Journal of Pharmacological and Toxicological Methods 78 (2016) 42–51

Contents lists available at ScienceDirect

Journal of Pharmacological and Toxicological Methods

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

Drug–target interaction prediction from PSSM based evolutionary information

Zaynab Mousavian [a], Sahand Khakabimamaghani [c], Kaveh Kavousi [b], Ali Masoudi-Nejad [a,*]

[a] Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran
[b] Laboratory of Biological Complex Systems and Bioinformatics (CBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran
[c] School of Computing Science, Simon Fraser University, Burnaby, Canada

ARTICLE INFO

Article history:
Received 3 June 2015
Received in revised form 1 November 2015
Accepted 8 November 2015
Available online 22 November 2015

Keywords:
Drug–target interaction
Learning
Classification
Position Specific Scoring Matrix (PSSM)

ABSTRACT

The labor-intensive and expensive experimental process of drug–target interaction prediction has motivated many researchers to focus on in silico prediction, which leads to the helpful information in supporting the experimental interaction data. Therefore, they have proposed several computational approaches for discovering new drug–target interactions. Several learning-based methods have been increasingly developed which can be categorized into two main groups: similarity-based and feature-based. In this paper, we firstly use the bi-gram features extracted from the Position Specific Scoring Matrix (PSSM) of proteins in predicting drug–target interactions. Our results demonstrate the high-confidence prediction ability of the Bigram-PSSM model in terms of several performance indicators specifically for enzymes and ion channels. Moreover, we investigate the impact of negative selection strategy on the performance of the prediction, which is not widely taken into account in the other relevant studies. This is important, as the number of non-interacting drug–target pairs are usually extremely large in comparison with the number of interacting ones in existing drug–target interaction data. An interesting observation is that different levels of performance reduction have been attained for four datasets when we change the sampling method from the random sampling to the balanced sampling.

1. Introduction

In the genomic drug discovery, detecting drug–target interactions is an important area of research, which can result in the identification of new drugs or novel targets for the current drugs (Masoudi-Nejad, Mousavian, & Bozorgmehr, 2013). Because a large amount of expense should be spent for the experimental prediction of drug–target interaction, the in silico prediction can be used as an alternative approach which provides helpful information in supporting the experimental interaction data. There are four types of target proteins which can be used in the pharmaceutical applications such as Enzymes, Ion Channels, G-protein Coupled Receptors (GPCR) and Nuclear Receptors, whose functions can be affected by interacting with compounds. There are several available databases including KEGG BRITE (Kanehisa et al., 2006), DrugBank (Wishart et al., 2008), GLIDA (Okuno et al., 2008), SuperTarget and Matador (Günther et al., 2008), BRENDA (Schomburg et al., 2004) and ChEMBL (Warr, 2009), which contain information about experimentally validated drug–target interactions and other related data such as genomic and chemical data. Since a rather small number of experimental drug–target interaction data has been recorded in public databases, a variety of computational methods have been proposed in the literatures for discovering new drug–target interactions based on the available interaction data.

Generally, the proposed computational methods can be classified into two main groups: the ligand-based methods and the target-based methods. In the ligand-based methods like QSAR (Quantitative Structure Activity Relationship), the learning-based methods have been developed for predicting interaction between a given ligand–target pair by comparing a given ligand to the known ligands of a given target protein (Butina, Segall, & Frankcombe, 2002; Byvatov et al., 2003). So, this approach is not applicable to proteins for which the small number of ligands is known. On the other hand, target-based methods or docking simulation is only applicable to proteins for which the 3D structures are available (Cheng et al., 2007; Donald, 2011; Morris et al., 2009). Because most of the membrane proteins such as Ion channels and G-Protein Coupled Receptors (GPCRs) have no available 3D structure in the public databases, hence the docking simulation method can't be applied on them.

To tackle the problems of these methods, several chemogenomic methods have been increasingly developed to integrate both the genomic space of target proteins and the chemical space of compounds into a unified space for detecting new drug–target pairs (Chen, Liu, & Yan, 2012; Cheng et al., 2012). Most of the existing approaches are learning-based methods which can be categorized into two main groups:

* Corresponding author.
E-mail addresses: zmousavian@ut.ac.ir (Z. Mousavian), amasoudin@ibb.ut.ac.ir (A. Masoudi-Nejad).
URL: http://LBB.ut.ac.ir (A. Masoudi-Nejad).

http://dx.doi.org/10.1016/j.vascn.2015.11.002
1056-8719/

similarity-based and feature-based (Mousavian & Masoudi-Nejad, 2014). The similarity-based methods have constructed the predictive models based on the similarity matrices among drugs and among targets (Bleakley & Yamanishi, 2009; Yamanishi et al., 2008). Furthermore, the feature-based methods have used various sets of discriminative features for representing drugs and targets (He et al., 2010; Tabei et al., 2012), so that these features have been used in building a learning model which can differentiate between the interacting and the non-interacting pairs. A number of machine learning algorithms such as support vector machine (Tabei et al., 2012; Wang et al., 2010), nearest neighbor (He et al., 2010) and random forest (Yu et al., 2012) have been used in predicting drug–target interaction. Despite the many advantages of the proposed methods, the imbalanced drug–target dataset can lead to a strong bias in the learning algorithms, so, the existing bias yields the over optimistic prediction results which are not realistic.

In this paper, we firstly use the bi-gram features extracted from the Position Specific Scoring Matrix (PSSM) of proteins, introduced recently by Sharma et al. (2013), in predicting drug–target interactions and compare the prediction ability of the Bigram-PSSM features with that of the Pseudo Amino Acid Composition (PAAC) features, the most widely used protein representation proposed by Chou, (2001). Because the PAAC features are defined with respect to a set of k amino acid properties, we use three amino acid properties including hydrophobicity value, hydrophilicity value and side chain mass in our work. Furthermore, a fingerprint containing 881 chemical substructures, defined in the PubChem database, is used for representing each drug compound.

As schematically represented in Fig. 1, our study consists of three main parts including: representing drug–target pairs, constructing positive and negative dataset, and building learning model. Because the aim of this study is to investigate the performance of the Bigram-PSSM descriptors in predicting drug–target interaction compared to the PAAC descriptors, these protein representations are combined with a fixed PubChem drug representation. Moreover, as the drug–target dataset is an unbalanced dataset, which needs to be balanced before model training, in this study we also investigate the effect of sampling methods on the accuracy of predictions. Therefore, two sampling methods including the simple random sampling and the balanced random sampling, proposed in Yu et al. (2010), are used for selection of negative samples from the non-interacting drug–target space and the constructed models, built on these datasets, are compared with respect to a set of performance indicators.

The rest of this paper is organized as follows: In Section 2, the materials and methods of this study including preparation of gold standard dataset, construction of positive and negative samples, features used for representation of drugs and targets, classification algorithm and performance evaluation method will be discussed. In Section 3, the prediction results of the built models on the gold standard dataset will be given and calculating different performance metrics will assess the performance of models. Furthermore, the effect of two sampling methods on the performance of model will be compared in this section. We also have a comparison between the Bigram-PSSM features and the AM-PSSM features, recently proposed by Nanni et al. on our dataset. Finally, the paper will be concluded in Section 4.

2. Materials and methods

2.1. Gold standard dataset

To be comparable with the previous studies, the gold standard dataset released by Yamanishi et al. (2008) is used as benchmark dataset in this study. In the gold standard dataset, the information about interacting or non-interacting drug–target pairs has been taken from the KEGG BRITE, BRENDA, SuperTarget and DrugBank databases. The proteins in this dataset are classified into 4 main classes including enzymes, ion channels, GPCRs and nuclear receptors, and there are 664, 204, 95 and 26 target proteins in these classes respectively. The number of known drugs targeting proteins in these classes is 445, 210, 223 and 54, respectively, and the number of known interacting drug–target pairs in each class is 2926, 1476, 635 and 90 respectively. More details about the curation of this dataset and some statistical properties of the four classes have been provided in Yamanishi et al. (2008).

The amino-acid sequence of target proteins, presented in the gold standard dataset, is retrieved from the Uniprot database using the Protr package (Nan et al.) Protr is an R package which has been implemented to extract most of the state-of-the-art features from the protein sequence. In addition, the all.sdf file in the DrugBank database includes the information about all drugs, and the chemical structure of drugs has been represented in this file using the SMILES format. The SMILES term is an acronym of Simplified Molecular-input Line-entry System, which encodes the chemical structure of molecules in a line notation. Accordingly, we extract the chemical structures of the drugs, presented in the gold standard dataset, from this file using their DrugBank identification indices.

2.2. Constructing positive and negative samples

Drug–target interaction network can be visualized as a bipartite graph, so that each edge connects a drug to a target protein. To extract positive and negative samples from this bipartite network for the classification framework, all known drug–target pairs in dataset (i.e. 5127 interacting drug–target pairs) are considered as positive samples and the remaining drug–target pairs correspond to the non-interacting drug–target pairs which can be considered as negative samples. Because the size of non-interacting pairs is not comparable with the size of interacting pairs, the constructed dataset is unbalanced. To resolve the bias caused by the unbalanced dataset, some approaches have randomly selected negative samples from the non-interacting pairs until their size reaches approximately to the size of positive samples (Wang et al., 2010; Wang et al., 2011; Cao et al., 2012).

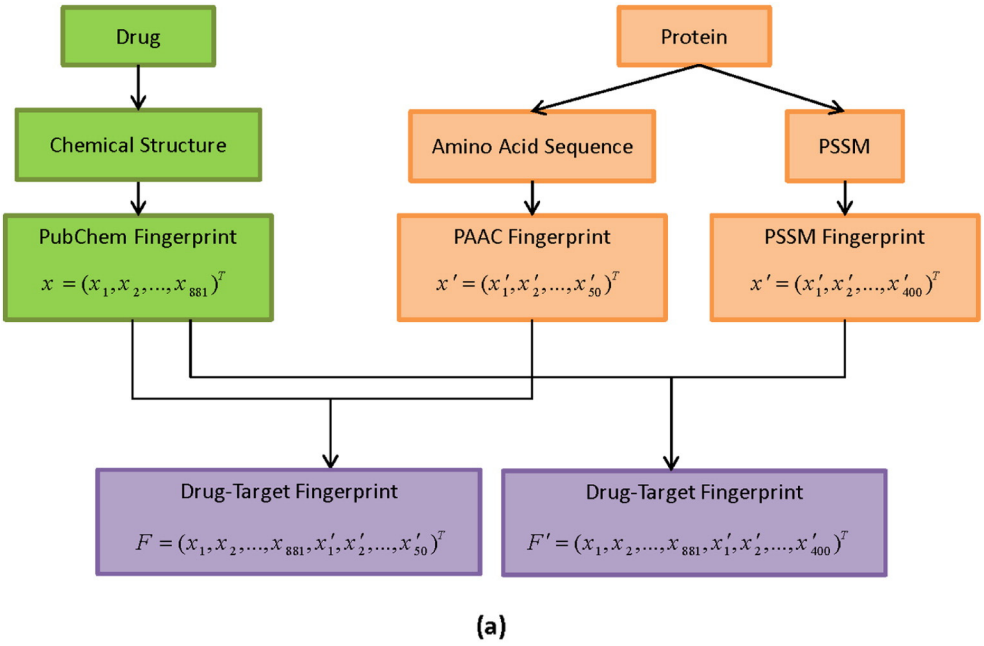
The mentioned random selection method is also applied in this study for construction of negative dataset from the non-interacting space. However, the sampling method may influence the prediction results, and selecting negative samples by a different approach may lead to a distinct accuracy of prediction. This is an important point especially when one is going to conduct performance comparisons between different models. To study the influence of sampling methods on the performance of classifier, we construct a second dataset containing positive and negative samples, in which the negative samples are selected using a different sampling method. This sampling method was firstly applied by Yu et al. for selection of negative samples in the problem of protein–protein interaction prediction and was called BRS-nonint method (Yu et al., 2010).

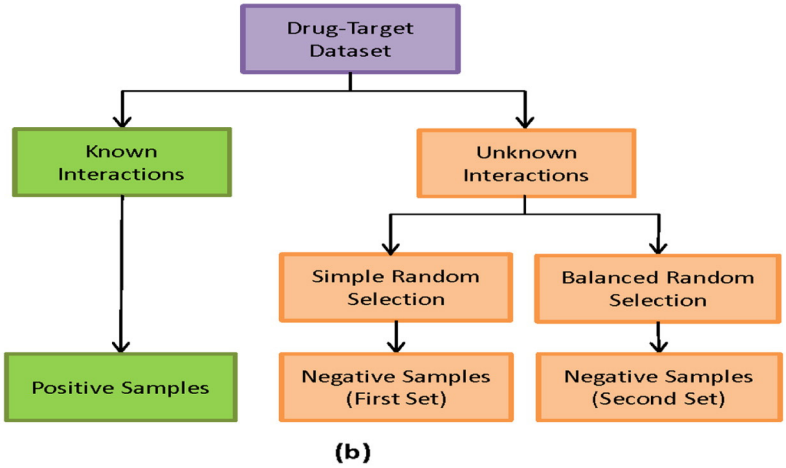
In the BRS-nonint method, non-interactions are randomly selected from the complement graph of protein–protein interactions until the degree of each protein in the negative dataset is equal to its degree in the positive dataset. We firstly use this method on drug–target data to create a negative dataset, so that all drugs and target proteins have the same degree distribution between positive and negative datasets. Despite protein–protein interaction networks, in drug–target network there are two types of nodes including proteins and drugs, and we expect to have only drug–target pairs in both positive and negative datasets. So, all drug-drug pairs and target–target pairs in the complement graph of drug–target interactions should be avoided in constructing negative dataset.

In the rest of this paper, we will refer to the randomly and non-randomly selected datasets as *random dataset* and balanced dataset respectively, and we will compare the prediction results of classifiers on both datasets in the results section (Section 3).

2.3. Representing drug molecules

For representing drug compounds, various types of descriptors can be defined on the basis of different types of drug properties. One of





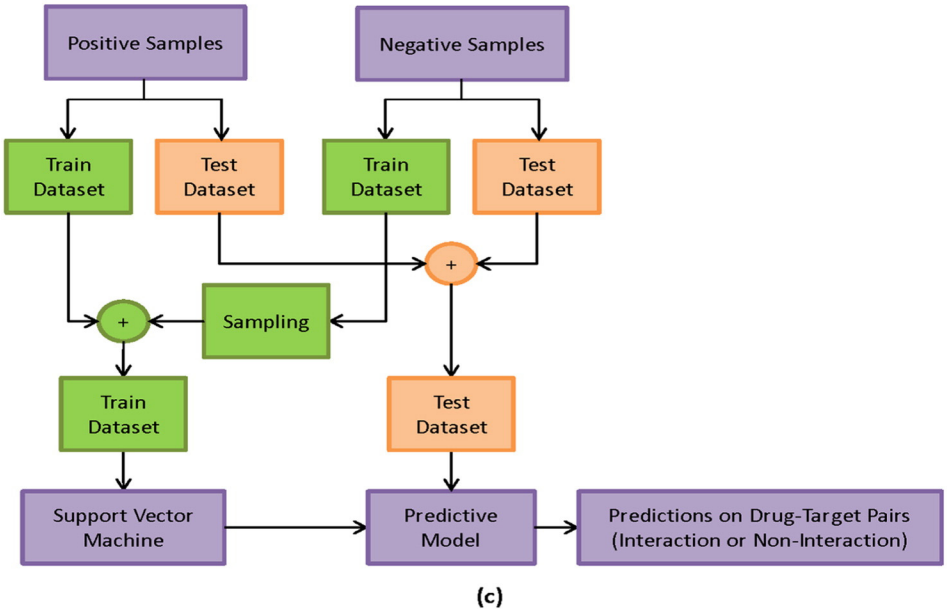


Fig. 1. Schematic representation of our study: a) representing drug–target pairs, b) constructing positive and negative datasets and c) building learning model.

these descriptors, that has been indicated in some studies to be effective in representing drugs, is the molecular substructure fingerprint (Tabei et al., 2012; Yamanishi et al., 2011; Tabei & Yamanishi, 2013). Each drug can be encoded by a molecular substructure fingerprint with a series of binary bits, which indicate the presence or absence of corresponding substructures in the molecule. There are 881 chemical substructures defined in the PubChem database, and each substructure can be assigned to a particular bit in the molecular fingerprint. Consequently, for any substructure which is seen in the drug molecule, the corresponding bit is set to 1 and for any substructure which is not seen in the drug, the corresponding bit is set to 0. Recently, researchers have developed many packages for measuring several kinds of descriptors for drug compounds. In our study, the rcdk package (Guha, 2007), an R package which has been developed to evaluate different types of descriptors for a chemical compound, is used for providing PubChem fingerprints. The chemical structure of each drug is given to the rcdk package using its SMILES format and then each drug is encoded in a 881-dimensional binary vector using this package.

2.4. Representing target proteins

Different kinds of protein descriptors extracted from various protein representations have been used by researchers in the protein-related learning problems including predicting secondary structure of proteins, predicting subcellular location of proteins and classifying proteins based on their functionality and so on. In this study, for representation of target proteins, we have used two types of descriptors, extracted from the amino acid sequence of protein and the PSSM of protein. The first one is a group of descriptors known as the PAAC descriptors, proposed by Chou (2001), which describes the structure information of proteins on the basis of their amino acid sequences. A set of *n* amino acid properties can be involved in the definition of PAAC descriptors. The original value of each amino acid property is normalized as follows:

$$H(i)=\frac{H^o(i)-\frac{1}{20}\sum_{i=1}^{20}H^o(i)}{\sqrt{\frac{\sum_{i=1}^{20}\left[H^o(i)-\frac{1}{20}\sum_{i=1}^{20}H^o(i)\right]^2}{20}}}$$

where $H^o(i)$ is the original value of an amino acid property for the i-*th* amino acid. Based on the normalized values of *n* amino acid properties, a correlation function can be defined between each pair of amino acids as:

$$\Theta(R_i,R_j)=\frac{1}{n}\sum_{k=1}^{n}\left[H_k(R_i)-H_k(R_j)\right]^2$$

where $H_k(R_i)$ and $H_k(R_j)$ refer to the normalized values of the k-*th* property in the amino acid property set for amino acids R_i and R_j respectively. Then for each protein, the following sequence-order correlated factors are defined as:

$$\theta_1=\frac{1}{N-1}\sum_{i=1}^{N-1}\Theta(R_i,R_{i+1})$$

$$\theta_2=\frac{1}{N-2}\sum_{i=1}^{N-2}\Theta(R_i,R_{i+2})$$

$$\vdots$$

$$\theta_\lambda=\frac{1}{N-\lambda}\sum_{i=1}^{N-\lambda}\Theta(R_i,R_{i+\lambda})$$

where N is the length of amino acid sequence and λ is a modifiable parameter that is less than N. Based on the defined factors, a set of 20 + λ descriptors, called the PAAC descriptors, are defined for a protein sequence as:

$$X_c=\frac{f_c}{\sum_{r=1}^{20}f_r+w\sum_{j=1}^{\lambda}\theta_j}\ (1<c<20)$$

$$X_c=\frac{w\theta_{c-20}}{\sum_{r=1}^{20}f_r+w\sum_{j=1}^{\lambda}\theta_j}\ (21<c<20+\lambda)$$

where f_c represents the normalized occurrence frequency of the i-*th* amino acid and w is a weighting factor which determines the sequence-order effect.

We use the Protr package (Xiao, Xu, & Cao, 2013) in our study to calculate the PAAC descriptors as described. Hydrophobicity value, the hydrophilicity value and the side chain mass, which are considered as amino acid properties, have been involved in this package by default. Using the default value of λ parameter in the protr package, which is equal to 30, a 50-dimensional real valued vector of descriptors is obtained for each protein:

$$X=(X_1,X_2,\ldots,X_{20+\lambda})^T.$$

Moreover, another type of protein descriptors based on the evolutionary information of proteins, firstly proposed by Sharma et al. (2013) for prediction of protein folding, is used in this study. These descriptors are the bi-gram probability features, which are extracted from the PSSM of proteins. The PSSM, which was firstly introduced by Gribskov, McLachlan, and Eisenberg (1987) for finding distantly related proteins, is generated from the multiple sequence alignment and used in the iterations of the Position Specific Iterated BLAST (PSI-BLAST). The PSSM of each protein is a matrix with *L* rows and 20 columns, where *L* is the total length of protein sequence. The element of PSSM in the i-*th* row and j-*th* column represents the probability of observing the j-*th* amino acid in the i-*th* location of protein sequence, which is obtained from the evolutionary information of protein. In our study, we calculate the PSSMs for all proteins using the Protr package. Subsequently, the bi-gram features of each proteins are measured based on the PSSM profile of protein as follows:

$$B_{m,n}=\sum_{i=1}^{N-1}P_{i,m}P_{i+1,n}(1\le m\le 20\ \text{and}\ 1\le n\le 20)$$

where $P_{i,j}$ denotes the element at the i-*th* row and the j-*th* column of the PSSM of protein. Therefore, a 400-dimensional bi-gram feature vector denotes each protein as:

$$B=\left(B_{1,1},B_{1,2},\ldots,B_{1,20},B_{2,1},\ldots,B_{2,20},\ldots,B_{20,1},\ldots,B_{20,20}\right)^T$$

where $B_{i,j}$ represents the frequency of transition from the i-*th* amino acid to the j-*th* amino acid in the primary sequence of protein. Avoiding zeros in the feature vector is the main advantage of bi-gram features extracted from the PSSM in comparison with the traditional bi-gram features extracted from the protein sequence, which makes it useful for protein representation in protein-related prediction and machine learning tasks.

2.5. Representation of drug–target pair

To encode each drug–target pair, the fingerprint of drug is concatenated to the fingerprint of target. Therefore, a fingerprint containing information about both drug and target is used for predicting interaction between drug and target. In this study, when we use the PAAC features for encoding proteins, each drug–target pair is encoded by a 931 dimensional vector, consists of 881 dimensions for representing drug and 50 dimensions for representing protein. When the Bigram-PSSM features are used for representation of proteins, a fingerprint consists of 1281

dimensions is used for encoding drug–target pair, in which 881 dimensions correspond to the drug and 400 dimensions correspond to the target.

2.6. Classification algorithm

In our study, we use Support Vector Machine (SVM) as a pattern classification engine which has been shown to be effective in solving many biological classification problems (Vapnik & Vapnik, 1998). For classification of instances which are linearly separable, SVM finds a hyperplane with a maximal margin to separate the instances of both classes in the multidimensional space with minimum classification error. Let $X_T = \{(x_i, y_i)\}$ be a set of training data, so that x_i refers to the feature vector of i-*th* instance and y_i denotes corresponding class label. The decision function of SVM is defined as follows:

$$f(x_i) = \mathrm{sgn}(w^t x_i + b).$$

In the above definition, *W* denotes the weight vector and *b* is a constant coefficient. For correct classification of training data, the parameters should also be satisfied in the following condition:

$$y_i(w^t x_i + b) \geq 1, i = 1, \ldots, N$$

SVM finds the parameters correspond to the optimal hyperplane by minimizing $\|w\|^2$, and the quadratic optimization method is used for solving the minimization problem.

2.7. Performance evaluation

In this study, we evaluate the performance of models by the 5-fold cross validation technique. In the 5-fold cross validation, dataset is split into five parts with roughly equal size. Four parts of dataset are used for training and one remaining part is used for testing. The training procedure is repeated five times, each time by a different set of folds, and each built model is evaluated using the remaining part. The total performance on whole dataset is measured by averaging the results over all hold-out parts. Here, with respect to the biased data, a modified 5-fold cross-validation method is used. Each time we select a hold-out part for test, the remaining training parts are sampled using either of the methods addressed in Section 2.2. The training is conducted on the sampled set, but the test is performed on the whole hold-out partition. Accordingly, whole drug–target pairs will be included in test, while only a sampled subset of them will be involved in learning.

A variety of performance indicators can be used to assess the performance of predictor model. Assume P (Positive) is the number of interacting samples, N (Negative) is the number of non-interacting samples, TP (True Positive) denotes the number of interacting samples classified correctly, TN (True Negative) denotes the number of non-interacting samples classified correctly, FP (False Positive) be the number of non-interacting samples classified erroneously as interacting, and FN (False Negative) indicates the number of interacting samples classified erroneously as non-interacting, then the performance indicators are defined as follows:

- Sensitivity or Recall (TPR or True Positive Rate) refers to the percentage of positive samples which are classified correctly. It is defined as:

$$\text{Sensitivity} = \frac{TP}{P} = \frac{TP}{TP + FN}.$$

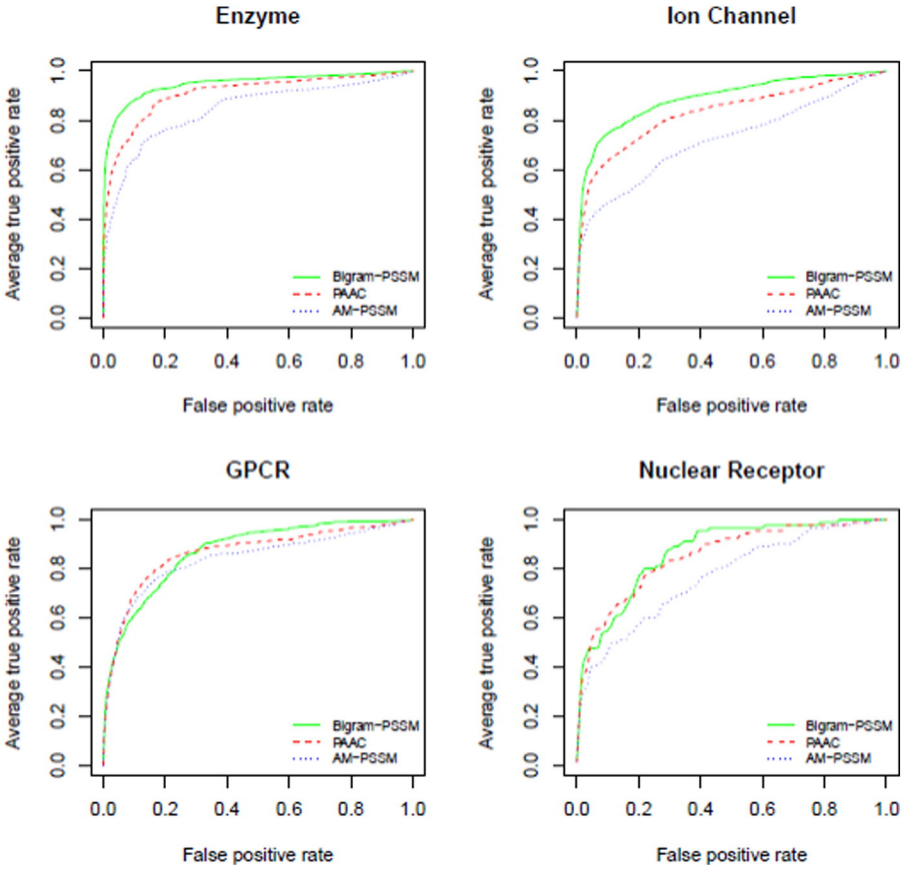


Fig. 2. ROC curves of the Bigram-PSSM, PAAC and AM-PSSM models for all classes of target proteins: enzymes, ion channels, GPCRs and nuclear receptors.

- Specificity (TNR or True Negative Rate) refers to the percentage of negative samples which are classified correctly. It is defined as:

$$\text{Specifity} = \frac{\text{TN}}{\text{N}} = \frac{\text{TN}}{\text{TN} + \text{FP}}.$$

- Precision refers to the percentage of positive predictions which truly belong to the positive class. It is defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

AUC and AUPR are two another important measures which can be used for evaluating the classification accuracy. AUC refers to the area under ROC curve, which is a plot of TPR against FPR at multiple threshold settings, and the AUPR represents the area under precision–recall curve, which shows the precision rate vs the recall rate for different thresholds. Due to the fact that the highly ranked false positive samples are punished by the AUPR much more than the AUC, the AUPR is a more accurate measure for evaluating performance in dealing with highly skewed dataset compared to the AUC.

3. Results and discussion

This section describes two experimental settings and their results as well as a comparative analysis between the proposed method and the other existing methods. The first setting compares the performance of classifier using the Bigram-PSSM protein features against its performance with the PAAC descriptors. Furthermore, in the second experiment, two sampling methods including random and balanced sampling are used for constructing the negative dataset and the influence of these methods is studied on the performance of model. In Section 3.3, an additional experiment is conducted to compare the Bigram-PSSM features with the AM-PSSM features recently proposed by Nanni, Lumini, and Brahnam (2014a). To have a comparison between all models, the ROC plots and the precision–recall curves of models including the Bigram-PSSM, the PAAC and the AM-PSSM models are shown in Figs. 2 and 3 respectively. In order to compare all models in more details, the average values of sensitivity, specificity and precision are also calculated for each model by considering several thresholds for prediction score. Because the prediction scores with high confidence are interesting in practical applications, we change the threshold from the upper one percentile in the prediction score to the upper ten percentile and compute the performance indicators based on each threshold. For all datasets, statistics of the prediction performance for all models are given in Table 1.

3.1. Bigram-PSSM vs. PAAC

In the first experiment, the performance of the bigram-PSSM protein descriptors versus the PAAC ones in identifying drug–target interactions is compared. Two predictor models are learned using each of protein descriptors in concatenation with a fixed PubChem drug descriptor, which called Bigram-PSSM and PAAC models in the rest of this paper. It should be mentioned that the *random dataset* detailed in Section 2.2 is used in this experiment.

As can be seen in Fig. 2, among the four datasets, the prediction ability of the Bigram-PSSM model is significantly greater than the PAAC one for enzymes and ion channels and the curves are also comparable for GPCR and nuclear receptors. To have a further comparison between the Bigram-PSSM and the PAAC models, we also compare the precision–recall curves of both models. As indicated in Fig. 3, the Bigram-PSSM model achieves improvements for enzymes and ion channels in terms of the area under precision–recall curve similar to

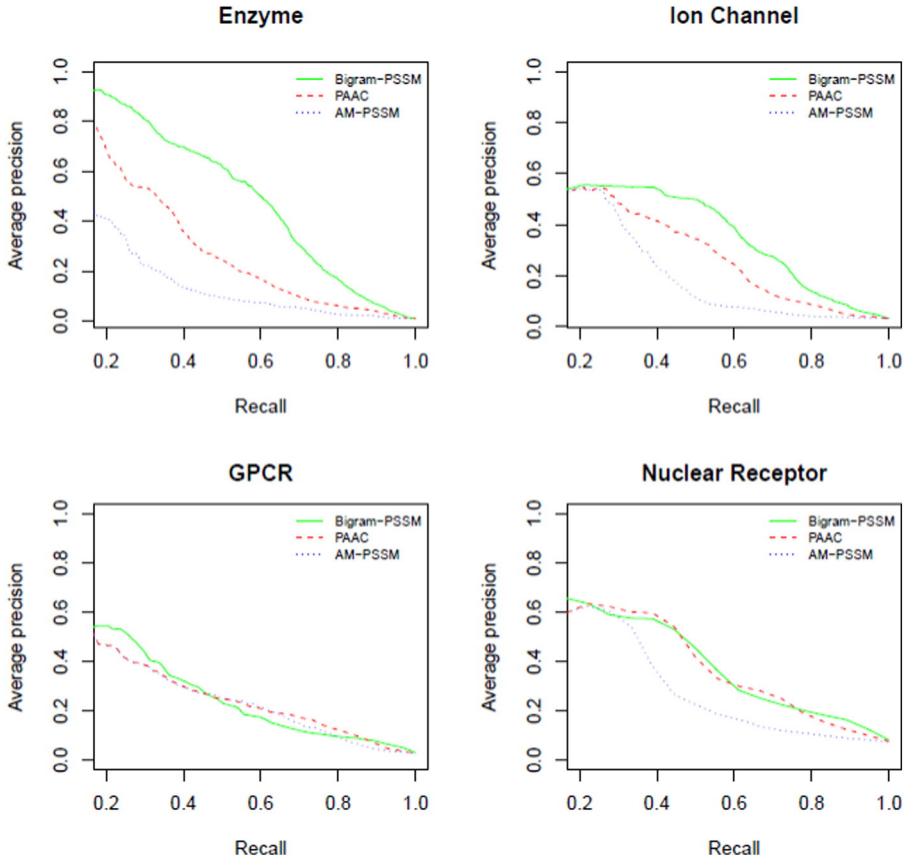


Fig. 3. Precision–recall curves of the Bigram-PSSM, PAAC and AM-PSSM models for all classes of target proteins: enzymes, ion channels, GPCRs and nuclear receptors.

Table 1
Statistics of the prediction performance for the Bigram-PSSM, the PAAC and the AM-PSSM models.

Data	Sampling	AUC	AUPR	Threshold	Sensitivity	Specificity	Precision
Enzyme	Bigram-PSSM	94.8	54.6	Top 1%	52.6	99.6	57.2
				Top 2%	55.1	99.6	55.9
				Top 5%	58.4	99.5	52.3
				Top 10%	60.9	99.4	48.9
	PAAC	91	34.7	Top 1%	33.9	99.6	48.2
				Top 2%	36.2	99.5	43.9
				Top 5%	39.5	99.3	36.7
				Top 10%	42.5	99	31.1
	AM-PSSM	84.3	21.9	Top 1%	13.8	99.9	49.2
				Top 2%	15.6	99.8	45.1
				Top 5%	20.6	99.7	41.8
				Top 10%	24	99.6	35.6
Ion Channel	Bigram-PSSM	88.9	39	Top 1%	29.9	99.1	55
				Top 2%	36.2	98.9	54
				Top 5%	42.8	98.6	51.6
				Top 10%	49.5	98.2	49
	PAAC	83.1	31	Top 1%	20.5	99.3	52.6
				Top 2%	22.4	99.3	52.9
				Top 5%	24.7	99.1	50.1
				Top 10%	29.1	98.8	47.6
	AM-PSSM	72.2	23.6	Top 1%	9.6	99.7	32.2
				Top 2%	13.3	99.6	43.6
				Top 5%	19.4	99.4	52.4
				Top 10%	23.4	99.2	52
GPCR	Bigram-PSSM	87.2	28.2	Top 1%	0	100	0
				Top 2%	0.9	100	38
				Top 5%	15.5	99.6	55
				Top 10%	30.9	98.6	42.8
	PAAC	86.5	29.1	Top 1%	18.1	99.4	53.3
				Top 2%	19.7	99.3	50.4
				Top 5%	22.7	99.1	46.1
				Top 10%	26.8	98.8	42.7
	AM-PSSM	83.9	27.4	Top 1%	15	99.6	56.4
				Top 2%	16.6	99.5	54.3
				Top 5%	18.8	99.4	49.8
				Top 10%	21.3	99.2	46.8
Nuclear Receptor	Bigram-PSSM	86.9	41.1	Top 1%	0	99.8	0
				Top 2%	7.7	99.4	29.7
				Top 5%	18.9	99.2	70.9
				Top 10%	33.3	98.4	61.4
	PAAC	85.2	39.7	Top 1%	16.7	99.1	60
				Top 2%	22.2	98.9	60.8
				Top 5%	31.1	98.5	59.6
				Top 10%	35.6	98.1	57.3
	AM-PSSM	76.7	31.7	Top 1%	7.7	99.6	31
				Top 2%	15.6	99.4	50
				Top 5%	17.8	99	54.9
				Top 10%	25.6	98.8	59

the area under the ROC curve. It seems that the Bigram-PSSM model has a good performance for enzymes and ion channels, followed by nuclear receptors and GPCR.

Furthermore, as seen in Table 1, for all classes of drug–target interactions except GPCR, the AUC measure of Bigram-PSSM model is slightly better than that of the PAAC model. Furthermore, when we compare the two models based on the AUPR measure, which is a more important performance measure than the AUC, the Bigram-PSSM model provides larger improvements for enzymes, ion channels and nuclear receptors respectively, and for GPCR, the AUPR score of this model is also comparable with that of the PAAC model. Moreover, the specificity of both models is quite high for all threshold settings on the four datasets, but the sensitivity and the precision values vary. Same as the previous observations, in the case of enzyme and ion channel datasets, the Bigram-PSSM model significantly outperforms the PAAC model in terms of sensitivity and precision, but for nuclear receptor and GPCR datasets, the PAAC model achieves higher sensitivity and precision values for all threshold settings except the threshold of prediction score is set to the upper 10%. Here, it seems that there is a trade-off between the high-confidence score and the high sensitivity and specificity values which can be attained by the Bigram-PSSM model for nuclear receptor and GPCR datasets.

Totally it can be concluded that, for the current machine-learning task, the bi-gram features extracted from the evolutionary information of proteins summarized in the PSSM can be more informative than the pseudo amino acid features extracted from the raw sequence of proteins. The good performance of different kinds of protein descriptors extracted from the evolutionary information has been shown in many classification problems related to proteins (Sharma et al., 2013; Nanni et al., 2014a; Dehzangi et al., 2015; Nanni, Lumini, & Brahnam, 2014b), and the results of this experiment also confirm this fact in drug–target interaction prediction.

3.2. Random vs. balanced

As mentioned earlier, the method of BRS-nonint has been firstly proposed by Yu et al. for selecting non-interacting pairs in the learning task of protein–protein interaction prediction Yu et al. (2010) have suggested this sampling method to maintain the degree distribution for each protein in both positive and negative datasets. They have shown that the simple features extracted from the protein sequence can't predict protein–protein interaction and they have shown that the AUC of some models would be drastically decreased by changing the negative sampling method from the random sampling to the balanced sampling. This study has motivated us to investigate the effect of negative sampling method on the performance of the model in the area of drug–target interaction prediction.

The *balanced dataset* introduced before (Section 2.2) is used in this experiment and the Bigram-PSSM model is evaluated on this new dataset to compare its results with the obtained results from the *random dataset* (also shown in Table 1). Fig. 4 shows the ROC plots of model on both datasets for the four classes of drug–target interactions. As indicated in Fig. 4, changing the strategy of negative selection, from the random sampling to the balanced sampling, decreases the performance of model specifically in the case of GPCR and nuclear receptor datasets.

To further study the effect of sampling method on the prediction results of Bigram-PSSM model, several statistics are given in Table 2. For the four datasets, a detailed comparison between two sampling methods shows that incorporating the balanced sampling leads to lower sensitivity and precision values in most cases of threshold settings. The AUC and the AUPR measures of the model using the random dataset are greater than those of the model on the balanced dataset for all classes of target proteins. Among the four datasets, the influence of sampling method on the performance indicators can be better observed for GPCR and nuclear receptors. For GPCR and nuclear receptors, when the upper 10% in the prediction score is considered as interaction, the sensitivity of Bigram-PSSM model decreases from 30.9 and 33.3 to 3.8 and 2.2 respectively. This observation is also seen for the sensitivity and precision values in other threshold settings. Totally, the reduction in sensitivity measure is the most significant on GPCR and nuclear receptors and the least significant on ion channels and enzymes. To further investigate the effect of sampling method on the performance of model, we also evaluate the PAAC model and the AM-PSSM model (the AM-PSSM model will be introduced in Section 3.3) on the balanced dataset. The obtained results confirm that the performance of models is drastically decreased using the balanced dataset specifically for GPCR and nuclear receptor datasets. The results of this experiment can be found in Tables S1–S2 and Figs. S1–S2 of the Supplementary materials.

As seen in the results above, the choice of which drug–target pairs are considered as non-interacting pairs in model training has a noticeable impact on the performance of model, and in most cases the performance of model decreases by changing the sampling method from a random sampling method to a balanced sampling method. However, all relevant studies to date have used the simple random sampling method to construct an unbiased negative dataset for training and

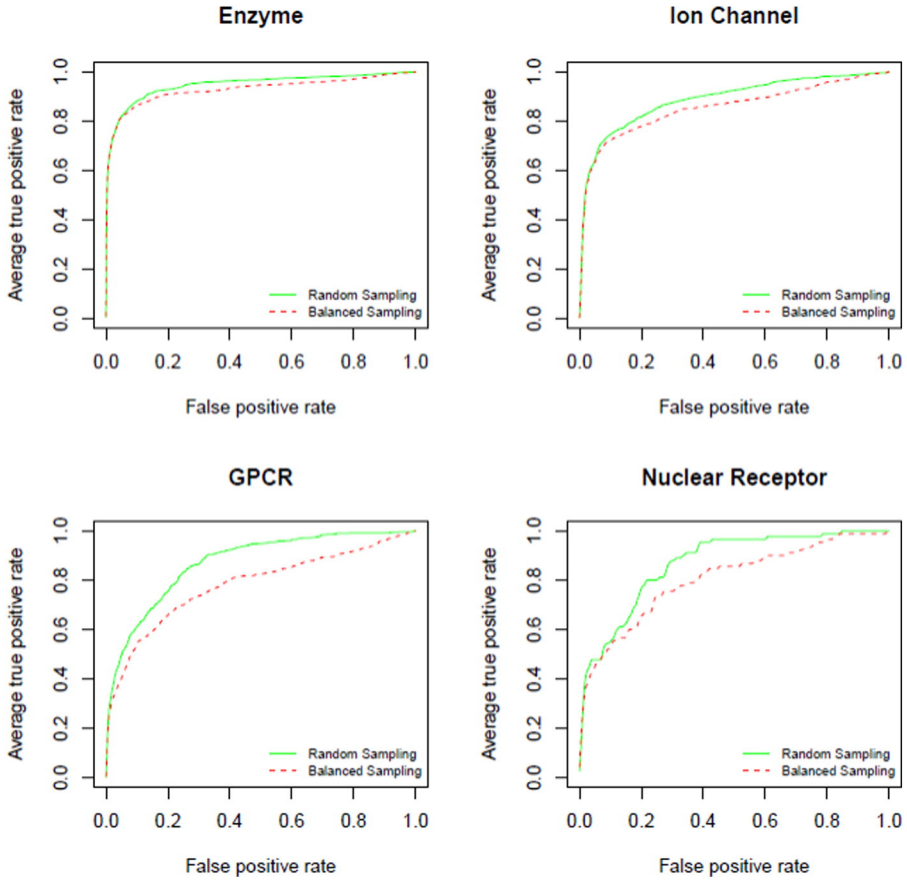


Fig. 4. ROC curves of the Bigram-PSSM model using random and balanced datasets for all classes of target proteins: enzymes, ion channels, GPCRs and nuclear receptors.

achieved very close-to-optimal results. Our results suggest that the reported results in these studies might be an over-estimation of real performances. Accordingly, we conclude that more care should be taken when evaluating methods. It should be noted that this does not mean that the comparative results are wrong (i.e., if a method is reported better using random samples, this is a probably true result); what we should be concerned about is, the magnitude of reported measures.

3.3. Comparison with existing research

As far as we know, limited studies have been done in the area of drug–target interaction prediction by utilizing PSSM-based features. Among them, we can refer to the recent study which has been published by Nanni et al. (2014a), in which the PSSM profiles of proteins have been used for extracting different sets of protein features. In this study Nanni et al. (2014a) have shown that the Autocovariance Matrix features extracted from the PSSM profile (called AM-PSSM features) have the minimum error rate in predicting drug–target interactions compared to the other proposed features. Accordingly, additional experiments are conducted to compare the performance of Bigram-PSSM features with the AM-PSSM features on the same dataset (i.e., the random dataset introduced in this paper).

To have a fair comparison between features, similar to the Bigram-PSSM model, the fixed PubChem drug descriptors are used in the AM-PSSM model for drug representation and all settings involving the SVM classifier and the cross-validation method are as same as the ones used in the Bigram-PSSM model. Only the Bigram-PSSM features are replaced by the AM-PSSM features in the new model. As it can be observed in Fig. 1, the prediction ability of the AM-PSSM model is much worse than that of the Bigram-PSSM and PAAC models for all datasets except GPCR. To further compare the Bigram-PSSM model with the AM-PSSM one, the average of AUC, AUPR, sensitivity, specificity and precision for both models can also be seen in Table 1.

As seen in Table 1, the Bigram-PSSM model performs better than the AM-PSSM model from the viewpoint of the AUC and the AUPR metrics. An interesting observation is that for the four datasets, different levels of improvements have been attained by the Bigram-PSSM model compared to the AM-PSSM model. In the case of enzymes and ion channels, it is clear that the sensitivity and precision of the Bigram-PSSM model is higher than those of the AM-PSSM model for all threshold settings. Furthermore for GPCR and nuclear receptors, the Bigram-PSSM model outperforms the AM-PSSM model in higher-sensitivity thresholds.

As mentioned before, the learning methods proposed for prediction of drug–target interactions can be categorized into two classes: similarity-based and feature-based. Most of the studies (as well as our study) have used the dataset proposed by Yamanishi et al. (2008) to assess the prediction ability of their proposed methods. In the following, we compare the AUC of the Bigram-PSSM model with that of some well known methods including DBSI (Cheng et al., 2012), KBMF2K (Gönen, 2012), and NetCBP (Chen & Zhang, 2013), and the proposed methods in Yamanishi et al. (2008), Wang et al. (2010) and Yamanishi et al. (2010) for the four classes of target proteins. As shown in Table 3, the AUC of the Bigram-PSSM model is superior in comparison with the AUC of DBSI (Cheng et al., 2012), KBMF2K (Gönen, 2012), and NetCBP (Chen & Zhang, 2013) and the method proposed by Yamanishi et al. (2010) for the four datasets. Table 3 also indicates that the Bigram-PSSM model outperforms the other compared methods in most of the datasets.

4. Conclusion

In this paper, a new learning model based on the features extracted from the evolutionary information of proteins has been proposed to

Table 2
Statistics of the prediction performance for the Bigram-PSSM model using random and balanced sampling.

Data	Sampling	AUC	AUPR	Threshold	Sensitivity	Specificity	Precision
Enzyme	Random	94.8	54.6	Top 1%	52.6	99.6	57.2
				Top 2%	55.1	99.6	55.9
				Top 3%	56.8	99.5	54.7
				Top 4%	58.1	99.5	53.5
				Top 5%	58.4	99.5	52.3
				Top 10%	60.9	99.4	48.9
	Balanced	92.8	56	Top 1%	46.3	99.8	68.2
				Top 2%	48.6	99.7	65.8
				Top 3%	50.2	99.7	64.1
				Top 4%	51.4	99.7	63
				Top 5%	52.4	99.7	62.8
				Top 10%	57.6	99.6	59.1
Ion Channel	Random	88.9	39	Top 1%	29.9	99.1	55
				Top 2%	36.2	98.9	54
				Top 3%	39.9	98.8	53.6
				Top 4%	41.7	98.7	52.8
				Top 5%	42.8	98.6	51.6
				Top 10%	49.5	98.2	49
	Balanced	85.5	37.6	Top 1%	23.5	99.3	55.5
				Top 2%	32	99.1	55.4
				Top 3%	35.9	98.9	54.5
				Top 4%	38.5	98.8	53.3
				Top 5%	39.7	98.7	52.1
				Top 10%	47.3	98.3	49.8
GPCR	Random	87.2	28.2	Top 1%	0	100	0
				Top 2%	0.9	100	38
				Top 3%	1.7	99.9	29.8
				Top 4%	9.3	99.7	42.8
				Top 5%	15.5	99.6	55
				Top 10%	30.9	98.6	42.8
	Balanced	78	23.5	Top 1%	0	100	0
				Top 2%	0	100	0
				Top 3%	0	100	0
				Top 4%	0	100	0
				Top 5%	0	100	0
				Top 10%	3.8	99.9	47.2
Nuclear Receptor	Random	86.9	41.1	Top 1%	0	99.8	0
				Top 2%	7.7	99.4	29.7
				Top 3%	10	99.4	33.6
				Top 4%	13.3	99.3	47.9
				Top 5%	18.9	99.2	70.9
				Top 10%	33.3	98.4	61.4
	Balanced	80.3	39.2	Top 1%	0	100	0
				Top 2%	0	100	0
				Top 3%	0	100	0
				Top 4%	0	99.8	0
				Top 5%	0	99.8	0
				Top 10%	2.22	99.7	6.67

infer potential drug–target interactions. We here firstly used the bigram features extracted from the Position Specific Scoring Matrices (PSSM) in the problem of drug–target interaction prediction. To compare the performance of the Bigram-PSSM features against the PAAC features, a fixed drug representation such as PubChem fingerprint has been used in each model for drug representation. The obtained results demonstrated that the Bigram-PSSM model is significantly better than the PAAC model in terms of several performance statistics. Based on the ROC and the precision–recall curves, it can be concluded that the Bigram-PSSM features of proteins is more informative than the PAAC ones in predicting new drug–target pairs specifically for enzymes and ion channels.

Because of the unbalanced property of the dataset used in this study, the second contribution of this paper lies in studying the effect of negative selection strategy on the prediction ability of the model. From the obtained results, it is clear for us that the balanced sampling method reduces the model performance, and this degradation of performance is more significant for GPCR and nuclear receptors datasets. However, the other related studies have only used the simple random selection method to provide the negative dataset, and the impact of this method on the predicted performance has not been investigated before our study.

In this study, a gold standard dataset published by Yamanishi et al. (2008) has been used and the results of some studies on this dataset have been compared with our results. The better performance of the Bigram-PSSM model against the compared methods has been shown in terms of the AUC measure. The Bigram-PSSM model has also been compared with the AM-PSSM model, proposed recently by Nanni et al. (2014a), and the significant improvements in the performance indicators have been achieved by the Bigram-PSSM model in all classes of drug–target interactions.

Although the Bigram-PSSM model shows a good performance compared to most of the existing methods, there are a few studies like Xia et al. (2010) which have reported better performance indicators on the gold standard dataset (Yamanishi et al., 2008). Xia et al. (2010) have obtained better results for predicting drug–target interactions using semi-supervised learning methods like LapRLS and NetLapRLS. Although the previously defined chemical and genomic spaces by Yamanishi et al. (2008) have been used by Xia et al. (2010), utilizing semi-supervised learning methods has led to better results in comparison with those of Yamanishi et al. (2008). It is worth noting that the main purpose of our study is to demonstrate the good performance of the Bigram-PSSM features for prediction of drug–target interaction, and a general method like SVM was used for making a classifier. Even though the Bigram-PSSM features performed well using a commonly used classifier, some further improvements are conceivable by utilizing a semi-supervised learning method in the future.

Acknowledgments

The authors would like to thank Dr. Javad Zahiri for his useful discussions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.vascn.2015.11.002.

References

Bleakley, K., & Yamanishi, Y. (2009). Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*, *25*(18), 2397–2403.
Butina, D., Segall, M.D., & Frankcombe, K. (2002). Predicting ADME properties in silico: methods and models. *Drug Discovery Today*, *7*(11), S83–S88.
Byvatov, E., et al. (2003). Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of Chemical Information and Computer Sciences*, *43*(6), 1882–1889.

Table 3
The comparison of the Bigram-PSSM model with existing approaches in terms of the AUC.

Dataset	DBSI (Cheng et al., 2012)	KBMF2K (Gönen, 2012)	NetCBP (Chen & Zhang, 2013)	Yamanishi et al. (2008)	Yamanishi et al. (2010)	Wang et al. (2010)	Bigram-PSSM
Enzyme	80.75	83.2	82.51	90.4	89.2	88.6	94.8
Ion Channel	80.29	79.9	80.34	85.1	81.2	89.3	88.9
GPCR	80.22	85.7	82.35	89.9	82.7	87.3	87.2
Nuc. Rec.	75.78	82.4	83.94	84.3	83.5	82.4	86.9

Cao, D. -S., et al. (2012). Large-scale prediction of drug–target interactions using protein sequences and drug topological structures. *Analytica Chimica Acta*, *752*, 1–10.
Chen, H., & Zhang, Z. (2013). A semi-supervised method for drug–target interaction prediction with consistency in networks. *PloS One*, *8*(5), e62975.
Chen, X., Liu, M. -X., & Yan, G. -Y. (2012). Drug–target interaction prediction by random walk on the heterogeneous network. *Molecular BioSystems*, *8*(7), 1970–1978.
Cheng, A.C., et al. (2007). Structure-based maximal affinity model predicts small-molecule druggability. *Nature Biotechnology*, *25*(1), 71–75.
Cheng, F., et al. (2012). Prediction of drug–target interactions and drug repositioning via network-based inference. *PLoS Computational Biology*, *8*(5), e1002503.
Chou, K.C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure, Function, and Bioinformatics*, *43*(3), 246–255.
Dehzangi, A., et al. (2015). Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. *Journal of Theoretical Biology*, *364*, 284–294.
Donald, B.R. (2011). *Algorithms in structural molecular biology*. The MIT Press.
Gönen, M. (2012). Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics*, *28*(18), 2304–2310.
Gribskov, M., McLachlan, A.D., & Eisenberg, D. (1987). Profile analysis: Detection of distantly related proteins. *Proceedings of the National Academy of Sciences*, *84*(13), 4355–4358.
Guha, R. (2007). Chemical informatics functionality in R. *Journal of Statistical Software*, *18*(5), 1–16.
Günther, S., et al. (2008). SuperTarget and Matador: Resources for exploring drug–target relationships. *Nucleic Acids Research*, *36*(Suppl. 1), D919–D922.
He, Z., et al. (2010). Predicting drug–target interaction networks based on functional groups and biological features. *PloS One*, *5*(3), e9603.
Kanehisa, M., et al. (2006). From genomics to chemical genomics: New developments in KEGG. *Nucleic Acids Research*, *34*(Suppl. 1), D354–D357.
Masoudi-Nejad, A., Mousavian, Z., & Bozorgmehr, J.H. (2013). Drug–target and disease networks: Polypharmacology in the post-genomic era. *In Silico Pharmacology*, *1*(1), 1–4.
Morris, G.M., et al. (2009). AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, *30*(16), 2785–2791.
Mousavian, Z., & Masoudi-Nejad, A. (2014). Drug–target interaction prediction via chemogenomic space: Learning-based methods. *Expert Opinion on Drug Metabolism & Toxicology*, *10*(9), 1273–1287.
Nanni, L., Lumini, A., & Brahnam, S. (2014a). A set of descriptors for identifying the protein–drug interaction in cellular networking. *Journal of Theoretical Biology*.
Nanni, L., Lumini, A., & Brahnam, S. (2014b). An empirical study of different approaches for protein classification. *The Scientific World Journal*, *2014*.
Okuno, Y., et al. (2008). GLIDA: GPCR–ligand database for chemical genomics drug discovery—Database and tools update. *Nucleic Acids Research*, *36*(Suppl. 1), D907–D912.
Schomburg, I., et al. (2004). BRENDA, the enzyme database: Updates and major new developments. *Nucleic Acids Research*, *32*(Suppl. 1), D431–D433.
Sharma, A., et al. (2013). A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. *Journal of Theoretical Biology*, *320*, 41–46.
Tabei, Y., & Yamanishi, Y. (2013). Scalable prediction of compound–protein interactions using minwise hashing. *BMC Systems Biology*, *7*(Suppl 6), S3.
Tabei, Y., et al. (2012). Identification of chemogenomic features from drug–target interaction networks using interpretable classifiers. *Bioinformatics*, *28*(18), i487–i494.
Vapnik, V.N., & Vapnik, V. (1998). *Statistical learning theory. Vol. 2*, New York: Wiley.
Wang, Y. -C., et al. (2010). Computationally probing drug–protein interactions via support vector machine. *Letters in Drug Design & Discovery*, *7*(5), 370–378.
Wang, Y. -C., et al. (2011). Kernel-based data fusion improves the drug–protein interaction prediction. *Computational Biology and Chemistry*, *35*(6), 353–362.
Warr, W.A. (2009). ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute outstation of the European Molecular Biology Laboratory (EMBL-EBI). *Journal of Computer-Aided Molecular Design*, *23*(4), 195–198.
Wishart, D.S., et al. (2008). DrugBank: A knowledge base for drugs, drug actions and drug targets. *Nucleic Acids Research*, *36*(Suppl. 1), D901–D906.
Xia, Z., et al. (2010). Semi-supervised drug–protein interaction prediction from heterogeneous biological spaces. *BMC Systems Biology*, *4*(Suppl. 2), S6.
Xiao, N., Xu, Q., & Cao, D. (2013). *Protr: Protein Sequence Feature Extraction with R*. R package version 0.2-0.
Yamanishi, Y., et al. (2008). Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, *24*(13), i232–i240.
Yamanishi, Y., et al. (2010). Drug–target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, *26*(12), i246–i254.
Yamanishi, Y., et al. (2011). Extracting sets of chemical substructures and protein domains governing drug–target interactions. *Journal of Chemical Information and Modeling*, *51*(5), 1183–1194.
Yu, H., et al. (2012). A systematic prediction of multiple drug–target interactions from chemical, genomic, and pharmacological data. *PloS One*, *7*(5), e37608.
Yu, J., et al. (2010). Simple sequence-based kernels do not predict protein–protein interactions. *Bioinformatics*, *26*(20), 2610–2614.