Comparison of Cutoff Strategies for Geometrical Features in Machine Learning-Based Scoring Functions

Shirley W.I. Siu, Thomas K.F. Wong, and Simon Fong

Department of Computer and Information Science
University of Macau
Macau, China
{shirleysiu,mb15404,ccfong}@umac.mo

Abstract. Countings of protein-ligand contacts are popular geometrical features in scoring functions for structure-based drug design. When extracting features, cutoff values are used to define the range of distances within which a protein-ligand atom pair is considered as in contact. But effects of the number of ranges and the choice of cutoff values on the predictive ability of scoring functions are unclear. Here, we compare five cutoff strategies (one-, two-, three-, six-range and soft boundary) with four machine learning methods. Prediction models are constructed using the latest PDBbind v2012 data sets and assessed by correlation coefficients. Our results show that the optimal one-range cutoff value lies between 6 and 8 Å instead of the customary choice of 12 Å. In general, two-range models have improved predictive performance in correlation coefficients by 3-5%, but introducing more cutoff ranges do not always help improving the prediction accuracy.

Keywords: scoring function, protein-ligand binding affinity, geometrical features, machine learning, structure-based drug design.

1 Introduction

With the advances in biophysical experiments in recent years, the amount of known molecular structures has been increased rapidly. Fast and accurate structure-based computational methods to identify putative drug molecules from a large database of small ligand molecules become a crucial step in modern drug discovery [1]. In structure-based drug design (SBDD), the preferred conformation of a ligand molecule in the active site of the target protein is predicted first by a docking algorithm, then the biological activity of the protein-ligand complex in terms of binding constant or binding free energy is estimated using a scoring function [2]. While current docking algorithms are able to generate docked conformations reasonably close to the native complexes, the problem lies in the difficulty to accurately predict the binding affinities of the docked complexes in order to distinguish the active ligands from decoys. In addition, highly

M. Yao et al. (Eds.): ADMA 2013, Part II, LNAI 8347, pp. 336–347, 2013.

accurate scoring functions are essential for lead optimization in the later stage of the drug discovery process.

Despite years of effort, the performance of scoring functions is still far from satisfactory. A recent comparative assessment of scoring functions on a benchmark data set by Cheng et al. [3] has shown that the "ranking power" of the top scoring functions have merely 50% success rate and the correlation coefficients between experimental binding values and predicted scores (the so-called "scoring power") ranged from 0.545 to 0.644 only. When applying an updated list of scoring functions to a new validation data set by Huang et al., the same conclusion was obtained [2]. Both Cheng and Huang's studies highlight the need for new scoring functions with improved prediction accuracy and higher reliability.

One promising alternative to the conventional scoring functions is to apply machine learning (ML) algorithms to construct models for protein-ligand binding prediction. Since ML algorithms learn the theory directly from data through the process of fitting and inference without prior assumption of the statistical distribution in the data, ML is ideally suited for handling biological data that is noisy, complex, and lack of comprehensive theory. Only in the last two years, studies applying ML techniques to construct scoring functions have been seen to emerge. Ballester and Mitchell trained a random forest model (RF-Score) from simple geometrical features of protein-ligand atom-type pairs [5]. Li et al. applied support vector regression (SVR) techniques to learn features derived from knowledge-based pairwise potentials (SVR-KB) and physicochemical properties (SVR-EP) [6]. Another scoring function (NNScore) combined energetic terms from the popular AutoDock Vina program and the BINANA binding characteristics [7] to construct a single-hidden-layer neural network [8]. Recently, an extensive assessment of ML-based scoring functions was carried out by Ashtawy and Mahapatra in which five ML techniques, a combination of three sets of features, and in total 42 different prediction models were constructed [4]. By comparing these ML-based scoring functions to the conventional scoring functions, they showed that the ranking power of the best ML model reaches 62.5% success rate whereas the best conventional model has only 57.8% [4]. Their study is a valuable proof-of-concept that ML is the method of choice for the next generation scoring functions for SBDD. Innovative ML-based scoring functions can be produced by applying new ML algorithms and feature selection methods, or by combining a number of independent ML-models to create consensus scoring functions.

The success of these ML-based scoring functions relies on the correct choice of structural or physiochemical features which can capture the patterns of binding interactions between protein and ligand molecules. In particular, geometrical features such as occurrence count of element pairs are commonly adopted in scoring functions [4,5,6,7]. However, a geometrical feature usually requires a predefined cutoff value to distinguish "interacting" from "non-interacting" atom pairs by distances of the pairs. Often, this value seems to be chosen quite arbitrarily without clear justification. For example, a cutoff of 2.5 Å was used in the BINANA algorithm [7], whereas a cutoff of 12 Å was used by RF-Score [5] and Ashtawys

ML-models adopting the RF-Score features [4]. Kramer and Gedeck defines six cutoff values (3.0, 3.5, 4.0, 4.5, 6.0, 12.0 Å) to bin contact counts by distances of atom-type pairs and ignores all pairs falling out of 12 Å distance. Finer binning strategy is proposed by Hsu et al., where they use 10 cutoff distances (2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5, 7.0 Å) to count the total number of protein-ligand interactions of vdW force and another 10 cutoff distances (2.5, 2.6, 2.7, 2.8, 2.9, 3.0, 3.1, 3.2, 3.3, 3.4 Å) for hydrogen-bonding and electrostatic interactions [18].

To investigate the predictive abilities of different cutoff strategies in scoring functions, here we compare prediction models using features generated from six cutoff strategies used in literatures, we called them one-range, two-range, three-range, six-range, and soft boundary. Four popular ML techniques – random forests, weighted k-nearest neighbors, support vector machine, and multiple linear regression – are employed to construct in total 24 scoring functions to predict protein-ligand binding affinity. Finally, the best scoring function obtained in this study will be compared to existing conventional and ML-based scoring functions.

2 Materials and Methods

2.1 Data Sets

Data sets used in this study were obtained from the PDBbind database [10]. This manually curated database provides a collection of PDB structures of protein-ligand complexes with experimentally measured binding affinities. Each release of the database contains three data sets: The *general set* is the collection of all binding data in the database; the *refined set* is a selection of high resolution data of 2.5 Å or higher from the general set; the *core set* is a collection of complexes with the highest, medium, and lowest binding affinities from each cluster of protein structures in the refined set. The latest PDBbind v2012 database contains 2,897 complexes in the refined set and 201 complexes in the core set. In this study, the refined set with the core set data removed was used as the training data and the core set was used as the test data.

To compare the ML-based methods to existing scoring functions, we also trained our models using the PDBbind v2007 database, which was used as benchmark data in two recent comparative assessments of scoring functions [3,4].

2.2 Features

Occurrence counts of element pairs were shown to be powerful in predicting protein-ligand binding affinities [5]. Such features are straightforward to compute and are commonly used in combination with other energy or physiochemical features to build prediction models. The determinant to the predictive strength of these geometrical features is the distance criterion used to generate the feature data. For example, an one-range strategy with a single cutoff of 12.0 Å was used in RF-Score [5]; any pair with a distance beyond the cutoff was ignored.

Similarly, a one-range strategy with cutoff at 4.0 Å was adopted in NNScore [8]. A six-range strategy was chosen in Kramer's work; they suggested that different bin sizes should be used for close and distal interactions [17]. It is generally believed that the more the number of distance ranges to use in a model, the higher prediction accuracy would be achieved. However, one additional range introduced in the model will likely to increase the total number of features by double (if no feature selection is conducted), yet the gain in performance may be merely nominal. Therefore, it is instructive to compare the different cutoff strategies and to find out the optimal distance cutoffs which could maximize the predictive performance but minimize the number of required features.

In this work, we are going to investigate one-range, two-range, and three-range cutoff strategies systematically, and compare them to two other cutoff strategies proposed by Kramer [17] and Ouyang [9]. These cutoff strategies are defined as follows:

The one-range strategy counts the total number of a protein element that comes within a cutoff distance (d_0) of a ligand element. Each element-pair is one feature. Using a similar formulation to [5], the occurrence count for a protein-ligand element-pair x_{PL} is:

$$x_{PL} = \sum_{i=1}^{N_P} \sum_{j=1}^{N_L} \theta(d_0 - d_{ij}) \ , \tag{1}$$

where N_P is the number of protein atom which is an element P and N_L is the number of ligand atom which is an element L. d_{ij} is the distance between the i^{th} protein atom and the j^{th} ligand atom of the types in consideration. θ is the unit step function which returns 1 if the argument is positive, and zero if otherwise.

The two- and three-range cutoff strategies introduce more cutoff thresholds such that counts of protein-ligand atoms in different distance ranges are tallied separately. For example, in the two-range cutoff strategy, two cutoffs d_0 and d_1 are defined and so two features x^0_{PL} and x^1_{PL} are created for each element-pair:

$$x^0_{PL} = \sum_{i=1}^{N_P} \sum_{j=1}^{N_L} \theta(d_0 - d_{ij}) \tag{2}$$

and

$$x^1_{PL} = \sum_{i=1}^{N_P} \sum_{j=1}^{N_L} \theta(d_1 - d_{ij}) \times \theta(d_{ij} - d_0) \ . \tag{3}$$

For each of the cutoff strategies, the optimal cutoff value(s) was determined by performing a systematic search over the range of possible values during model construction (see below).

We also tested the six-range cutoff strategy suggested by Kramer [17] and the cutoff strategy with soft boundary from Ouyang [9]. The former uses six thresholds of 3.0, 3.5, 4.0, 4.5, 6.0, and 12.0 Å. For the latter, instead of simply

counting the protein-ligand element pairs, distance-based functions are used to convert each count into two separate contributions, namely repulsion and attraction. Also, instead of a sharp cutoff, a soft threshold for each element pair is determined from their van der Waals radii and a tailing function is introduced at the threshold boundary to account for the reduced intermolecular interaction contribution at large distances.

In this work, nine element types (C, N, O, F, P, S, Cl, Br, and I) are used for both protein and ligand. Therefore, in total there are 81, 162, 243 features for the one-, two-, three-range strategies respectively, 486 features for Kramer's six-range strategy, and 162 features for Ouyang's strategy. Nevertheless, because certain element types (such as F, P, Cl, Br, I) are rare in protein molecules, the maximum number of features for a complex is reduced to 36, 72, 98 for the one-, two-, three-range strategies, and 205, 72 for Kramer's strategy and Ouyang's strategy, respectively.

2.3 Scoring Functions

To compare the predictive performance of different cutoff strategies in scoring functions, we applied four machine learning (ML) techniques popularly used in bioinformatics applications to create prediction models. These ML techniques include random forests (RF) [12], support vector machine (SVM), weighted k-nearest neighbors (wkNN) [13], and multiple linear regression (MLR). All models were trained to predict the pK value.

To compare fairly all prediction models, each model was tuned to have the optimal parameters. The tuning procedure is as follows: For RF, the parameters to be evaluated include the number of trees to grow (*ntree*) and the number of features randomly sampled at each nodal split (*mtry*). Using the training data, ten RF models were generated for each *ntree* value between 500 and 8000 in 500-interval (using the default $mtry = p/3$ where p is the number of features). The optimal *ntree* was determined as the one with the minimum averaged out-of-bag (OOB) error. After deciding the value of *ntree*, similar procedure was conducted to search for the optimal *mtry* value in the range of 1 to the maximum number of features. For SVM models, we used the radial basis function (RBF) as the kernel function. Two parameters – the cost C and the width of the kernel γ – were optimized using grid search. Values which gave the lowest mean squared error in ten-fold cross validation were chosen. For weighted kNN, optimal value for the number of neighbors to consider (k) and the kernel function to convert distances into weights were determined using ten-fold cross validation. The distance metric to use was the Manhattan distance which was found to give better results than Euclidean distance in all cases. Finally, in MLR algorithm, a generalized linear model was fitted to obtain a vector of weights for the input features. Again, ten-fold cross validation was used to evaluate the model.

Training and testing of all prediction models were carried out using the statistical package R [16].

2.4 Performance Metrics

Commonly used metrics to assess the performance of a scoring function are the Pearson's (R_P), Spearman's (R_S) and Kendall's (R_K) correlation coefficients, and root mean squared error ($RMSE$). R_p measures the linear dependence between the predicted binding affinities and the experimental binding affinities:

$$R_p = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}} \ , \tag{4}$$

where y_i and $\hat{y}_i$ represent the experimental and the predicted binding affinities of sample i. The total number of samples is denoted by N.

R_S is the widely used metric for ranking correlation, i.e. it carries out on the ranks of the values rather than the values themselves. In $\mathrm{R_S}$, it compares the position of a sample when ranked by the predicted binding affinity to its position when ranked by the experimental value:

$$R_s = 1 - \frac{6\sum_{i=1}^{N}(d_i)^2}{N(N^2-1)} \ , \tag{5}$$

where d_i is the difference in the two ranks for sample i.

R_K is another ranking correlation. It measures the similarity of the two rankings by counting the total number of concordant C (when the order of two samples in the predicted ranking agrees with the order in the experimental ranking) and discordant pairs D (when it disagrees):

$$R_K = \frac{2(C-D)}{N(N-1)} \ . \tag{6}$$

To measure how well a scoring function predicts the absolute value of the binding affinity, the mean squared error (MSE) or the square root of MSE (RMSE) is used:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i}(y_i - \hat{y}_i)^2 \ . \tag{7}$$

3 Results and Discussion

3.1 ML-Models Using Different Cutoff Strategies

One-range cutoff strategy is the simplest strategy where only one cutoff value is used. We built one-range prediction models with cutoff values ranging from 3 to 30 Å and calculated MSE as a function of different cutoffs. As shown in Fig. 1, all models show a change of MSE in a consistent manner: The MSEs are large for models using small cutoff values (typically less than 5 Å). Between 5 and 13 Å, the MSEs reach the minima and vary slightly. Beyond 13 Å, there is a slow but distinctive trend of increase in MSEs. The optimal cutoff for each ML algorithm can be identified as the one with the lowest MSE. Comparison of four

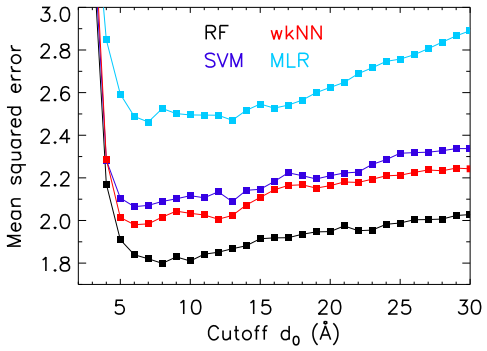

Fig. 1. The mean squared error (MSE) analysis of prediction models using one-range cutoff strategy

ML models on the test data is reported in Table 1. The best scoring function using the one-range cutoff strategy is RF with optimal cutoff at 8 Å. It achieves a correlation coefficient R_P of 0.703, ranking correlations R_S of 0.692 and R_K of 0.504, and RMSE of 1.803.

In the two-range strategy, element-pair counts for short-range interactions are separated from long-range interactions, so two cutoff values d_0 (for the short-range) and d_1 (for the long-range) are required. To find these, we tested prediction models with d_1 between 8 and 14 Å and d_0 between 3 and $(d_1 - 1)$ Å. The result from internal validation is shown in Fig. 2. As expected, all two-range models (color lines) demonstrate improved performances by yielding lower MSEs (except for two cases in MLR) comparing to one-range models (black lines). The optimal cutoff values found in all models are in the ranges of 4-7 Å for d_0 and 11-14 Å for d_1. Again, the RF model achieves the lowest MSE at $d_0 = 7.0$ Å and $d_1 = 11.0$ Å, which attains the correlation coefficient R_P of 0.727, ranking correlations R_S of 0.718 and R_K of 0.527. Compared to one-range models, the predictive performance of two-range models measured by correlation coefficients is increased by 3-5%.

In the three-range strategy, three distance cutoffs are used to separately binning the counts of short-range, mid-range, and long-range interactions. To construct a three-range model, we based on the optimal two-range models to evaluate the addition of a third cutoff threshold in the range of 3 to $(d_1 - 1)$ Å using internal validation procedure. As shown in Table 1, the optimal three-range cutoffs found for the ML-based models are 3-4 Å for d_0, 6-7 Å for d_1, 11-14 Å for d_2. Interestingly, while there is little or no improvement for prediction models using RF, wkNN, and SVM algorithms, MLR has an increase of 5-7% in correlation coefficients compared to its two-range model.

We have also tested a six-range cutoff strategy used by Kramer as part of the descriptor set in their protein-ligand binding scoring function [17]. In this strategy, a smaller bin width is used for counting the short-range interaction and a larger bin width for the long-range interaction. It should be pointed out that

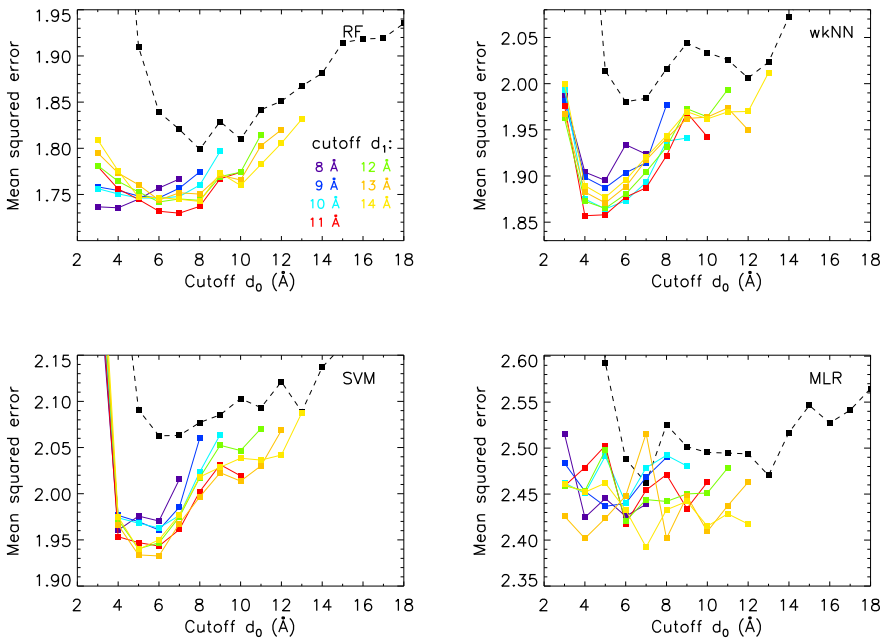

Fig. 2. The mean squared error (MSE) analysis of prediction models using two-range cutoff strategy. Cutoff d_1 varies from 8 to 14 Å and cutoff d_0 from 3 to $d_1 - 1$. Results of one-range cutoff strategy are also shown in black dashed lines for reference.

Kramer assigned atom types to the protein and ligand atoms using Crippen atom typing scheme whereas no typing scheme is applied here. Our purpose is to test if the introduction of more cutoff values would improve the prediction, therefore, all four ML algorithms were applied on six-range features generated using Kramer's cutoff strategy. Except a small improvement of 2% of the wkNN model (compared to its two-range model), in general Kramer's models give worse performance than models using one-range, two-range and three-range cutoff strategies.

The final strategy we tested is the soft threshold method introduced by Ouyang et al. [9]. Unlike the aforementioned strategies where the same cutoff values are used for binning the counts regardless of atom types of the interacting pair, Ouyang introduces a specified cutoff distance for each protein-ligand element pair which is calculated as the sum of their van der Waals radii. This specified cutoff distance defines distance ranges in which the contribution of the pairwise occurrence will be counted in full (a value of 1) or partially (between 0 and 1) as a function of the measured distance. Two values are computed from the so-called membership functions representing the pair's contribution to the repulsive component and the attractive component of the total protein-ligand binding energy. Tailing functions are introduced at the cutoff boundaries such that the repulsive and attractive function at the upper boundary goes smoothly from 1 to 0, and the attractive function at the lower boundary goes from 0 to 1.

Table 1. Performance Comparison of Different Scoring Functions Against the PDBbind v2012 Test Set

Cutoff strategy	ML	Optimal cutoff (Å)	R_P	R_S	R_K	RMSE
One-range	RF	8.0	0.703	0.692	0.504	1.803
	wkNN	6.0	0.691	0.671	0.491	1.778
	SVM	6.0	0.674	0.668	0.479	1.831
	MLR	7.0	0.577	0.587	0.410	2.002
Two-range	RF	7.0, 11.0	0.727	0.718	0.527	1.759
	wkNN	4.0, 11.0	0.682	0.670	0.482	1.796
	SVM	6.0, 13.0	0.676	0.674	0.481	1.802
	MLR	7.0, 14.0	0.582	0.596	0.416	1.990
Three-range	RF	3.0, 7.0, 11.0	0.728	0.720	0.524	1.760
	wkNN	4.0, 7.0, 11.0	0.693	0.681	0.499	1.768
	SVM	4.0, 6.0, 13.0	0.655	0.659	0.474	1.831
	MLR	3.0, 7.0, 14.0	0.611	0.634	0.446	1.942
Kramer	RF	3.0, 3.5, 4.0, 4.5, 6.0, 12.0	0.706	0.700	0.506	1.798
	wkNN		0.690	0.671	0.493	1.769
	SVM		0.652	0.657	0.475	1.860
	MLR		0.556	0.578	0.413	2.013
Ouyang	RF	–	0.717	0.712	0.517	1.771
	wkNN	–	0.669	0.658	0.476	1.811
	SVM	–	0.684	0.688	0.494	1.785
	MLR	–	0.563	0.595	0.419	2.016

Optimal parameters for the ML-based scoring functions are as follows: For RF models, ntree/mtry values are 3000/8 for one-range, 3000/13 for two-range, 4000/15 for three-range, and 3000/55 for Kramer. For wkNN models, triangular kernel is the best kernel in all cases when using Manhattan distance. The optimal k is 11. For SVM models, fine-tuned gamma/cost for radial basis kernel are 0.25/1.414214 for one-range, 0.08838835/2 for two-range, 0.04419417/2.828427 for three-range, 0.015625/2 for Kramer, 0.0625/2 for Ouyang.

To compare how this strategy performs with other cutoff strategies, all four ML algorithms were applied on Ouyang's feature data. The result is also reported in Table 1. Compared to the two-range models which have the same number of features, Ouyang models perform 1-3% worse in correlation coefficients using RF, wkNN, and MLR algorithms, but slightly better using SVM.

Overall, results of the comparative assessment of six different cutoff strategies in this section indicate that the predictive performance of one-range, two-range, and three-range models are within 5% of one another. Introducing more cutoffs with the purpose of attaining finer resolution for contact counts is shown to be unnecessary since the increased model complexity results in poorer predictions. Also, the use of soft boundary to avoid the sharp cutoff do not improve performance significantly. Among models using different ML methods, RF models always outperform other ML models by 1-7%. Therefore, taken into account both the predictive ability and model complexity, we consider the two-range RF model to be the best binding affinity prediction model obtained in this study.

3.2 Comparison with State-of-the-Art Models

To the best of our knowledge, no existing scoring functions have been tested with the new data sets employed in this work. Nevertheless, it is interesting to know how our scoring function is compared to existing conventional and ML-based scoring functions. To this end, two conventional scoring functions, X-Score [14] and DSX [15], were selected for comparison. They are chosen because of their outstanding performances reported in [3]. Programs of these functions are freely available [14] and can be applied directly to predict binding affinities in the test data. As shown in Table 2, the two-range RF outperforms X-Score by 21-25%, 14-29% and 17-35%, in R_P, R_S and R_K, respectively. In recent years, ML-based scoring functions have been showed significant improvements over conventional scoring functions on the previous version of the PDBbind database. Here, we trained our two-range RF scoring function using the PDBbind v2007 refined data set and tested on the PDBbind v2007 core data set. With this, the model can be compared directly to ML-based scoring functions developed with these data sets. These include RF-Score [5], CScore [9], and the five scoring functions from [4] (RF::XA, BRT::AR, SVM::XAR, kNN::XR, MLR::XR, MARS::AR).

Table 2. Performance Comparison of Conventional Scoring Functions Against the PDBbind v2012 Test Set

Scoring function	R_P	R_S	R_K	RMSE
X-Score (HPScore)	0.595	0.625	0.448	1.987
X-Score (HMScore)	0.598	0.631	0.451	1.956
X-Score (HSScore)	0.582	0.604	0.431	2.019
X-Score (AvgScore)	0.600	0.627	0.449	1.969
DSX (PDB)	–	0.594	0.421	–
DSX (Pharm)	–	0.557	0.390	–
Two-range RF (this work)	0.727	0.718	0.527	1.759

Since DSX predicts a score for a protein-ligand complex instead of the binding affinity, R_P and RMSE values of DSX scoring functions were not assessed.

Table 3. Performance Comparison of Existing ML-based Scoring Functions Against the PDBbind v2007 Test Set

Scoring function	R_P	R_S	Ref
CScore	0.801	–	[9]
BRT::AR	0.793	0.782	[4]
Two-range RF	0.787	0.777	this work
RF::XA	0.777	0.771	[4]
RF-Score	0.776	0.762	[5]
SVM::XAR	0.768	0.792	[5]
kNN::XR	0.727	0.720	[5]
MLR::XR	0.641	0.731	[5]
MARS::AR	0.681	0.665	[5]

Parameters for two-range RF: ntree=3000, mtry=13. Results of other scoring functions are taken from literatures.

As shown in Table 3, the performance of the two-range RF model using simple geometrical features is comparable to the top ranked scoring functions using combination of different physiochemical features, and it is ranked as the third best.

4 Conclusion

The development of scoring functions to accurately predict binding affinities of protein-ligand complexes is a daunting task in structure-based drug design. The problem lies in the selection of predictive geometrical or physiochemical features to capture patterns of binding interactions. Among the geometrical features, counting of protein-ligand atomic contacts is a simple yet effective choice as shown in the work of Ballester and Mitchell [5] where only a single cutoff of 12 Å is used to extract contact counts. Numerous cutoff strategies have been employed in other works but used as part of the feature set, so the predictive performance of these cutoff strategies is unclear. In this study, our aim is to compare the predictive abilities of models using different cutoff strategies, namely one-range, two-range, three-range, Kramer's six-range, and the cutoff with soft boundary from Ouyang, and to find out the optimal cutoff values for binding affinity prediction. Prediction models were constructed using four state-of-the-art ML techniques (RF, wkNN, SVM, MLR) with the latest PDBbind v2012 data sets and the models were assessed by three correlation coefficients and RMSD. Our results show that the optimal one-range cutoff value lies between 6 and 8 Å instead of the customary choice of 12 Å . In general, two-range models have improved predictive performance of 3-5% as measured by correlation coefficients, but introducing additional cutoff ranges (three-, six-range) do not always help improving the prediction accuracy. We also show that the two-range RF model (the best model in this work) is able to outperform the best conventional scoring functions and performs comparably to other top-ranked ML-based scoring functions against the PDBbind v2007 benchmark data set. Results of this work are helpful to the selection of geometrical features and cutoff values in the development of scoring functions for structure-based drug design.

Acknowledgments. The authors would like to thank the Information and Communication Technology Office of the University of Macau (UM) for their support of high performance computing facilities. This work is funded by the Research and Development Administration Office of UM (grant SRG022-FST13-SWI).

References

1. Kitchen, D.B., Decornez, H., Furr, J.R., Bajorath, J.: Docking and scoring in virtual screening for drug discovery: Methods and applications. Nat. Rev. 3, 935–949 (2004)
2. Huang, S.Y., Grinter, S.Z., Zou, X.: Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. Phys. Chem. Chem. Phys. 12, 12899–12908 (2010)

3. Cheng, T., Li, X., Li, Y., Liu, Z., Wang, R.: Comparative assessment of scoring functions on a diverse test set. J. Chem. Inf. Model. 49, 1079–1093 (2009)
4. Ashtawy, H.M., Mahapatra, N.R.: A comparative assessment of ranking accuracies of conventional and machine-learning-based scoring functions for protein-ligand binding affinity prediction. IEEE/ACM Trans. Comput. Biol. Bioinf. 9, 1301–1312 (2012)
5. Ballester, P.J., Mitchell, J.B.O.: A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. Bioinf. 26, 1169–1175 (2010)
6. Li, L., Wang, B., Meroueh, S.O.: Support vector regression scoring of receptor-ligand complexes for rank-ordering and virtual screening of chemical libraries. J. Chem. Inf. Model. 51, 2132–2138 (2011)
7. Durrant, J.D., McCammon, J.A.: BINANA: A novel algorithm for ligand-binding characterization. J. Mol. Graphics. Modell. 29, 888–893 (2011)
8. Durrant, J.D., Mc Cammon, J.A.: NNScore 2.0: A neural-network receptor-ligand scoring function. J. Chem. Inf. Model. 51, 2897–2903 (2011)
9. Ouyang, X., Handoko, S.D., Kwoh, C.K.: CScore: A simple yet effective scoring function for protein-ligand binding affinity prediction using modified CMAC learning architecture. J. Bioinf. Comput. Biol. 9, 1–14 (2011)
10. Wang, R., Fang, X., Lu, Y., Wang, S.: The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. J. Med. Chem. 47, 2977–2980 (2004)
11. Muegge, I., Martin, Y.C.: A general and fast scoring function for protein-ligand interactions: a simplified potential approach. J. Med. Chem. 42, 791–804 (1999)
12. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)
13. Hechenbichler, K., Schliep, K.: Weighted k-nearest-neighbor techniques and ordinal classification. Discussion paper 399, SFB 386 (2004)
14. Wang, R., Lai, L., Wang, S.: Further development and validation of empirical scoring functions for structure-based binding affinity prediction. J. Comput.-Aided Mol. Des. 16, 11–26 (2002), The program X-Score v1.2, `http://sw16.im.med.umich.edu/software/xtool` (August 2013)
15. Neudert, G., Klebe, G.: DSX: a knowledge-based scoring function for the assessment of protein-ligand complexes. J. Chem. Inf. Model. 51, 2731–2745 (2011), The program DSX 0.89, `http://pc1664.pharmazie.uni-marburg.de/drugscore/dsx_download.php` (August 2013)
16. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2012)
17. Kramer, C., Gedeck, P.: Global free energy scoring functions based on distance-dependend atom-type pair descriptors. J. Chem. Inf. Model. 51, 707–720 (2011)
18. Hsu, K.-C., Chen, Y.-F., Yang, J.-M.: GemAffinity: a scoring function for predicting binding affinity and virtual screening. Int. J. Data Mining and Bioinformatics 6, 27–41 (2012)