A comparative study of family-specific protein–ligand complex affinity prediction based on random forest approach

Yu Wang · Yanzhi Guo · Qifan Kuang · Xuemei Pu · Yue Ji · Zhihang Zhang · Menglong Li

Received: 10 September 2014 / Accepted: 16 December 2014 / Published online: 20 December 2014

Abstract The assessment of binding affinity between ligands and the target proteins plays an essential role in drug discovery and design process. As an alternative to widely used scoring approaches, machine learning methods have also been proposed for fast prediction of the binding affinity with promising results, but most of them were developed as all-purpose models despite of the specific functions of different protein families, since proteins from different function families always have different structures and physicochemical features. In this study, we proposed a random forest method to predict the protein–ligand binding affinity based on a comprehensive feature set covering protein sequence, binding pocket, ligand structure and intermolecular interaction. Feature processing and compression was respectively implemented for different protein family datasets, which indicates that different features contribute to different models, so individual representation for each protein family is necessary. Three family-specific models were constructed for three important protein target families of HIV-1 protease, trypsin and carbonic anhydrase respectively. As a comparison, two generic models including diverse protein families were also built. The evaluation results show that models on family-specific datasets have the superior performance to those on the generic datasets and the Pearson and Spearman correlation coefficients (R_p and Rs) on the test sets are 0.740, 0.874, 0.735 and 0.697, 0.853, 0.723 for HIV-1 protease, trypsin and carbonic anhydrase respectively. Comparisons with the other methods further demonstrate that individual representation and model construction for each protein family is a more reasonable way in predicting the affinity of one particular protein family.

Keywords Protein–ligand binding affinity prediction · Family-specific model · Generic model · Random forest

Introduction

Structure-based drug design methods, such as docking, have become a common tool in the drug discovery process over the past decade [1–3]. One of the most important issues in structure-based drug design methods is the screening of available ligands with their relevant target proteins. In most cases, the stronger a ligand binds with its target protein, it would more probably affect the physiological function of the protein, and as a consequence, it will be likely a suitable drug candidate [4]. Therefore, the assessment of the binding affinity between a ligand and its target protein plays an essential role in drug discovery and design process. The study on the relationship between the descriptors of a given protein–ligand complex and its binding affinity becomes very important in modern drug discovery process since the binding affinity is mainly determined by the interaction between the ligand and the relevant macromolecular target [5, 6].

The most widely used methods for predicting the binding affinity of protein–ligand complex are based on docking and scoring functions that can identify the binding

Electronic supplementary material The online version of this article (doi:10.1007/s10822-014-9827-y) contains supplementary material, which is available to authorized users.

Y. Wang · Y. Guo (✉) · Q. Kuang · X. Pu · Y. Ji · Z. Zhang · M. Li (✉)
College of Chemistry, Sichuan University,
Chengdu 610064, Sichuan, People's Republic of China
e-mail: yzguo@scu.edu.cn

M. Li
e-mail: liml@scu.edu.cn

modes of the ligands and estimate the strength of the protein–ligand interaction. Traditionally, the common scoring functions in molecular docking can be roughly divided into three different types: force field based methods (e.g. DOCK [7], GOLD [8], SIE [9], and LIE [10]), knowledge based potentials (e.g. DrugScore [11], PMF [12, 13], DFIRE [14], and 3DDFT [15]) and empirical scoring functions (e.g. X-Score [16], FlexX score [17], SCORE [18, 19], and SODOCK [20]).

Nevertheless, although a few scoring functions such as X-Score [16] achieves a remarkable performance on the PDBbind benchmark, despite improvements over the last years, most scoring functions still suffer from a rather poor correlation with experimental binding affinity [22, 23]. Besides, docking and scoring approaches are not easy implementation and often take a long time. For that reason, as an alternative to widely used docking and scoring approach, some other in silico methods such as Hi-PLS [24] and novel geometrical descriptors-based methods [25, 26] based on the structures of ligands and the relevant proteins are also proposed for the fast prediction of the binding affinity. These methods firstly use the molecular descriptors calculated from the structures of the ligand and its target, and then use machine learning methods to develop prediction model. Ballester and Mitchell [21] reported a machine-learning scoring function called random forest (RF)-score that employed RF and it outperformed all other scoring functions when tested on the core set in PDBbind V2007 by using protein–ligand complex descriptors and a nonlinear learning algorithm. Compared to the docking and scoring functions based methods, these methods have shown the obvious advantages such as easy implementation, fast prediction process and strong predictive ability.

Since successes have been achieved by the methods mentioned above, models constructed by these methods were mostly on large functionally and structurally diverse datasets. We can call them as generic models because they are based on diverse protein–ligand complexes despite of the functions of different target families. It is obvious that generic models do not take the functional specificity of each target family into account. It is widely believed that specific models are superior to generic ones because of the specificity, which have been proved by previous researches. For example, to address the problem that generic models overlooked the difference among the actual physiological states in different tissues, Zhao and Huang [27] reconstructed a human heart-specific metabolic network, Wang et al. [28] generated a heart-specific DM1 mouse model. Lewalle et al. [29], Heil et al. [30] and Xu et al. [31] constructed species-specific models rather than generic models. As for binding affinity prediction, Saranya and Selvaraj [32] developed QSAR models to predict the binding affinity only for HIV-1 protease inhibitors and achieved a good performance. Xue et al. [33] successfully developed a kinase target-specific scoring function to assess the binding of ATP-competitive kinase inhibitors.

Proteins belonging to different function families always have different structures and physicochemical features [33]. Therefore, in our work, three specific models were constructed for three different target families of HIV-1 protease, trypsin and carbonic anhydrase respectively. As a comparison, two generic models on diverse protein–ligand complexes were also built. Each protein–ligand complex was characterized by using a comprehensive feature set covering all aspects of each complex, including protein sequence, binding pocket, ligand structure and intermolecular interaction. From the feature importance evaluation and selection, the selected important features of each family are very different from each other because of their different functions. The large feature sets were observed in the generic models due to the larger protein–ligand complex sample space compared to the specific models. Moreover, the specific models yield a better performance than the generic models, which demonstrates that we should take specificity of different functions of protein families into account when predicting the affinity of the protein–ligand complex and it would be more reasonable to construct the specific model for the specific family.

Materials and methods

Dataset

All of the protein–ligand complex information was extracted from the refined set of PDBbind database [34]. The PDBbind database is the largest data collection of the protein–ligand complexes, with information for both binding affinities and known 3D crystal structures. Being updated every year, the version 2012 includes 2,897 protein–ligand complexes with experimentally measured binding affinity data. The refined set is composed by retrieving proteins that bind only one known drug like ligand, excluding those with a molecular weight higher than 1000 and both carbohydrates and nucleic acids. Then, compounds with cofactors and those with X-ray structure determined at a resolution higher than 2.5 Å were also excluded. Finally, the complexes with known dissociation constants (K_d) or inhibition constants (K_i) were considered. Since the binding affinity values range from 1.2 pM to 10.1 mM, we used the negative logarithm of K_d and K_i values in this study.

Here, five datasets of protein–ligand complexes were respectively constructed based on the refined set. Three are family-specific datasets for three important target families which are 170 complexes of HIV-1 protease, 110 complexes of trypsin and 126 complexes of carbonic anhydrase. The three protein families are also most populated in the PDBbind refined set. The other two are generic datasets that includes functionally and structurally diverse protein–ligand complexes. The number of protein–ligand complexes publicly available in the PDBbind database has grown from 1300 complexes in 2007 to 2897 complexes in 2012. Most of the researches published to predict the binding affinity of protein–ligand complexes applied the data of PDBbind database version 2007 or some version even older. In order to validate our model, we constructed two generic datasets from version 2012 and 2007 respectively. We found that four complexes from the refined set in the PDBbind database V2007 have been abrogated or replaced in the Protein Data Bank (PDB) now. As a consequence, in our work, there are 2897 protein–ligand complexes in version 2012 and 1296 protein–ligand complexes in version 2007, named as V2012 and V2007 respectively. The PDB IDs of the five datasets are listed in supplementary information S1.

Li et al. [35] proposed a strategy to generate data partitions using uniform sampling on a round-robin basis. Though this partitioning method is not thoroughly random, it has an obvious advantage that each partition could span the largest range of binding affinities and incorporates the largest structural diversity of different protein families. Similarly, in order to select the training samples that can fully represent the whole sample space in each dataset, affinities of the protein–ligand complexes were sorted from low to high and then divided into several subsets according to the affinity value intervals. According to the ratio (4:1) of the numbers of training samples versus testing ones, we randomly select the training samples at each sample interval. As a consequence, the HIV-1 protease dataset contains 136 complexes in the training set and 34 in the test set, the trypsin dataset includes 88 and 22 samples in the training and test set and the carbonic anhydrase dataset includes 100 and 26 complexes in the training and test set respectively. For the two generic datasets, V2007 and V2012 contain 1037 and 259 complexes, 2318 and 579 complexes in the training and test set, respectively. The training set and test set extracting process was randomly repeated five times for three family-specific datasets, and ten times for V2007 and V2012, since these two generic datasets have large amount of samples. So we built five models for each family-specific dataset and ten models for the two generic datasets respectively. A summary of all the datasets is shown in Table 1.

Methods

Feature extraction

The affinity of a protein–ligand complex is commonly decided by features from the target protein, ligand and their interaction. In this paper, we proposed a comprehensive feature set to represent all aspects of a protein–ligand complex. Each protein–ligand complex was described by descriptors from four blocks: protein sequence, binding pocket, ligand structure and intermolecular interaction. These four blocks of descriptors could cover the major information related to the specificity and the binding affinity.

Table 1 A summary of affinity range and the number of samples in each dataset used in this study

Protein family	pKd/pKi range	Number of complexes	
		Training set	Test set
HIV-1 protease	4.30–11.59	136	34
Trypsin	2.27–7.96	88	22
Carbonic anhydrase	3.90–10.52	100	26
V2007	0.49–13.96	1037	259
V2012	2.00–11.92	2318	579

Block 1: Descriptors based on protein sequence

The FASTA format sequences of all proteins were collected from PDB. Then, the structural and physicochemical features of proteins were computed from amino acid sequences using the web-version of PROFEAT software [36]. Seven types of features were generated, which are (1) amino acid and dipeptide composition, (2) normalized Moreau-Broto autocorrelation, (3) Moran autocorrelation, (4) Geary autocorrelation, (5) composition, transition, distribution, (6) sequence order and (7) Pseudo amino acid composition ($\lambda = 30$), respectively. At last, 1,080 descriptors were achieved.

Block 2: Descriptors from binding pocket

Binding pockets are the surface concavities of proteins where a substrate might bind. The binding pocket in the PDBbind database in each case was defined as the residues on the protein within 10 Å from the bound ligand observed in the crystal structure. According to international conventions, the distance cutoff of 10 Å means the distance from any atoms of the amino residue to any atoms of the ligand in a protein–ligand complex. Since the capability of a pocket to interact with small molecules determines the biological function of a protein, binding pocket descriptors are important to characterize the interaction between a protein and its ligand. Before calculating the binding pocket descriptors, the binding pocket structures of the protein–ligand complexes from the PDBbind database were first added with hydrogen atoms and then minimized to the lowest energy conformation. After that, 30 descriptors were generated including 27 CPSA (charged partial surface area) features, a FINGERPRINT feature, a MOLPROP_VOLUME feature, and a MOL_WEIGHT feature by Sybyl-X (Version 1.1).

Block 3: Descriptors of ligand structures

The following 6,122 structural descriptors of ligands were obtained using PowerMV software (Version 0.61) [37], including 546 atom pair descriptors, 4,662 atom pair descriptors, 735 fragment pair descriptors, 147 pharmacophore fingerprints descriptors, 24 weighted Burden number descriptors and 8 properties descriptors.

Block 4: Intermolecular interaction features

The intermolecular interaction features published by Ballester and Mitchell [21] were used. Briefly, the number of occurrences of Ligand_atom–Protein_atom pairs in a radius of 12 Å for the elements C, N, O, F, P, S, Cl, Br, and I of the ligand and C, N, O, and S of each protein are counted. Therefore, each complex will be characterized by a vector with 36 variables. The 12 Å cut-off distance was suggested in PMF [12] to incorporate the solvation effects to the maximum extent.

Feature pre-processing and principal component analysis(PCA)

Finally 7,268 feature variables were obtained from the four blocks of descriptors of protein–ligand complexes. Ballester et al. [38] found that a more precise chemical description of the protein–ligand complex does not generally lead to a more accurate prediction of binding affinity. Actually, when the number of descriptors is large, the feature set probably contains irrelevant and redundant variables that cause the dimensionality problem and make the model difficult to interpret [39]. This "curse of dimensionality" can also lead to model overfitting [40], so it is necessary to implement feature selection and compression. Here a principal component analysis (PCA) was employed to perform objective feature selection before model building.

Firstly, for each model, before separating them into the training set and test set, a rigorous pre-processing was performed on features of each block respectively. Descriptors in each block were checked for constant or near constant values and those detected were removed from the original feature vector. Then these descriptors were filtered to remove the redundant variables whose pair correlation coefficients were higher than 0.9.

Then considering the limited number of samples, PCA was applied to compress the features into principal properties and hence new information-rich orthogonal latent variables with reduced noise levels were obtained. The central idea of PCA is to reduce the dimensionality of a dataset consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs) which are uncorrelated and ordered so that the first few retain most of the variation present in all of the original variables [41]. In this work, accounting for ≥90 % variance of the original information, the significant PCs were obtained for the three specific models and two generic models.

The operations above were carried out on each data set. Finally, the features after feature selection and compression from the four blocks are merged into a new feature vector for every instance. After that, we implemented the partitioning method to separate the data set into the training set and test set. A workflow of feature processing and compression was shown in Fig. 1 and detailed information was shown in Table 2.

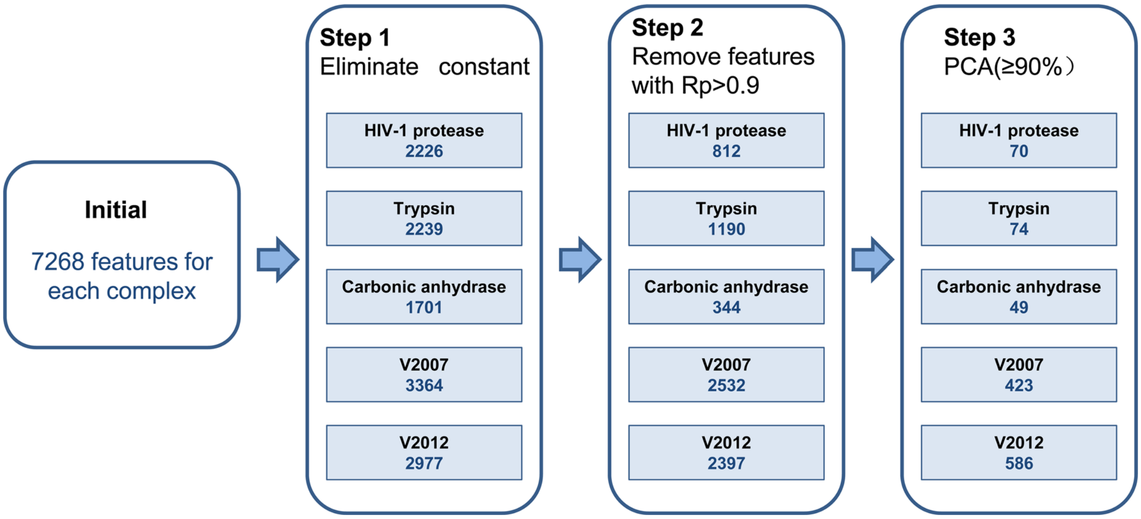

Fig. 1 The workflow of feature processing and compression

Table 2 Detailed information about feature processing and compression in each feature block for five models

Protein family	Feature block[a]	Feature numbers				Final features
		Initial	Constant out	Correlated Rp > 90 % out	90 % PCA	
HIV-1 protease	1	1,080	810	275	8	70
	2	30	30	13	5	
	3	6,122	1,351	515	53	
	4	36	35	9	4	
Trypsin	1	1,080	1,080	837	38	74
	2	30	30	13	4	
	3	6,122	1,093	339	31	
	4	36	36	1	1	
Carbonic anhydrase	1	1,080	956	88	4	49
	2	30	30	9	4	
	3	6,122	687	243	39	
	4	36	28	4	2	
V2007	1	1,080	1,080	824	113	423
	2	30	30	9	6	
	3	6,122	2,218	1,690	301	
	4	36	36	9	3	
V2012	1	1,080	1,080	818	146	586
	2	30	30	11	6	
	3	6,122	1,831	1,558	431	
	4	36	36	10	3	

[a] Block 1: descriptors based on protein sequence and structure. Block 2: descriptors from binding pocket. Block 3: descriptors of ligand structures. Block 4: intermolecular interaction features

Random forest modeling

In this study, we employed a RF model to establish the correlations between descriptors and binding affinities of the protein–ligand complexes. RF is a machine-learning method which is based on an ensemble of decision trees generated from bootstrap samples of training data, with prediction calculated by consensus over all trees. It has been shown to perform very well in non-linear regression [42]. Svetnik et al. [43] applied RF to investigate structure–activity relationships of pharmaceutical molecules. Polishchuk et al. [44] used RF to implement QSAR prediction of aquatic toxicity. In RF, a bootstrap sample was produced from the whole training set to form a subset for building each tree. The samples that are not used to build the current tree are placed in the out-of-bag (OOB) set. Each tree is trained on a different subset of the training set (approximately 60 %) and at every splitting node with a different subset of variables. This adds variability to the model and is the main reason for the improved robustness of RF compared to a single decision tree. The parameter m_{try}, the number of variables used at each splitting node, is the only tunable parameter that significantly influences the performance of the model. Each tree is then grown without pruning. The final model is chosen by the lowest error for prediction of the OOB set and only after that resulting model was applied for prediction of external test set. In addition, RF can be also used to estimate variable importance to identify those variables that contribute the most to the binding affinity prediction across known complexes. Here, the RF models were generated by the RF package in the R version 2.15.3 [45].

Model evaluation

Once the models were built, only the prediction results of training set was insufficient to prove the predictive ability of the model. Therefore, we implemented the internal validation and external validation to test the robustness of the model.

In RF, the standard way of assessing the predictive power is OOB validation. It is a type of cross-validation in parallel with the training step by using the so-called OOB set [46]. In OOB validation, the model training process is repeated *n* times with a randomly chosen subset, and the samples which are not used for training are predicted by the generated model. Usually, the model training process is repeated many times, far more than the number of the randomly chosen subsets, so that each sample can be predicted several times with different models. The overall prediction accuracy is then assessed from the average prediction on each sample. The OOB estimate is obtained by considering the OOB part of the data for the *i*th tree, denoted by D_i^{OOB}. The *i*th tree is used to predict the property of the observations in D_i^{OOB}. It has been shown [47] that on average each tree uses approximately 2/3 of the whole data set and hence the size of D_i^{OOB}, is on average 1/3 of the dataset. This implies that each observation will be in the OOB data about 1/3 of the time. Consequently, the OOB estimates can be aggregated to provide an ensemble prediction for each observation. This result is an OOB estimate of the mean square error (MSE) that can be used to approximate the MSE for the entire ensemble of trees. The MSE expressed in terms of the OOB samples is computed by Eq. (1) [42].

$$MSE \approx MSE^{OOB} = n^{-1}\sum_{i=1}^{n}\left\{\widehat{Y}^{OOB}(X_i) - Y_i\right\}^2 \tag{1}$$

In addition, external validation provides a more objective evaluation on the performance of the model. The data that are not used during the model development is the test set in our study and the test set were used for the external validation. The performance of regression model can be measured by Pearson correlation coefficient (R_p), Spearman correlation coefficient (R_s) and root mean squared error (RMSE):

$$R_p = \frac{N\sum_{n=1}^{N}p^{(n)}y^{(n)} - \sum_{n=1}^{N}p^{(n)}\sum_{n=1}^{N}y^{(n)}}{\sqrt{\left(N\sum_{n=1}^{N}(p^{(n)})^2\right)\left(N\sum_{n=1}^{N}(y^{(n)})^2 - \left(\sum_{n=1}^{N}y^{(n)}\right)^2\right)}} \tag{2}$$

$$R_s = \frac{N\sum_{n=1}^{N}p_r^{(n)}y_r^{(n)} - \sum_{n=1}^{N}p_r^{(n)}\sum_{n=1}^{N}y_r^{(n)}}{\sqrt{\left(N\sum_{n=1}^{N}(p_r^{(n)})^2 - \left(\sum_{n=1}^{N}p_r^{(n)}\right)^2\right)\left(N\sum_{n=1}^{N}(y_r^{(n)})^2 - \left(\sum_{n=1}^{N}y_r^{(n)}\right)^2\right)}} \tag{3}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(p^{(n)} - y^{(n)})^2} \tag{4}$$

where $y^{(n)}$ and $p^{(n)}$ are the values of experimentally determined affinity and estimated affinity of the *n*th complex out of N complexes in the test set, respectively, $\{y_r^{(n)}\}$ and $\{p_r^{(n)}\}$ are the rankings of $\{y^{(n)}\}$ and $\{p^{(n)}\}$, respectively.

Results and discussion

Feature compression and evaluation

After pre-processing, the remaining features in each block were compressed by PCA for each dataset. In this work, accounting for ≥90 % variance of the original information, the significant PCs were obtained for the three specific models and two generic models. From Table 2, we can see that in total, the original features of each model were efficiently compressed by PCA. Totally, the number of PCs for the three specific models of HIV-1 protease, trypsin and carbonic anhydrase are 70, 74 and 49 respectively, all much lower than 100. For the two generic models of V2007 and V2012, the number of compressed variables (PCs) is 423 and 586 respectively, which is only one-tenth of the number of original features. Although the features for V2007 and V2012 were also efficiently compressed, compared to the three specific models, the number of PCs is much higher than those of the three specific models because V2007 and V2012 datasets include much more diverse samples which distribute much larger feature space, so much more features are needed to cover the larger feature space, but the features of one protein family have stronger specificity than those of diverse function families. It can also be seen that the number of samples in V2012 dataset is nearly 2.5 times the number of samples in V2007, so the PC variables in V2012 model is more than those in V2007 model.

In order to further evaluate the variables in four blocks, all features were further analyzed by RF and the importance scores were achieved and represented as '%IncMSE'. '%IncMSE' is an estimate of the importance of the given descriptor for binding affinity prediction across the training data and it indicates the increase of the mean standard error after the permutation of one descriptor. A larger score suggests that a descriptor should contribute to protein–ligand binding affinity prediction remarkably. Figure 2 plots the average importance scores of the features in four blocks for the five models. For HIV-1 protease complexes, the permutation of features in ligand structure block increases MSE by 4.08 % on average, so ligand structure features contribute the most to HIV-1 protease-specific model. Moreover, according to importance score ranking, the top 10 descriptors with the highest scores are all from this block. After investigation, we find that most of the ligands of HIV-1 proteases are peptide-like ligands and these ligands have significant difference in structure with those of other proteins, so the effective characterization of ligand structure is most important for the binding affinity

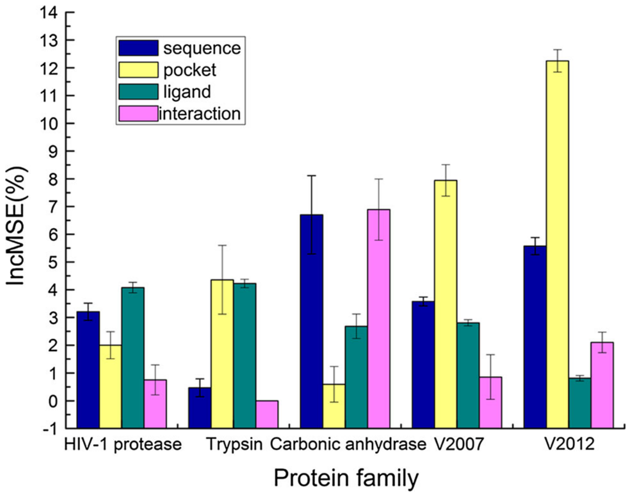

Fig. 2 The average importance scores of features in four blocks for five models

prediction of HIV-1 protease. As for trypsin complexes, the features of the binding pocket block contribute a great deal. Under the permutation test by RF, its MSE can increase 4.36 % on average. Among the top 10 descriptors, 9 are from ligand structure block and 1 is binding pocket feature. For carbonic anhydrase complexes, features from the intermolecular interaction block contribute the most and the permutation of features in this block increases MSE by 6.89 % on average. After investigating the interactions schematic plot in the PDB database, we find that the major interactions between carbonic anhydrase and their ligands are the hydrogen bonds, so those intermolecular features correlated with hydrogen bonds are contribute much more than other features.

In addition, an interesting finding can be seen that although the average score of features in protein sequence block is relatively high in three specific models, none of them ranks top 10 because the family-specific proteins have high sequence similarity. However, for V2007 and V2012 complexes, features in binding pocket block contribute the most and the permutation of features in this block increases MSE by 7.94 and 12.25 % respectively, on average. But for V2012 model, the top 10 descriptors include 8 from protein sequence block, but they are of relatively low importance scores and 2 from binding pocket block with the highest scores. As above, features that contribute the most to the prediction model are both from binding pocket block for two generic models. From the above discussion, we can conclude that important features for proteins of different functions are different. Due to the generic model contains functionally and structurally diverse protein–ligand complexes, the variables are substantially large in amount so that they can cover the whole sample space. So it is necessary to make specific characterization of family-specific proteins and for binding affinity prediction, it is more reasonable to construct specific models.

Prediction performance

In order to effectively test the performance of the method, the internal validation on the training set and external validation in the independent test set were implemented. The prediction results of the three specific and two generic models on test sets are shown in Fig. 3a–e and detailed prediction results including the training sets are listed in Table S1 in the supplementary information S2. As shown in Fig. 3 and Table S1, all models give a good internal performance on the training set with R_p higher than 0.97, indicating a very high linear dependence between these variables over the training set. However the three specific models obviously outperform the two generic models on the external validation with R_p higher than 0.72, especially the trypsin-specific model yields a very promising prediction result with R_p and R_s as high as 0.87 and 0.85 on the test set, but the two generic models yield R_p and R_s lower than 0.70.

As an excellent method for predicting protein–ligand complex affinity, RF-score [21] achieved a good performance by using only intermolecular interaction features and the nonlinear RF model. In order to further demonstrate the validity of our method, comparisons between our method and RF-score were implemented and the comparison results are shown in Fig. 4 and detailed information are listed in Table S2 in the supplementary information S2. The process of the selection of training set and testing set were repeated ten times for V2007 and V2012 datasets and five times for three specific datasets because of the relatively small size of datasets of the latter. In this study we used exactly the same training sets and the same test sets in order to make a fair comparison between our method and the RF-Score. So based on the same datasets, it is the fair comparison between our feature set and that of RF-Score. From Fig. 4, on average, RF-score also yields a good performance with R_p higher than 0.93 on the training sets for all five models. For the independent test sets, the two generic models by our method give a comparative performance with those by RF-score and the average R_p and R_s for V2007 and V2012 on the test sets are 0.69, 0.68 and 0.70, 0.68 by our method, 0.69, 0.71 and 0.69, 0.71 by RF-score respectively. However, for the three specific models our method performs better than RF-score. The average R_p and R_s for HIV-1 protease, trypsin, and carbonic anhydrase on the test sets are 0.74, 0.87, 0.74 and 0.70, 0.85, 0.72 by our method, 0.68, 0.67, 0.68 and 0.61, 0.58, 0.63 by RF-score respectively. The comparison results indicate that the four blocks of descriptors can more comprehensively represent the binding information between the ligand and the

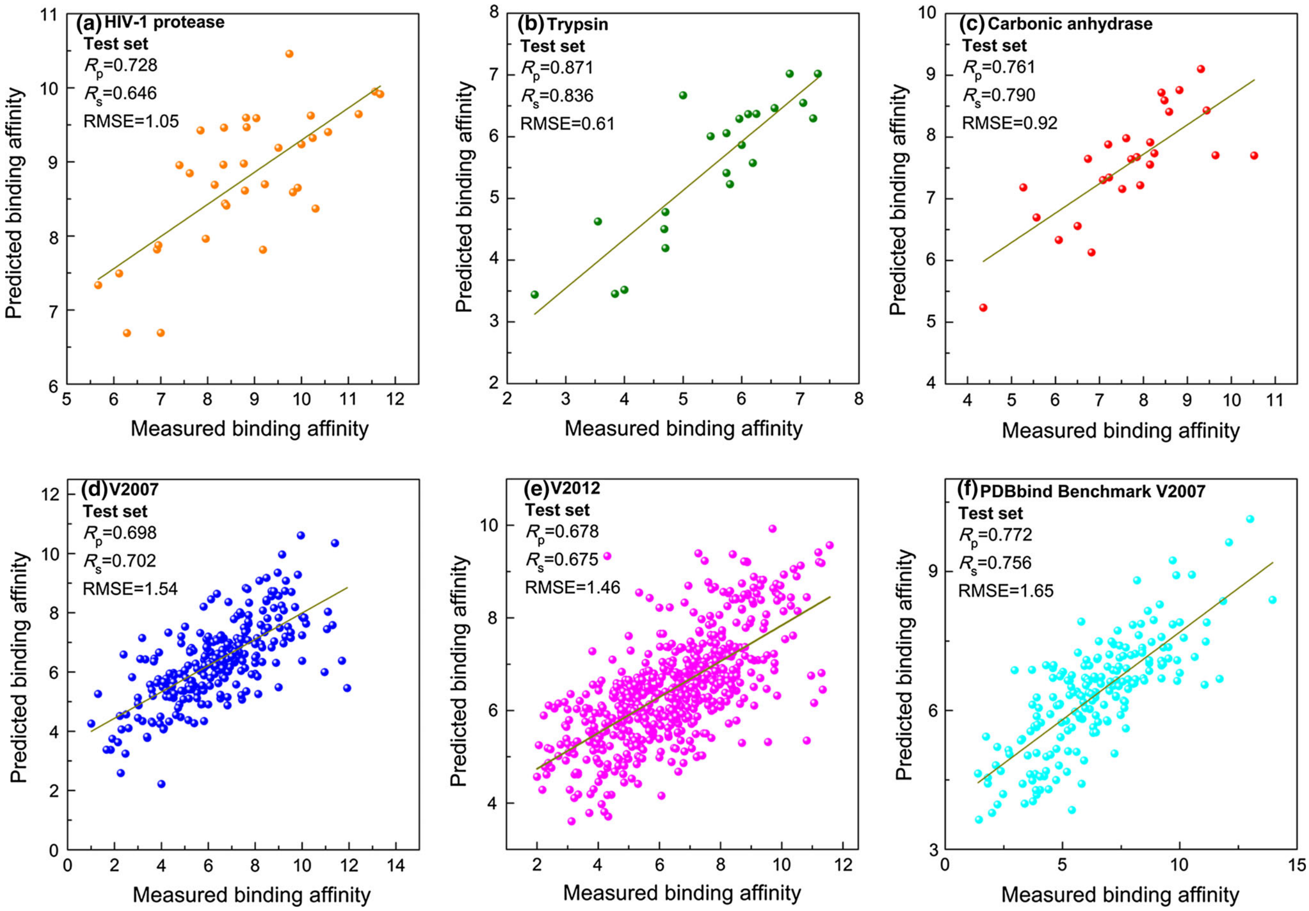

Fig. 3 Scatter plots of predicted versus measured binding affinity values of the three specific and two generic models along with PDBbind benchmark V2007

target protein, rather than only intermolecular interaction features. It is more reasonable to construct the individual prediction model for a protein family rather than generic model of diverse protein families.

In addition, Cheng et al. [48] have conducted a comparative assessment for 16 popular scoring functions on PDBbind benchmark V2007 by using 195 protein–ligand complexes in the core set as the test set and the remaining 1105 complexes as the training set. In order to further demonstrate the predictive power of our method, the performance of our method on PDBbind benchmark was also achieved. We used the exactly same 195 protein–ligand complexes in PDBbind benchmark V2007 as the test set and the remaining 1105 complexes in PDBbind V2007 as the training set. Figure 3f also shows the prediction result of our method on the PDBbind benchmark V2007 and Table 3 presents the performance of our method and the RF-Score, including the RF-Score v2.0 which is also published by Ballester et al. [38] that performs better than the old version of RF-Score, along with 16 scoring functions on the PDBbind benchmark V2007. The performance results for the other 16 scoring functions shown in Table 3 were extracted from Cheng et al. [48]. Comparison results show that our generic model and RF-Score achieve the better performance than other 16 scoring functions, indicating the superiority of machine-learning scoring functions. Since the R_p and R_s of the generic model by our method are slightly lower than those of RF-Score v1.0 (<0.01), the family-specific models by our method still give a superior performance to RF-Score.

Furthermore, we tested our method and RF-Score on the exactly same family test sets as Cheng et al. [48] used in their work, including HIV protease (112 complexes), trypsin (73 complexes), carbonic anhydrase (44 complexes) and thrombin (38 complexes). By excluding the samples in each family test set, the generic model by our method and RF-score were respectively constructed using the remaining complexes from PDBbind V2012 as the training set. The training set and test set are non-overlapping but the training set contains the target complexes as well so the comparison is fair and valid as the classical scoring functions also include the target in their training set. The prediction results of our method and RF-Score along with the selected classical scoring functions

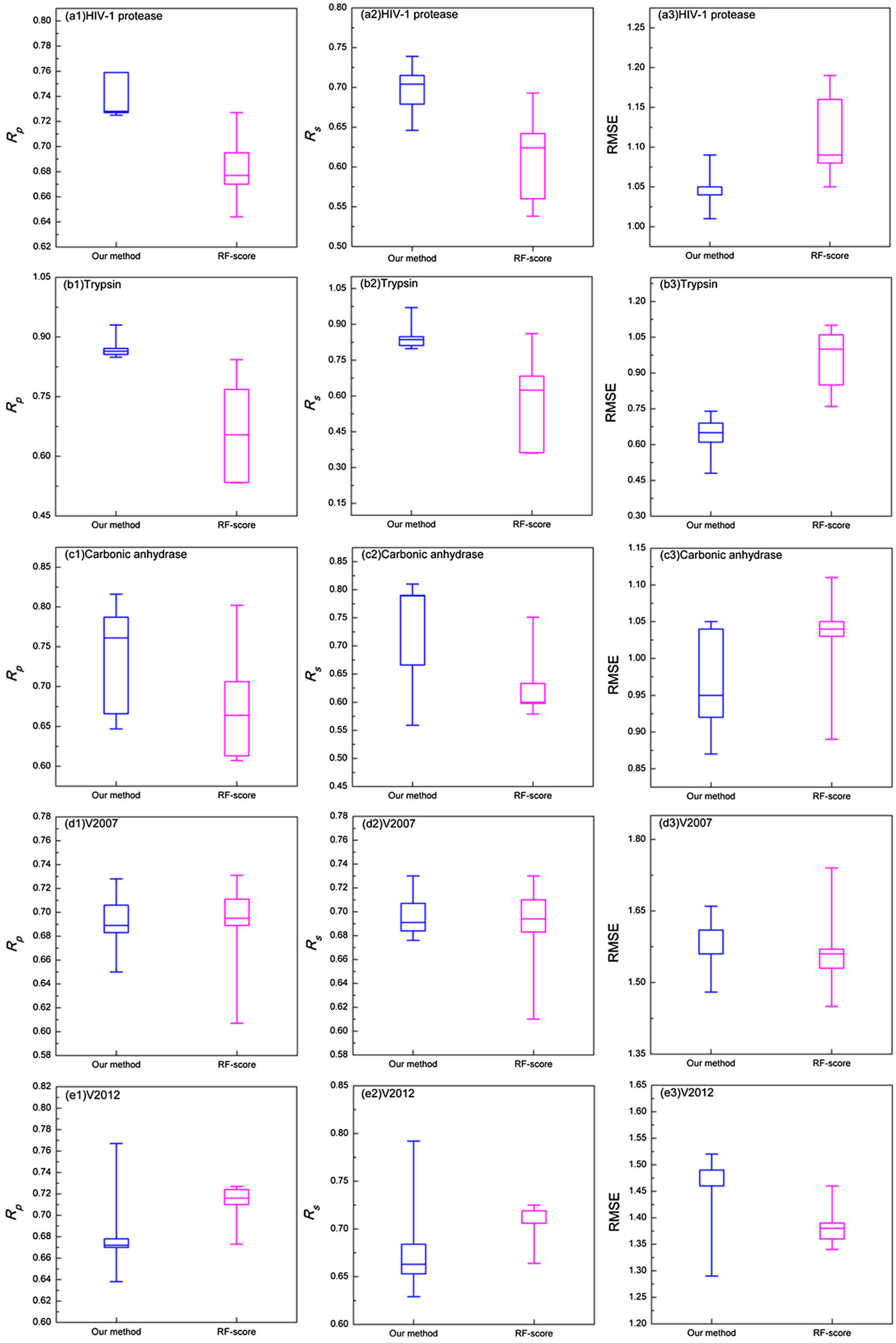

Fig. 4 Box plots of prediction performance of R_p, R_s, and RMSE for five models

Table 3 Performance of our method, RF-score and other 16 scoring functions on the PDBbind benchmark V2007

Scoring function	R_p	R_s	SD
RF-Score::Elem-v2	0.803	0.797	1.54
RF-Score::Elem-v1	0.776	0.762	1.58
Our method	0.772	0.756	1.65
X-Score::HMScore	0.644	0.705	1.83
DrugScore^{CSD}	0.569	0.627	1.96
SYBYL::ChemScore	0.555	0.585	1.98
DS::PLP1	0.545	0.588	2.00
GOLD::ASP	0.534	0.577	2.02
SYBYL::G-Score	0.492	0.536	2.08
DS::LUDI3	0.487	0.478	2.09
DS::LigScore2	0.464	0.507	2.12
GlideScore-XP	0.457	0.435	2.14
DS::PMF	0.445	0.448	2.14
GOLD::ChemScore	0.441	0.452	2.15
SYBYL::D-Score	0.392	0.447	2.19
DS::Jain	0.316	0.346	2.24
GOLD::GoldScore	0.295	0.322	2.29
SYBYL::PMF-Score	0.268	0.273	2.29
SYBYL::F-Score	0.216	0.243	2.35

listed by Cheng et al. are shown in Table 4. The prediction results for the selected classical scoring functions shown in Table 4 were extracted from Cheng et al. [48]. From Table 4, RF-Score and our method do not perform better than other classical scoring functions in every family target and our generic model only gives the best performance on the HIV protease test set, which is consistent with the conclusion by Cheng et al. that the performance of each method is case-dependent because different target protein families have the different intrinsic characteristics. However, the comparison was carried out between only our generic model and other methods. Because the prediction model was trained using functionally and structurally diverse protein–ligand complexes but test on one particular family target, the generic model would weaken the specificity of one particular family target, so the performance of such a model on complexes of this particular protein type would be probably poor. In fact, a satisfactory result has been achieved when we used the family-specific model by our method for each family test set, as shown in Fig. 3. According to the comparison results from Fig. 3 and Table 4, the main conclusion was further addressed that individual representation for each protein family is necessary and it is more reasonable to construct the individual

Table 4 Performance of our method, RF-Score and other selected classical scoring functions on the four family-specific test sets from PDBbind V2007

HIV protease (N = 112)			
Scoring functions	R_s	R_p	SD
Our method	0.522	0.532	1.71
RF-Score	0.393	0.488	1.47
By NHA[a]	0.140	0.172	1.62
A: X-Score::HPScore	0.339	0.341	1.54
B: SYBYL::ChemScore	0.228	0.276	1.58
C: DS::PMF04	0.200	0.183	1.61
D: DrugScore^{PDB}::PairSurf	0.170	0.225	1.60
A + B	0.304		
A + C	0.291		
A + D	0.266		
B + C	0.225		
B + D	0.205		
C + D	0.194		

Trypsin (N = 73)			
Scoring functions	R_s	R_p	SD
Our method	0.573	0.633	1.35
RF-Score	0.712	0.738	1.14
By NHA[a]	0.603	0.655	1.28
A: X-Score::HSScore	0.824	0.817	0.97
B: DS::Ludi2	0.791	0.823	0.96
C: DS::PLP2	0.774	0.797	1.02
D: SYBYL::ChemScore	0.773	0.829	0.95
A + B	0.845		
A + C	0.814		
A + D	0.818		
B + C	0.831		
B + D	0.808		
C + D	0.812		

Carbonic anhydrase (N = 44)			
Scoring functions	R_s	R_p	SD
Our method	0.626	0.677	1.59
RF-Score	0.415	0.646	1.61
By NHA[a]	0.273	0.443	1.25
A: DS::PLP2	0.772	0.800	0.84
B: SYBYL::G-Score	0.646	0.706	0.99

Thrombin (N = 38)			
Scoring functions	R_s	R_p	SD
Our method	0.407	0.580	1.80
RF-Score	0.545	0.619	1.77
By NHA[a]	0.555	0.622	1.66
A: DS::PLP1	0.672	0.692	1.53
B: SYBYL::G-Score	0.626	0.667	1.58

Table 4 continued

Carbonic anhydrase (N = 44)				Thrombin (N = 38)			
Scoring functions	R_s	R_p	SD	Scoring functions	R_s	R_p	SD
C: SYBYL::ChemScore	0.631	0.699	1.00	C: DrugScore^{CSD}::Pair	0.622	0.651	1.61
D: SYBYL::PMF-Score	0.618	0.627	1.09	D: X-Score::HSScore	0.586	0.666	1.58
A + B	0.780			A + B	0.699		
A + C	0.757			A + C	0.653		
A + D	0.763			A + D	0.666		
B + C	0.686			B + C	0.641		
B + D	0.713			B + D	0.601		
C + D	0.735			C + D	0.644		

[a] Using the number of heavy atoms on each ligand as the only variable in correlation

prediction model for a protein family rather than generic model of diverse protein families.

Conclusions

In this study, we developed a machine learning method to predict the binding affinity for both family-specific and generic protein–ligand complexes. A comprehensive characterization covering all aspects of each complex was proposed based on descriptors in four blocks of protein sequence, binding pocket, ligand structure and intermolecular interaction. Compared with the scoring function based methods, the machine learning methods have shown some obvious advantages such as easy implementation, fast prediction process and strong predictive ability. Through feature analysis and evaluation, the important features in different family-specific models are different, which indicate the necessity of individual representation for each protein family. Moreover, the prediction results on the external validation show that family-specific models are far superior to the generic models, because family-specific models take the structural and functional specificity of each target family into account. It is practical to develop specific models to improve the accuracy in binding affinity prediction. Finally, comparisons between our method and RF-score were implemented. Both of them used RF to build the prediction model but with different features. The superior performance of our method on the family-specific models indicates that the four blocks of descriptors are more comprehensive for characterizing the family-specific protein–ligand complexes. The good performance of specific models make us believe that our method can be a useful tool for predicting binding affinity of the three family-specific protein families.

Acknowledgments This work was funded by the National Natural Science Foundation of China (No. 21175095, 21273154, 21375090).

References

1. Coupez B, Lewis RA (2006) Docking and scoring-theoretically easy, practically impossible. Curr Med Chem 13:2995–3003
2. Kroemer RT (2007) Structure-based drug design: docking and scoring. Curr Protein Pept Sci 8:312–328
3. Jain AN (2006) Scoring functions for protein–ligand docking. Curr Protein Pept Sci 7:407–420
4. Li SY, Xi LL, Wang CQ, Li JZ, Lei BL, Liu HX, Yao XJ (2009) A novel method for protein–ligand binding affinity prediction and the related descriptors exploration. J Comput Chem 30: 900–909
5. Betz M, Saxena K, Schwalbe H (2006) Biomolecular NMR: a chaperone to drug discovery. Curr Opin Chem Biol 10:219–225
6. Diercks T, Coles M, Kessler H (2001) Applications of NMR in drug discovery. Curr Opin Chem Biol 5:285–291
7. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A geometric approach to macromolecule-ligand interactions. J Mol Biol 161:269–288
8. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:727–748
9. Naim M, Bhat S, Rankin KN, Dennis S, Chowdhury SF, Siddiqi I, Drabik P, Sulea T, Bayly CI, Jakalian A, Purisima EO (2007) Solvated interaction energy (SIE) for scoring protein–ligand binding affinities. 1. Exploring the parameter space. J Chem Inf Model 47:122–133
10. Aqvist J, Luzhkov VB, Brandsdal BO (2002) Ligand binding affinities from MD simulations. Acc Chem Res 35:358–365
11. Gohlke H, Hendlich M, Klebe G (2000) Knowledge-based scoring function to predict protein–ligand interactions. J Mol Biol 295:337–356
12. Muegge I, Martin YC (1999) A general and fast scoring function for protein–ligand interactions: a simplified potential approach. J Med Chem 42:791–804
13. Muegge I (2006) PMF scoring revisited. J Med Chem 49:5895–5902
14. Zhang C, Liu S, Zhu Q, Zhou Y (2005) A knowledge-based energy function for protein–ligand, protein–protein, and protein-dna complexes. J Med Chem 48:2325–2335
15. Imai T, Hiraoka R, Seto T, Kovalenko A, Hirata F (2007) Three-dimensional distribution function theory for the prediction of protein–ligand binding sites and affinities: application to the binding of noble gases to hen egg-white lysozyme in aqueous solution. J Phys Chem B 111:11585–11591
16. Gehlhaar DK, Verkhivker GM, Rejto PA, Sherman CJ, Fogel DR, Fogel LJ, Freer ST (1995) Molecular recognition of the inhibitor

AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. Chem Biol 2:317–324

17. Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261:470–489
18. Wang R, Lui L, Lai L, Tang Y (1998) Score: a new empirical method for estimating the binding affinity of a protein–ligand complex. J Mol Model 4:379–394
19. Wang R, Lai L, Wang S (2002) Further development and validation of empirical scoring functions for structure-based binding affinity prediction. J comput-Aided Mol Des 16:11–26
20. Chen HM, Liu BF, Huang HL, Hwang SF, Ho SY (2007) SODOCK: swarm optimization for highly flexible protein–ligand docking. J Comput Chem 28:612–623
21. Ballester PJ, Mitchell JBO (2010) A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics 26:1169–1175
22. Smith RD, Dunbar JB, Ung PMU, Esposito EX, Yang CY, Wang S, Carlson HA (2011) CSAR benchmark exercise of 2010: combined evaluation across all submitted scoring functions. J Chem Inf Model 51:2115–2131
23. Sotriffer C, Matter H (2011) The challenge of affinity prediction: scoring functions for structure-based virtual screening. In: Sotriffer C (ed) virtual screening: principles, challenges, and practical guidelines. Wiley-VCH, Weinheim
24. Linusson A, Lindstrom A, Pettersson F, Almqvist F, Berglund A, Kihlberg J (2006) Hierarchical PLS modeling for predicting the binding of a comprehensive set of structurally diverse protein–ligand complexes. J Chem Inf Model 46:1154–1167
25. Zhang S, Golbraikh A, Tropsha A (2006) Development of quantitative structure—binding affinity relationship modelsbased on novel geometrical chemical descriptors of the protein–ligand interfaces. J Med Chem 49:2713–2724
26. Deng W, Breneman C, Embrechts MJ (2004) Predicting protein−ligand binding affinities using novel geometrical descriptors and machine-learning methods. J Chem Inf Comput Sci 44:699–703
27. Zhao YQ, Huang JF (2011) Reconstruction and analysis of human heart-specific metabolic network based on transcriptome and proteome data. Biochem Biophys Res Commun 415:450–454
28. Wang GS, Kearney DL, De Biasi M, Taffet G, Cooper TA (2007) Elevation of RNA-binding protein CUGBP1 is an early event in an inducible heart-specific mouse model of myotonic dystrophy. J Clin Investig 117:2802–2811
29. Lewalle A, Niederer S, Smith N (2014) Species-specific comparison of the cardiac sodium/potassium pump based on a minimal biophysical model. Biophys J 106:117a
30. Heil F, Hemmi H, Hochrein H, Ampenberger F, Kirschning C, Akira S, Lipford G, Wagner H, Bauer S (2004) Species-specific recognition of single-stranded RNA via toll-like receptor 7 and 8. Science 303:1526–1529
31. Xu W, McDonough MC, Erdman DD (2000) Species-specific identification of human adenoviruses by a multiplex PCR assay. J Clin Microbiol 38:4114–4120
32. Saranya N, Selvaraj S (2012) QSAR studies on HIV-1 protease inhibitors using non-linearly transformed descriptors. Curr Comput-Aid Drug 8:10–49
33. Xue MZ, Zheng MY, Xiong B, Li YL, Jiang HL, Shen JK (2010) Knowledge-based scoring functions in drug design. 1. Developing a target-specific method for kinase-ligand interactions. J Chem Inf Model 50:1378–1386
34. Wang R, Fang X, Lu Y, Wang S (2004) The PDBbind database: collection of binding affinities for protein–ligand complexes with known three-dimensional structures. J Med Chem 47:2977–2980
35. Li HJ, Leung KS, Wong MH, Ballester PJ (2014) Substituting random forest for multiple linear regression improves binding affinity prediction of scoring functions: cyscore as a case study. BMC Bioinform 15:291
36. Li ZR, Lin HH, Han LY, Jiang L, Chen X, Chen YZ (2006) PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res 34:W32–W37
37. Liu K, Feng J, Young SS (2005) PowerMV: a software environment for molecular viewing, descriptor generation, data analysis and hit evaluation. J Chem Inf Model 45:515–522
38. Ballester PJ, Schreyer A, Blundell TL (2014) Does a more precise chemical description of protein–ligand complexes lead to more accurate prediction of binding affinity? J Chem Inf Model 54:944–955
39. Moody JE, Hanson SJ, Lippmann RP (1992) Advances in neural information processing systems 4. Morgan Kaufmann, Denver
40. Smith M (1993) Neural networks for statistical modeling. Van Nostrand Reinhold, New York
41. Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, New York
42. Svetnik V (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 43:1947–1958
43. Svetnik V, Liaw A, Tong C, Wang T (2004) Application of Breiman's random forest to modeling structure–activity relationships of pharmaceutical molecules. In: Roli F, Kittler J, Windeatt T (eds) Lecture notes in computer science, vol 3077. Springer, Berlin, pp 334–343
44. Polishchuk PG, Muratov EN, Artemenko AG, Kolumbin OG, Muratov NN, Kuz'min VE (2009) Application of random forest approach to QSAR prediction of aquatic toxicity. J Chem Inf Model 49:2481–2488
45. Core Team R (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
46. Breiman L (1996) Out-of-bag estimation. Technical report, UC Berkeley
47. Hastie T, Tibshirani R, Friedman J (2003) The elements of statistical learning. Springer, NewYork
48. Cheng TJ, Li X, Li Y, Liu ZH, Wang RX (2009) Comparative assessment of scoring functions on a diverse test set. J Chem Inf Model 49:1079–1093