCRs* and *Nuclear Receptors.*

Drug ID	Target ID	Rank in the drug's potential target proteins	Evidence
D00059	hsa:1814	4	KEGG
D00059	hsa:1816	10	KEGG
D00415	hsa:3355	5	Supertarget,Drugbank
D00419	hsa:5031	6	KEGG
D04625	hsa:154	3	KEGG
D02358	hsa:154	5	Drugbank
D00283	hsa:1814	7	Drugbank
D02349	hsa:154	5	Drugbank
D00182	hsa:2099	3	ChEMBL
D00898	hsa:2100	9	ChEMBL,KEGG
D00348	hsa:6258	7	ChEMBL
D00327	hsa:5915	8	ChEMBL
D00443	hsa:367	7	Supertarget
D00554	hsa:3174	10	KEGG
D00066	hsa:4306	2	Drugbank

CONCLUSION

In the post-genomic era, the knowledge of DTIs play a crucial role in discovering new drug candidates and finding new proteins to target for drug development. However, the experimental methods are expensive and inefficient. In this article, we report a computational model based on extremely randomized trees for predicting the drug-target interactions. An underlying idea of our proposed approach is that the structures of drug molecules and protein amino acids sequence have a great influence on the DTIs. In addition, we take the biological evolutionary information into account in the process of protein feature extraction. From the comparison with PseAAC descriptor, the evolutionary information of BLOSUM62 matrix proves to be useful for predicting DTIs. Our proposed method obtains good results preformed on four different datasets (*i.e., enzymes*, *ion channels*, *GPCRs* and *nuclear receptors*.), suggesting it has ability to predict drug-target interactions from large-scale datasets.

DECLARATIONS

The publication costs for this article were funded by the corresponding author's institution.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No Animals/Humans were used for studies that are base of this research.

CONSENT FOR PUBLICATION

Not applicable.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

This work is supported by National Science Foundation of China, under Grants 61772531, 61572506, 11631014 and Pioneer Hundred Talents Program of Chinese Academy of Sciences. The authors would like to thank all the guest editors and anonymous reviewers for their constructive advices.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's web site along with the published article.

REFERENCES

[1] Wishart, D.S.; Knox, C.; Guo, A.C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. *Nucleic Acids Res.,* **2008,** *36*(suppl 1), D901-D906.

[2] Chen, X.; Ji, Z.L.; Chen, Y.Z. *Nucleic Acids Res.,* **2002,** *30*(1), 412-415.

[3] Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K.F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. *Nucleic Acids Res.,* **2006,** *34*(suppl 1), D354-D357.

[4] Günther, S.; Kuhn, M.; Dunkel, M.; Campillos, M.; Senger, C.; Petsalaki, E.; Ahmed, J.; Urdiales, E.G.; Gewiess, A.; Jensen, L.J. *Nucleic Acids Res.,* **2008,** *36*(suppl 1), D919-D922.

[5] (a) Chen, X.; Yan, C.C.; Zhang, X.; Zhang, X.; Dai, F.; Yin, J.; Zhang, Y. *Briefings in bioinformatics* **2015**, bbv066; (b) Chen, X.; Liu, M.-X.; Yan, G.-Y. *Mol. BioSyst.,* **2012,** *8*(7), 1970-1978.

[6] Cheng, F.; Liu, C.; Jiang, J.; Lu, W.; Li, W.; Liu, G.; Zhou, W.; Huang, J.; Tang, Y. *PLoS Comput. Biol.,* **2012,** *8*(5), e1002503.

[7] Fakhraei, S.; Huang, B.; Raschid, L.; Getoor, L. Network-Based Drug-Target Interaction Prediction with Probabilistic Soft Logic. *IEEE/ACM Trans. Comput. Biol. Bioinform.,* **2014,** *11*(5), 775-787.

[8] Prado-Prado, F.; García-Mera, X.; Escobar, M.; Sobarzo-Sánchez, E.; Yañez, M.; Riera-Fernandez, P.; González-Díaz, H. 2D MI-DRAGON: A new predictor for protein-ligands interactions and theoretic-experimental studies of US FDA drug-target network, oxoisoaporphine inhibitors for MAO-A and human parasite proteins. *Eur. J. Med. Chem.,* **2011,** *46*(12), 5838-5851.

[9] Chen, L.; Lu, J.; Luo, X.; Feng, K.-Y. Prediction of drug target groups based on chemical-chemical similarities and chemical-chemical/protein connections. *Biochim. Biophys. Acta,* **2014,** *1844*(1), 207-213.

[10] 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 Res.,* **2004,** *32*(suppl 1), D431-D433.

[11] 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.

[12] (a) Li, X.; Zhao, Y.; Tian, B.; Jamaluddin, M.; Mitra, A.; Yang, J.; Rowicka, M.; Brasier, A.R.; Kudlicki, A. Modulation of gene expression regulated by the transcription factor NF-κB/RelA. *J. Biol. Chem.,* **2014,** *289*(17), 11927-11944; (b) Li, X.; Zhu, M.; Brasier, A.R.; Kudlicki, A.S. Inferring genome-wide functional modulatory network: A case study on NF-κB/RelA transcription factor. *J. Comput. Biol.,* **2015,** *22*(4), 300-312; (c) Yang, J.; Zhao, Y.; Kalita, M.; Li, X.; Jamaluddin, M.; Tian, B.; Edeh, C.B.; Wiktorowicz, J.E.; Kudlicki, A.; Brasier, A.R. 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. *Mol. Cell. Proteomics,* **2015,** *14*(10), 2701-2721.

[13] (a) Shen, J.; Cheng, F.; Xu, Y.; Li, W.; Tang, Y. Estimation of ADME properties with substructure pattern recognition. *J. Chem. Inf. Model.,* **2010,** *50*(6), 1034-1041; (b) Mannhold, R.; Kubinyi, H.; Folkers, G.; Jahnke, W.; Erlanson, D.A. *Fragment-based approaches in drug discovery*. John Wiley & Sons: **2006**; Vol. *34*.

[14] Jaakkola, T.; Diekhans, M.; Haussler, D. In: *Using the Fisher kernel method to detect remote protein homologies*, ISMB, **1999**; pp. 149-158.

[15] (a) Leslie, C.S.; Eskin, E.; Cohen, A.; Weston, J.; Noble, W.S. Mismatch string kernels for discriminative protein classification. *Bioinformatics* **2004,** *20*(4), 467-476; (b) Leslie, C.S.; Eskin, E.; Noble, W.S. In: *The spectrum kernel: A string kernel for SVM protein classification*, Pacific symposium on biocomputing, World Scientific: **2002**; pp. 566-575.

[16] Yu, X.; Zheng, X.; Liu, T.; Dou, Y.; Wang, J. Predicting subcellular location of apoptosis proteins with pseudo amino acid composition: Approach from amino acid substitution matrix and auto covariance transformation. *Amino Acids,* **2012,** *42*(5), 1619-1625.

[17] (a) Shi, M.-G.; Xia, J.-F.; Li, X.-L.; Huang, D.-S. Predicting protein-protein interactions from sequence using correlation coefficient and high-quality interaction dataset. *Amino Acids,* **2010,** *38*(3), 891-899; (b) Xia, J.-F.; Wu, M.; You, Z.-H.; Zhao, X.-M.; Li, X.-L. Prediction of beta-hairpins in proteins using physicochemical properties and structure information. *Protein Pept. Lett.,* **2010,** *17*(9), 1123-1128.

[18] Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Machine Learning,* **2006,** *63*(1), 3-42.

[19] (a) Maldonado, A.G.; Doucet, J.; Petitjean, M.; Fan, B.-T. Molecular similarity and diversity in chemoinformatics: From

theory to applications. *Mol. Divers.,* **2006,** *10*(1), 39-79; (b) Yang, X.G.; Chen, D.; Wang, M.; Xue, Y.; Chen, Y.Z. Prediction of antibacterial compounds by machine learning approaches. *J. Comput. Chem.,* **2009,** *30*(8), 1202-1211.

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

[21] Gönen, M. Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics,* **2012,** *28*(18), 2304-2310.

[22] Chen, H.; Zhang, Z.; Keskin, O. A semi-supervised method for drug-target interaction prediction with consistency in networks. *PLoS One,* **2013,** *8*(5), e62975.

[23] Gaulton, A.; Bellis, L.J.; Bento, A.P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.,* **2012,** *40*(D1), D1100-D1107.