Neurocomputing 228 (2017) 256–262

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

DrugRPE: Random projection ensemble approach to drug-target interaction prediction

Jun Zhang[a], Muchun Zhu[a], Peng Chen[a,*], Bing Wang[b,*]

[a] Institute of Health Sciences, College of Electrical Engineering and Automation, Anhui University, Hefei, Anhui 230601, China
[b] The Advanced Research Institute of Intelligent Sensing Network, School of Electronics and Information Engineering, The Key Laboratory of Embedded System and Service Computing, Tongji University, Shanghai 201804, China

ARTICLE INFO

2015 MSC:
00-01
99-00

Keywords:
Random projection
Drug-target interaction
REPTree
Ensemble system

ABSTRACT

Drug-target interaction is key in drug discovery. Since the determination of drug-target interactions is costly and time-consuming by in vitro experiments, computational method is a complement to determine the interactions. To address the issue, a random projection ensemble approach is proposed. First, drug-compounds are encoded with feature descriptors by software "PaDEL-Descriptor". Second, target proteins are encoded with physiochemical properties of amino acids, where the 34 relatively independent physiochemical properties are extracted from 544 properties in AAindex1 database. Random projection on the vector of drug-target pair with different dimensions can project the original space onto a reduced one and thus yield a transformed vector with a fixed dimension. Several random projections build an ensemble REPTree system. Experimental results show that our method significantly outperforms and runs faster than other state-of-the-art drug-target predictors, on the commonly used drug-target benchmark sets.

1. Introduction

Drug-target interaction is to identify whether a pair of drug and target can be interacted or not. It is a key in the drug discovery for specific disease [1]. Before a drug candidate was synthesized [2,3], several difficulties need to be overcame. The first difficulty is how to find out the drug effects to different people [4–6] and the second one is to trace and elucidate the drug effects along the biological interaction pathways in human beings [7]. Moreover, since drug discovery is costly and time-consuming as well as the number of new drug approvals is quite low per year, computational methods are complement to the drug discovery. Computational methods can be used to identify the sensitivity and toxicity before a drug candidate was approved [2,3], and they can save time and money to a great extent.

Many works have developed different computational methods for analyzing and identifying drug-target interactions. Such methods can be divided into various classes: docking simulations [8,9], literature text mining [10], methods combining chemical structure, genomic sequence and 3D structure information [11,12], kernel-based methods [13], and others [14]. The most commonly used machine learning methods have been widely applied to investigate drug-target interaction problem. Some focused on HIV protease cleavage site prediction [15], identification of GPCR (G protein-coupled receptors) type [16], protein sub-cellular location prediction [17,18], membrane protein type prediction [19], and a series of relevant web-server predictors as summarized in a recent review [20].

Machine learning methods are commonly used in protein interaction field [21–26]. Here we propose a random projection ensemble approach for drug-target interactions based on the REPTree algorithm [27] by using random projection [28,29] to project original data onto a rather smaller space. To encode the input to classifier ensemble, drug-compounds are encoded with feature descriptors by software "PaDEL-Descriptor", while target proteins are encoded with physiochemical properties of amino acids. From 544 properties in AAindex1 database, 34 relatively independent physiochemical properties are extracted. Random projection on the vector of drug-target pair with different dimensions can map the original space into a reduced one and thus yield a transformed vector with a fixed dimension. The protein targets for drugs are divided into enzymes, ion channels, GPCRs, and nuclear receptors in this study, the same as in references [11,12]. Several random projections build an ensemble REPTree system. Experimental results show that our method significantly outperforms and runs faster than other state-of-the-art drug-target predictors, on the commonly used drug-target benchmark sets.

* Corresponding author.
E-mail addresses: junzhang@ahu.edu.cn (J. Zhang), bigeagle@mail.ustc.edu.cn (P. Chen), wangbing@ustc.edu (B. Wang).

http://dx.doi.org/10.1016/j.neucom.2016.10.039
Received 29 December 2015; Received in revised form 4 April 2016; Accepted 24 October 2016
Available online 01 November 2016
0925-2312/

2. Methods

2.1. Feature vector representing a target protein

To encode target protein, AAindex1 database is used which contains 544 amino acid properties [30]. Most of them are relevant, so like as our previous work [31], irrelevant ones with a correlation coefficient (CC) of 0.5 are extracted, as did in AAindex1 itself who presents correlated properties with a CC of 0.8. The CC of each two properties is computed and the number of relevant properties is counted. Ranking the relevant number in descend, a list of properties is obtained. For the top one property, we remove all of the next properties related to the top one. Step by step, each property related to the previous one is removed from the list. Finally, 34 properties are retained, where each two properties have a CC less than 0.5 [31].

For the ith target protein chain, all residues in the whole chain are considered in this work. In order to investigate the evolution of protein residue in terms of physiochemical property, an encoding schema integrating amino acid properties and sequence profile is used to represent the residue. The sequence profile for one residue created by PSI-Blast with default parameters [32] is then multiplied by each amino acid property, where the property for one amino acid is multiplied by the score of the sequence profile for the same amino acid. That is to say, the profile SP^k for residue k and one amino acid property scale, Aap, are both vectors with 1×20 dimensions with the same orders of amino acids. Thereafter, $MSK^k = SP^k \times Aap$ for residue k represents the multiplication of the corresponding sequence profile by the scale, whose jth element $MSK^{k,j} = SP^{k,j} \times Aap^j$, $j = 1,\ldots,20$. The standard deviation of MSK^k, TD^k, is used to represent the kth residue. As a result, the ith target protein is vectorized as $TD = [TD^1,\ldots,TD^k,\ldots,TD^{lenSeq}]^T$, where $lenSeq$ is the length of the target sequence. A similar vector representation can be found in our previous work [31,33,34].

2.2. Feature vector representing a drug candidate

Moreover, in order to encode drug candidate, PaDEL-Descriptor software is used. PaDEL-Descriptor is a software for calculating molecular descriptors and fingerprints which currently calculates different descriptors (1D, 2D descriptors, and 3D descriptors) and 10 types of fingerprints [35]. A molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into an useful number or the result of some standardized experiment [36]. In this work, 1D and 2D descriptors are used, meanwhile salt is removed from a molecule which assumes that the largest fragment is the desired molecule. In addition, aromaticity information is removed and aromaticity is automatically detected in the molecule before the calculation of descriptors. The used 1D and 2D descriptors are listed in the following Table 6.

As a result, 1444 descriptors are used to encode one drug molecule. So the ith drug candidate can be formulated as $D^i = [D_1^i, D_2^i,\ldots,D_{1444}^i]^T$. These 1D and 2D descriptors as well as fingerprints are calculated mainly using The Chemistry Development Kit [35]. These descriptors include atom type electrotopological state descriptors, McGowan volume, molecular linear free energy relation descriptors, ring counts, and count of chemical substructures identified by Laggner [37].

For the ith pair of drug-target, DT^i, whose target is encoded by the AAindex1 property Aap, it can be formulated as a $(1444 + lenSeq)$-D vector given by

$$V^{i,Aap} = [D^i, TD_A^i ap] = [D_1^i, D_2^i,\ldots,D_{1444}^i, TD_1^{i,Aap}, TD^{i,Aap},\ldots,TD_{lenSeq}^{i,Aap}]^T, \quad (1)$$

where $lenSeq$ is the length of the target sequence.

The corresponding target value T^i is 1 or 0, denoting whether the drug-target pair is in interaction or not. Actually, our method expects to learn the relationship between input matrix V^{Aap} and the corresponding target array T, and to try to make its output as close to the target array T as possible, where Aap denotes that the target is encoded based on the irrelevant AAindex1 property Aap.

2.3. Random projection on REPTree

Random projection is a data reduction technique that projects a high dimensional data onto an low-dimensional subspace [38–41]. Given the original data vector, $X \in \Re^{N\times L1}$, the linear random projection is to multiply the original vector by a random matrix $R \in \Re^{L1\times L2}$. The projection

$$X^R = XR \sum_i x_i r_i \quad (2)$$

yields a dimensionality-reduced vector $X^R \in \Re^{N\times L2}$, where x_i is the ith sample of the original data, r_i is the ith column of the random matrix, and $L2 \ll L1$. The matrix R consists of random values and each column has been normalized to unity. In the Eq. (1), each original data sample with dimension L_1 has been replaced by a random, non-orthogonal direction L_2 in the reduced-dimensional space [39]. Therefore, the dimensionality of original data is reduced from $(1444 + lenSeq)$ to a rather small value.

REPTree is a fast tree learner that uses reduced-error pruning [27], based on information gain/variation reduction as the splitting principle, and optimizes for speed by sorting values for numeric attributes. This work adopts the default numFolds parameter of the REPTree (default 3 in WEKA software) that determines the size of pruning set: the data is divided equally into that number of parts and the last one used as an independent test set to estimate the error at each node.

Previous results showed that the generalization error caused by one classifier can be compensated by other classifiers, therefore using tree ensemble can yield significant improvement in prediction accuracy [42]. For the drug-target interaction prediction problem, the ensemble of simple trees votes for the most popular class of drug-target interaction. Given the set of training data $V_{tr}^{k,Aap} = \{(X_i^{R^k,Aap}, Y_i)\}|_{i=1}^N$ in terms of AAindex1 property Aap, after multiplied by the random projection R^k, let the number of training instances be N, and the number of features in the classifier be $L2$. Then the data $V^{k,Aap}$ is generated as an input to a REPTree and thus it builds a classifier $CF_{k,Aap}(x)$, where x is a training instance.

After all of REPTree classifiers with random projection are generated, they vote for the most popular class and thus the prediction of the ensemble is,

$$Pred(X) = majority\ vote\ \{CF_{k,Aap}(x)\}_{Aap=1}^{34}, \quad (3)$$

where x is a query instance.

Results showed that the majority vote with independent classifiers can often make a dramatic improvement [43,44]. Here a pair of drug-target is labelled as interacting if all of the classifiers identified it as positive class 1, otherwise it is identified as a pair without in drug-target interaction Table 1.

3. Materials

3.1. Data sets

We used the drug-target datasets in Ref. [12] for our study. It excluded drug-target pairs that lack experimental information and finally contains a total of 4797 pairs, of which 2719 for enzymes, 1372 for ion channels, 630 for GPCRs, and 86 for nuclear receptors. The lists of the pairs can be found in Ref. [12] and the details can be obtained from KEGG [45]. All these datasets were regarded as the positive ones in this work.

Table 1
The used 1D and 2D descriptors of PaDEL-Descriptor.

Descriptor type	Number*	Descriptor type	Number
Acidic group count	1	Barysz matrix	91
ALOGP	3	APol	1
Aromatic atoms count	1	Aromatic bonds count	1
Atom count	14	Autocorrelation	346
Basic group count	1	BCUT	6
Bond count	10	BPol	1
Burden modified eigenvalues	96	Eccentric connectivity index	1
Carbon types	9	Chi chain	10
Chi cluster	8	Chi path cluster	6
Chi path	32	Constitutional	12
Crippen logP and MR	2	Detour matrix	11
Vertex adjacency information (magnitude)	1	Atom type electrotopological state	489
FMFDescriptor	1	Fragment complexity	1
Hbond acceptor count	4	Hbond donor count	2
Hybridization ratio	1	Information content	42
Kappa shape indices	3	Largest chain	1
Largest Pi system	1	Longest aliphatic chain	1
Mannhold LogP	1	McGowan volume	1
Molecular distance edge	19	Molecular linear free energy relation	6
Path counts	22	Petitjean number	1
Ring count	68	Rotatable bonds count	4
Rule of five	1	Topological	3
Topological charge	21	Vander Waals volume	1
Topological distance matrix	11	Topological polar surface area	1
Walk counts	20	Weight	2
Weighted path	5	Wiener numbers	2
XLogP	1	Zagreb index	1
Extended topochemical atom	43		

*The number of descriptors in each type.

Table 2
Details of the drug-target dataset.

Dataset	Drugs	Targets	Interactions – positive pairs	Negative pairs
Enzymes	419	643	2719	5438
Ion channels	203	198	1372	2744
GPCRs	217	92	620	1240
Nuclear receptors	53	25	86	172
In total	892	958	4797	9588

Table 3
Prediction performance of the REPTree classifier ensemble with majority vote technique, i.e., the ensemble system predicts a drug-target pair to be interacting if all of REPTree classifiers in the ensemble predict it to be interacting.

Dataset	Target type	Rec	Acc	Prec	F1
Training[a]	Enzymes	0.970	0.944	0.876	0.921
	Ion channels	0.986	0.886	0.751	0.853
	GPCRs	0.994	0.892	0.758	0.860
	Nuclear receptors	0.709	0.812	0.722	0.716
Test[b]	Enzymes	0.972	0.900	0.782	0.867
	Ion channels	0.993	0.89	0.755	0.858
	GPCRs	1.000	0.852	0.693	0.818
	Nuclear receptors	0.837	0.911	0.889	0.862

[a] Prediciton on the training dataset $\aleph_{tr}$.
[b] Prediciton on the test dataset $\aleph_{ts}$.

Table 4
Performance comparison of our method under different projection dimensions, $L2$ in Eq. (1)[a].

$L2$	Target type	Rec	Acc	Prec	F1
3	Enzymes	0.972	0.900	0.782	0.867
	Ion channels	0.993	0.890	0.755	0.858
	GPCRs	1.000	0.852	0.693	0.818
	Nuclear receptors	0.837	0.911	0.889	0.862
	Average	0.951	0.888	0.780	0.851
5	Enzymes	0.827	0.868	0.994	0.903
	Ion channels	0.741	0.810	1.000	0.851
	GPCRs	0.384	0.527	0.971	0.550
	Nuclear receptors	0.876	0.609	0.653	0.748
	Average	0.707	0.704	0.905	0.762
10	Enzymes	0.838	0.813	0.908	0.872
	Ion channels	0.530	0.500	0.763	0.625
	GPCRs	0.849	0.461	0.541	0.661
	Nuclear receptors	0.748	0.679	0.815	0.780
	Average	0.741	0.613	0.757	0.735
20	Enzymes	0.771	0.706	0.830	0.799
	Ion channels	0.724	0.595	0.734	0.729
	GPCRs	0.781	0.700	0.815	0.797
	Nuclear receptors	0.477	0.576	0.932	0.631
	Average	0.688	0.644	0.828	0.739
50	Enzymes	0.830	0.687	0.763	0.795
	Ion channels	0.428	0.512	0.883	0.576
	GPCRs	0.465	0.566	0.930	0.620
	Nuclear receptors	0.815	0.710	0.800	0.807
	Average	0.635	0.619	0.844	0.700
100	Enzymes	0.819	0.670	0.751	0.784
	In channels	0.737	0.724	0.882	0.803
	GPCRs	0.535	0.610	0.920	0.676
	Nuclear receptors	0.739	0.694	0.842	0.787
	Average	0.708	0.675	0.485	0.763

[a] Prediction on the test dataset $\aleph_{ts}$.

Table 5
Performance comparison in accuracy of our method with two methods on the same datasets.

Method	Type	enzymes	ion channels	GPCRs	Nuclear receptors
Our method	REPTree	0.900	0.890	0.852	0.911
Ref. [12]	kNN	0.855	0.808	0.785	0.857
*-Drug		0.910[a]	0.873[b]	0.855[c]	0.892[d]
Random predictor		0.489	0.489	0.488	0.488

[a] See Ref. [48] for the iEzy-Drug predictor and its reported success rates.
[b] See Ref. [49] for the iCDI-Drug predictor and its reported success rates.
[c] See Ref. [50] for the iGPCR-Drug predictor and its reported success rates.
[d] See Ref. [51] for the iNR-Drug predictor and its reported success rates.

Algorithm 1. Prediction of drug-target interaction by random projection:

Require: Training drug-target set $\aleph_{tr}$ and test set $\aleph_{ts}$ by applying 10-fold cross-validation technique to V of each type of drug-target interactions
Ensure: Prediction Acc
for running times $1 \sim 100$ **do**
Obtain a random projection R^k;
for AAindex1 property $Aap = 1 \sim 34$ **do**
Run the REPTree classifier on $\aleph_{tr}^{k,Aap}$ by cross-validation;
Obtain the prediction $Pred(X)^{k,Aap}$;
end for

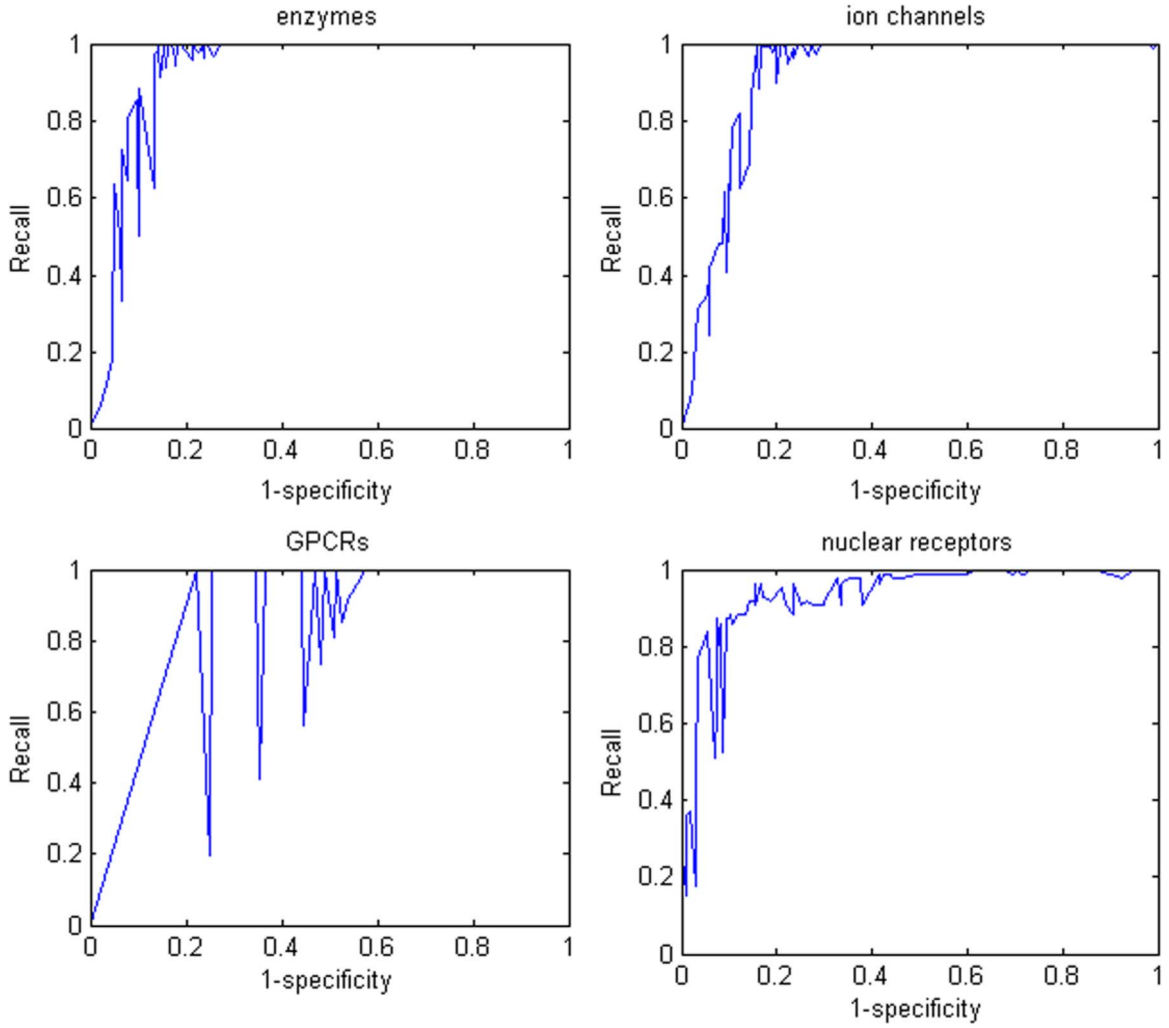


Fig. 1. ROC performance of our method for classes of enzymes, ion channels, GPCRs, and nuclear receptors.

Majority vote $Pred(X)^k$ to the predictions in terms of AAindex1 properties;
end for
Sort $Pred(X)$ and obtain the random projections with top accuracy;
Apply the top random projections to the test set $\aleph_{ts}$;
Calculate the performance on $\aleph_{ts}$.

The corresponding negative datasets were obtained by the same selection steps in Ref. [12]. The selection steps are: (1) separate the pairs in the above positive dataset into single drugs and proteins; (2) re-couple these singles into pairs in a way that none of them occurs in the corresponding positive dataset; (3) randomly picked the negative pairs thus formed until they reached the number two times as many as the positive pairs [12]. The drug-target interaction pairs are divided in terms of protein target family. The total of 4797 drug-target pairs are grouped into four families: enzymes, GPCRs, icon channels, and nuclear receptors. Finally, the four datasets contain 8157, 4116, 1860 and 258 pairs for enzymes, ion channels, GPCRs and nuclear receptors, respectively. Table 2 lists the details of the four datasets.

3.2. Drug-target interaction prediction evaluation

In this work we adopted four evaluation measures to show the ability of our model objectively, criteria of Recall (Rec), Precision (Prec), F-measure (F1), and Accuracy (Acc) [33,46,47]. They are defined as follows:

$$Rec = \frac{TP}{TP + FN} Prec = \frac{TP}{TP + FP} Acc = \frac{TP + TN}{TP + FN + FP + FN} F1 = 2 \times \frac{Prec \times Sen}{Prec + Sen}, \quad (4)$$

where TP (True Positive) is the number of correctly predicted drug-target pairs; FP (False Positive) is the number of false positives (incorrectly over predicted non drug-target pairs); TN (True Negative) is the number of correctly predicted non drug-target pairs; and FN (False Negative) is false negative, i.e., incorrectly under predicted drug-target pairs.

4. Results

4.1. Performance of drug-target interaction prediction

In this work, drug-target interactions are grouped into four types, enzymes, ion channels, GPCRs, and nuclear receptors. For each type, the proposed method is applied to it (see Algorithm 1) and the results are shown in the blow. The dataset of each type of drug-target interactions is divided into training data set $\aleph_{tr}$ and test one $\aleph_{ts}$ by 10-fold cross-validation. That is to say, the dataset is divided into 10 subsets with roughly the same number of instances, and one subset is regarded as the test set while the others are grouped as training set. The test subset is selected one-by-one and finally all of the instances are tested. Then different random projections are used to project the original dataset onto a rather lower space, in this work 5 dimension-

Table 6
The used 34 properties of AAindex1 database.

Accession	Data description	Type
ARGP820101	Hydrophobicity index (Argos et al., 1982)	
ARGP820102	Signal sequence helical potential (Argos et al., 1982)	
ARGP820103	Membrane-buried preference parameters (Argos et al., 1982)	Hydrophobicity
BULH740101	Transfer free energy to surface (Bull-Breese, 1974)	
BULH740102	Apparent partial specific volume (Bull-Breese, 1974)	
BIGC670101	Residue volume (Bigelow, 1967)	
BIOV880101	Information value for accessibility; average fraction 35% (Biou et al., 1988)	
BIOV880102	Information value for accessibility; average fraction 23% (Biou et al., 1988)	Residue volume
CHAM820101	Polarizability parameter (Charton-Charton, 1982)	
CHAM820102	Free energy of solution in water, kcal/mole (Charton-Charton, 1982)	
CHOC750101	Average volume of buried residue (Chothia, 1975)	
CHOC760101	Residue accessible surface area in tripeptide (Chothia, 1976)	
CHOC760102	Residue accessible surface area in folded protein (Chothia, 1976)	
BHAR880101	Average flexibility indices (Bhaskaran-Ponnuswamy, 1988)	
BROC820101	Retention coefficient in TFA (Browne et al., 1982)	Flexibility
BROC820102	Retention coefficient in HFBA (Browne et al., 1982)	
BEGF750101	Conformational parameter of inner helix (Beghin-Dirkx, 1975)	
BEGF750102	Conformational parameter of beta-structure (Beghin-Dirkx, 1975)	
BEGF750103	Conformational parameter of beta-turn (Beghin-Dirkx, 1975)	
BURA740101	Normalized frequency of alpha-helix (Burgess et al., 1974)	
BURA740102	Normalized frequency of extended structure (Burgess et al., 1974)	Secondary structure
CHAM830101	The Chou-Fasman parameter of the coil conformation (Charton-Charton, 1983)	
CHAM830102	A parameter defined from the residuals obtained from the best correlation of the Chou-Fasman parameter of beta-sheet (Charton-Charton, 1983)	
CHAM830103	The number of atoms in the side chain labelled 1+1 (Charton-Charton, 1983)	
CHAM830104	The number of atoms in the side chain labelled 2+1 (Charton-Charton, 1983)	
CHAM830105	The number of atoms in the side chain labelled 3+1 (Charton-Charton, 1983)	
CHAM830106	The number of bonds in the longest chain (Charton-Charton, 1983)	
ANDN920101	alpha-CH chemical shifts (Andersen et al., 1992)	
BUNA790101	alpha-NH chemical shifts (Bundi-Wuthrich, 1979)	
BUNA790102	alpha-CH chemical shifts (Bundi-Wuthrich, 1979)	
BUNA790103	Spin-spin coupling constants 3JHalpha-NH (Bundi-Wuthrich, 1979)	Steric parameter
CHAM810101	Steric parameter (Charton, 1981)	
CHAM830107	A parameter of charge transfer capability (Charton-Charton, 1983)	
CHAM830108	A parameter of charge transfer donor capability (Charton-Charton, 1983)	

ality space projected. For achieving better random projection, the training data set $\aleph_{tr}$ is divided into training subset $\aleph_{tr}^{sub}$ and test subset $\aleph_{ts}^{sub}$ by 10-fold cross-validation. Running the REPTree classifier by the random projection technique, predictions on the test subset $\aleph_{ts}^{sub}$ by the training subset $\aleph_{tr}^{sub}$ are obtained. Only random projections yielding top performance are retained. Running REPTree classifier, given the top random projection, on the training data set $\aleph_{tr}$ and test one $\aleph_{ts}$ yields the final predictions.

In details, there are 34 random projections R^k on the original data matrix in terms of the 34 independent AAindex1 properties. The ensemble of the 34 classifiers by random projections yields prediction for the training data subset $\aleph_{tr}^{sub}$ and the prediction for the test data subset $\aleph_{ts}^{sub}$. The 34 random projections are retained if the prediction accuracy is larger than 0.75 for $\aleph_{tr}$. Repeating the classifier ensemble by random projections R^k, several top predictions are obtained by random projections $R^k(k = 1 \sim K)$, on $\aleph_{tr}$ and $\aleph_{ts}$. Combining the K predictions yields final prediction. Table 3 shows the performance comparison of the ensembles for the four protein target classes. Here the dimensionality of the original data is reduced from $(1444 + lenSeq)$ to 5. For drugs, they are encoded into vectors with fixed length, 1444, while for protein targets with different sequence length, they can be encoded into vectors with different sequence length *lenSeq*. The longest sequence length is used for the original space dimensionality *maxLenSeq* of random projections. Target sequence with shorter length *lenSeq* is encoded as a subspace $\Re^{lenSeq}$ in the space $\Re^{maxLenSeq}$, i.e., $\Re^{lenSeq} \in \Re^{maxLenSeq}$. From the Table 3, it can be seen that the ensemble system tested on nuclear receptors class performs better than that on other classes. It yields an accuracy of 0.911 and a precision of 0.889 at a recall of 0.837.

4.2. Performance with respect to different projection dimension

We adopted random projection technique to search for the optimized feature space and further applied it to obtain the prediction of drug-target interactions. To do that, random projections with different projection dimensions were investigated. Table 4 lists performance comparison according to different $L2$ in Eq. (1). From the Table 4, experiments with projection dimension $L2 = 3$ performs the best than others and yields an accuracy of 0.888. It seems that experiments with smaller projection dimension perform better than those with larger projection dimension.

4.3. Comparison with other methods

We also compared our method with other two methods: the work in Ref. [12] and the random predictor on the same datasets. Table 5 shows the performance comparison in accuracy of our method with other two methods. The random predictor is implemented here and ran 100 times. The average performance is appended at the bottom of the table. Our method yields accuracies of 0.900, 0.89, 0.852, and 0.911 for classes of enzymes, ion channels, GPCRs, and nuclear receptors, respectively. Our method achieves *Acc* improvements of 4.5–8.2% than the work [12] for the four classes. In addition, our method performs comparatively to the four web-servers: iEzy-Drug, iCDI-Drug, iGPCR-Drug and iNR-Drug. Moreover results showed that our method outperforms the random predictor by 2 times of *Acc* score.

The performance of ensemble classifier with majority vote is illustrated in Fig. 1. Although it is much difficult to identify drug-target pairs in GPCRs class, our method yields good predictions.

4.4. Description of 34 properties

Since 34 independent properties of amino acids are extracted from AAindex1 database and applied in the drug-target prediction, the details of the 34 properties are listed in the Table 6. Data description of each property of amino acids is shown in the Table. Some properties are for hydrophobicity index, some for residue volume, some for flexibility index, some for secondary structure, and some for atom-atom interactions. These properties of amino acids are important for encoding protein sequence in that they represent protein sequences by different environmental features. The encoding schema aims to apply

various statistic features to recover real interactions among amino acid residues.

5. Conclusions

This paper proposes an ensemble of REPTree classifiers by random projection to identify drug-target interactions. For each independent AAindex1 property, the original encoders for drug-target interaction transformed by different random projections are input into a REPTree classifier. There are 34 REPTree classifiers with respect to AAindex1 property. The ensemble of these REPTree classifiers can yield good prediction on drug-target interactions. Therefore, our method is simple for only statistical amino acid properties are applied. Moreover, the dimensionality reduction of random projection is adopted here to reduce the original encoder space. More importantly, the random projection technique can handle protein chains with different numbers of amino acids and get unified encoder space. Actually, the random projection technique provides a useful mechanism such that it reduces the high dimensional original data and makes the data more diverse and thus, the method yields a good prediction on drug-target interactions. Results show that our method outperforms other state-of-the-art methods in the prediction of drug-target interactions.

Acknowledgments

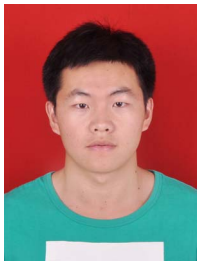
This work was supported by the National Natural Science Foundation of China (Nos. 61672035, 61300058, 61472282 and 61271098).

References

[1] J. Knowles, G. Gromo, A guide to drug discovery, target selection in drug discovery, Nat. Rev. Drug Discov. 2 (1) (2003) 63–69. http://dx.doi.org/10.1038/nrd986.
[2] Johnson, Wolfgang, Predicting human safety screening and computational approaches, Drug Discov. Today 5 (10) (2000) 445–454.
[3] S. Sirois, G. Hatzakis, D. Wei, Q. Du, K.-C. Chou, Assessment of chemical libraries for their druggability, Comput. Biol. Chem. 29 (1) (2005) 55–67. http://dx.doi.org/10.1016/j.compbiolchem.2004.11.003.
[4] A.J. Wood, W.E. Evans, H.L. McLeod, Pharmacogenomics drug disposition, drug targets, and side effects, New Engl. J. Med. 348 (6) (2003) 538–549.
[5] J.-F. Wang, D.-Q. Wei, C. Chen, Y. Li, K.-C. Chou, Molecular modeling of two cyp2c19 SNPs and its implications for personalized drug design, Protein Pept. Lett. 15 (1) (2008) 27–32.
[6] J.-F. Wang, D.-Q. Wei, K.-C. Chou, Pharmacogenomics and personalized use of drugs, Curr. Top. Med. Chem. 8 (18) (2008) 1573–1579.
[7] S. Mizutani, E. Pauwels, V. Stoven, S. Goto, Y. Yamanishi, Relating drug-protein interaction network with drug side effects, Bioinformatics 28 (18) (2012) i522–i528.
[8] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, A fast flexible docking method using an incremental construction algorithm, J. Mol. Biol. 261 (3) (1996) 470–489. http://dx.doi.org/10.1006/jmbi.1996.0477.
[9] A.C. Cheng, R.G. Coleman, K.T. Smyth, Q. Cao, P. Soulard, D.R. Caffrey, A.C. Salzberg, E.S. Huang, Structure-based maximal affinity model predicts small-molecule druggability, Nat. Biotechnol. 25 (1) (2007) 71–75. http://dx.doi.org/10.1038/nbt1273.
[10] S. Zhu, Y. Okuno, G. Tsujimoto, H. Mamitsuka, A probabilistic model for mining implicit chemical compound-generelations from literature, Bioinformatics 21 (Suppl. 2) (2005) ii245–ii251.
[11] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, M. Kanehisa, Prediction of drug-target interaction networks from the integration of chemical and genomic spaces, Bioinformatics 24 (13) (2008) i232–i240. http://dx.doi.org/10.1093/bioinformatics/btn162.
[12] Z. He, J. Zhang, X.-H. Shi, L.-L. Hu, X. Kong, Y.-D. Cai, K.-C. Chou, Predicting drug-target interaction networks based on functional groups and biological features, PLoS One 5 (3) (2010) e9603. http://dx.doi.org/10.1371/journal.pone.0009603.
[13] Y.-C. Wang, C.-H. Zhang, N.-Y. Deng, Y. Wang, Kernel-based data fusion improves the drug-protein interaction prediction, Comput. Biol. Chem. 35 (6) (2011) 353–362.
[14] N. Nagamine, T. Shirakawa, Y. Minato, K. Torii, H. Kobayashi, M. Imoto, Y. Sakakibara, Integrating statistical predictions and experimental verifications for enhancing protein-chemical interaction predictions in virtual screening, PLoS Comput. Biol. 5 (6) (2009) e1000397.
[15] K.-C. Chou, A vectorized sequence-coupling model for predicting hiv protease cleavage sites in proteins, J. Biol. Chem. 268 (23) (1993) 16938–16948.
[16] X. Xiao, P. Wang, K.-C. Chou, Gpcr-ca: a cellular automaton image approach for predicting g-protein-coupled receptor functional classes, J. Comput. Chem. 30 (9) (2009) 1414–1423.
[17] K.-C. Chou, H.-B. Shen, Cell-ploc: a package of web servers for predicting subcellular localization of proteins in various organisms, Nat. Protoc. 3 (2) (2008) 153–162.
[18] X. Xiao, J.-L. Min, P. Wang, K.-C. Chou, Predict drug-protein interaction in cellular networking, Curr. Top. Med. Chem. 13 (14) (2013) 1707–1712.
[19] K.-C. Chou, D.W. Elrod, Prediction of membrane protein types and subcellular locations, Proteins: Struct. Funct. Bioinform. 34 (1) (1999) 137–153.
[20] K.-C. Chou, H.-B. Shen, et al., Review: recent advances in developing web-servers for predicting protein attributes, Nat. Sci. 1 (02) (2009) 63.
[21] L. Zhu, Z.-H. You, D.-S. Huang, B. Wang, t-lse: a novel robust geometric approach for modeling protein-protein interaction networks, PLoS One 8 (4) (2013) e58368. http://dx.doi.org/10.1371/journal.pone.0058368.
[22] D.-S. Huang, H.-J. Yu, Normalized feature vectors: a novel alignment-free sequence comparison method based on the numbers of adjacent amino acids, IEEE/ACM Trans. Comput. Biol. Bioinform. 10 (2) (2013) 457–467. http://dx.doi.org/10.1109/TCBB.2013.10.
[23] D.-S. Huang, L. Zhang, K. Han, S. Deng, K. Yang, H. Zhang, Prediction of protein-protein interactions based on protein-protein correlation using least squares regression, Curr. Protein Pept. Sci. 15 (6) (2014) 553–560.
[24] B. Wang, D.-S. Huang, C. Jiang, A new strategy for protein interface identification using manifold learning method, IEEE Trans. Nanobiosci. 13 (2) (2014) 118–123. http://dx.doi.org/10.1109/TNB.2014.2316997.
[25] L. Zhu, S.-P. Deng, D.-S. Huang, A two-stage geometric method for pruning unreliable links in protein-protein networks, IEEE Trans. Nanobiosci. 14 (5) (2015) 528–534. http://dx.doi.org/10.1109/TNB.2015.2420754.
[26] S.-P. Deng, L. Zhu, D.-S. Huang, Mining the bladder cancer-associated genes by an integrated strategy for the construction and analysis of differential co-expression networks, BMC Genom. 16 (Suppl. 3) (2015) S4. http://dx.doi.org/10.1186/1471-2164-16-S3-S4.
[27] F. Esposito, D. Malerba, G. Semeraro, V. Tamma, The Effects of Pruning Methods on the Predictive Accuracy of Induced Decision Trees, 1999.
[28] X.Z. Fern, C.E. Brodley, Random projection for high dimensional data clustering: a cluster ensemble approach, in: ICML, vol. 3, 2003, pp. 186–193.
[29] A. Schclar, L. Rokach, Random projection ensemble classifiers, in: Enterprise Information Systems, Springer, 2009, pp. 309–316.
[30] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, Aaindex: amino acid index database, progress report 2008, Nucleic Acids Res. 36 (2008) D202–D205. http://dx.doi.org/10.1093/nar/gkm998 (Database issue).
[31] P. Chen, J. Li, L. Wong, H. Kuwahara, J.Z. Huang, X. Gao, Accurate prediction of hot spot residues through physicochemical characteristics of amino acid sequences, Proteins 81 (8) (2013) 1351–1362. http://dx.doi.org/10.1002/prot.24278.
[32] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped blast and psi-blast: a new generation of protein database search programs, Nucleic Acids Res. 25 (17) (1997) 3389–3402.
[33] P. Chen, J. Li, Sequence-based identification of interface residues by an integrative profile combining hydrophobic and evolutionary information, BMC Bioinform. 11 (2010) 402. http://dx.doi.org/10.1186/1471-2105-11-402.
[34] P. Chen, L. Wong, J. Li, Detection of outlier residues for improving interface prediction in protein heterocomplexes, IEEE/ACM Trans. Comput. Biol. Bioinform. 9 (4) (2012) 1155–1165. http://dx.doi.org/10.1109/TCBB.2012.58.
[35] C.W. Yap, Padel-descriptor: an open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (7) (2011) 1466–1474. http://dx.doi.org/10.1002/jcc.21707.
[36] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, vol. 11, John Wiley & Sons, 2008.
[37] J. Klekota, F.P. Roth, Chemical substructures that enrich for biological activity, Bioinformatics 24 (21) (2008) 2518–2525. http://dx.doi.org/10.1093/bioinformatics/btn479.
[38] C.H. Papadimitriou, P. Raghavan, H. Tamaki, S. Vempala, Latent Semantic Indexing: a Probabilistic Analysis, 1998.
[39] S. Kaski, Dimensionality reduction by random mapping: fast similarity computation for clustering, in: Proceedings of the Neural Networks, IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on, vol. 1, 1998, pp. 413–418 http://dx.doi.org/10.1109/IJCNN.1998.682302 ⟨http://ieeexplore.ieee.org/stamp/stamp.jsp?Arnumber=682302⟩.
[40] Z. Wang, W. Jie, S. Chen, D. Gao, Random projection ensemble learning with multiple empirical kernels, Knowl. Based Syst. 37 (2013) 388–393.
[41] A. Ahmad, G. Brown, Random projection random discretization ensembles–ensembles of linear multivariate decision trees, IEEE Trans. Knowl. Data Eng. 26 (5) (2014) 1225–1239. http://dx.doi.org/10.1109/TKDE.2013.134 ⟨http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6574846⟩.
[42] P. Chen, J.Z. Huang, X. Gao, LigandRFs: random forest ensemble to identify ligand-binding residues from sequence information alone, BMC Bioinform. 15 (Suppl. 15) (2014) S4. http://dx.doi.org/10.1186/1471-2105-15-S15-S4.
[43] P. Chen, J. Li, Prediction of protein long-range contacts using an ensemble of genetic algorithm classifiers with sequence profile centers, BMC Struct. Biol. 10 (Suppl. 1) (2010) S2. http://dx.doi.org/10.1186/1472-6807-10-S1-S2.
[44] L.I. Kuncheva, C.J. Whitaker, R.P.W. Duin, Limits on the Majority Vote Accuracy in Classifier Fusion, 2003.
[45] M. Kanehisa, The Kegg Database, Novartis Found. Symp., vol. 247, 2002 91–101; discussion 101–3, 119–28, 244–52.
[46] P. Chen, C. Liu, L. Burge, J. Li, M. Mohammad, W. Southerland, C. Gloster, B. Wang, DomSVR: domain boundary prediction with support vector regression

from sequence information alone, Amino Acids 39 (3) (2010) 713–726. http://dx.doi.org/10.1007/s00726-010-0506-6.
[47] B. Wang, P. Chen, D.-S. Huang, J.-j. Li, T.-M. Lok, M.R. Lyu, Predicting protein interaction sites from residue spatial sequence profile and evolution rate, FEBS Lett. 580 (2) (2006) 380–384. http://dx.doi.org/10.1016/j.febslet.2005.11.081.
[48] J.-L. Min, X. Xiao, K.-C. Chou, Iezy-drug: a web server for identifying the interaction between enzymes and drugs in cellular networking, Biomed. Res. Int. 2013 (2013) 701317. http://dx.doi.org/10.1155/2013/701317.
[49] X. Xiao, J.-L. Min, P. Wang, K.-C. Chou, iCDI-PseFpt: identify the channel-drug interaction in cellular networking with pseaac and molecular fingerprints, J. Theor. Biol. 337 (2013) 71–79. http://dx.doi.org/10.1016/j.jtbi.2013.08.013.
[50] X. Xiao, J.-L. Min, P. Wang, K.-C. Chou, igpcr-drug: a web server for predicting interaction between gpcrs and drugs in cellular networking, PLoS One 8 (8) (2013) e72234. http://dx.doi.org/10.1371/journal.pone.0072234.
[51] Y.-N. Fan, X. Xiao, J.-L. Min, K.-C. Chou, Inr-drug predicting the interaction of drugs with nuclear receptors in cellular networking, Int. J. Mol. Sci. 15 (3) (2014) 4915–4937. http://dx.doi.org/10.3390/ijms15034915.

Peng Chen specialises in machine learning and data mining with applications to bioinformatics, drug discovery, computer vision, etc. He has published about 40 high quality referred papers in international conferences and journals. He is an Professor in the Institute of Health Sciences, Anhui University, Hefei, China. He received his Bachelor degree from Electronic Engineering Institute, PLA, Master degree from Kunming University of Science and Technology, and Ph.D. degree from University of Science and Technology of China. Prior to joining Anhui University, he served in City University of Hong Kong (2006, as senior research associate), Howard University, USA (2008–2009, as Postdoc Fellow), Nanyang Technological University, Singapore (2009–2010, as Research fellow), and King Abdullah University of Science and Technology (KAUST), Saudi Arabia (2012–2014, as Postdoc Fellow). From 2011–2013, he was an Associate Professor in Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, China.



Muchun Zhu is a master student in the Institute of Health Sciences, Anhui University. His research fields are bioinformatics and software applications.

Jun Zhang was born in Anhui Province, Chin, in 1971. He received M.S. degree in Pattern Recognition & Intelligent System in 2004, from Institute of Intelligent Machines, Chinese Academy of Sciences. He received the Ph.D. degree from University of Science and Technology of China, Hefei, China in 2007. Currently, Dr. Zhang is associate professor in the School of Electrical Engineering and Automation, Anhui University, China. His research interests focus on deep learning, ensemble learning and cheminformatics.

Bing Wang received the B.S. and M.S degree from Hefei University of Technology, Hefei, China in 1998 and 2004 respectively. He received the Ph.D. degree from University of Science and Technology of China, Hefei, China in 2006. Currently, Dr. Wang is serving as a research professor in the School of Electronics and Information Engineering, Tongji University, Shanghai, China. His research interests mainly focus on machine learning, computational biology and cheminformatics.