Supervised Consensus Scoring for Docking and Virtual Screening

Reiji Teramoto* and Hiroaki Fukunishi

Fundamental and Environmental Research Laboratories, NEC Corporation, 34, Miyukigaoka, Tsukuba, Ibaraki 305-8501, Japan

Received November 8, 2006

Docking programs are widely used to discover novel ligands efficiently and can predict protein−ligand complex structures with reasonable accuracy and speed. However, there is an emerging demand for better performance from the scoring methods. Consensus scoring (CS) methods improve the performance by compensating for the deficiencies of each scoring function. However, conventional CS and existing scoring functions have the same problems, such as a lack of protein flexibility, inadequate treatment of salvation, and the simplistic nature of the energy function used. Although there are many problems in current scoring functions, we focus our attention on the incorporation of unbound ligand conformations. To address this problem, we propose supervised consensus scoring (SCS), which takes into account protein−ligand binding process using unbound ligand conformations with supervised learning. An evaluation of docking accuracy for 100 diverse protein−ligand complexes shows that SCS outperforms both CS and 11 scoring functions (PLP, F-Score, LigScore, DrugScore, LUDI, X-Score, AutoDock, PMF, G-Score, ChemScore, and D-score). The success rates of SCS range from 89% to 91% in the range of rmsd < 2 Å, while those of CS range from 80% to 85%, and those of the scoring functions range from 26% to 76%. Moreover, we also introduce a method for judging whether a compound is active or inactive with the appropriate criterion for virtual screening. SCS performs quite well in docking accuracy and is presumably useful for screening large-scale compound databases before predicting binding affinity.

1. INTRODUCTION

Protein−ligand docking is widely used to discover novel ligands efficiently in structure-based drug design. Over the past 15 years, various docking programs have been developed, and their performance has been evaluated in detail.[1−20] These docking programs attempt to predict the binding conformation of a ligand and the protein−ligand binding affinity. They involve two computational steps: docking and scoring. In the docking step, many ligand conformations are generated. There are several conformation sampling methods, such as genetic algorithms, Monte Carlo simulation, and simulated annealing. All sampling methods are guided by a function that evaluates the fitness between the protein and ligand. In the scoring step, a scoring function is used to evaluate the protein−ligand affinity. The scoring functions are important, because the final predicted conformations are selected according to the scores.

There are three groups of scoring functions: force-field-based methods, empirical scoring functions, and knowledge-based potentials. Force-field-based scoring functions apply classical molecular mechanics energy functions. They approximate the binding free energy of protein−ligand complexes by a sum of van der Waals and electrostatic interactions. Solvation is usually taken into account using a distance-dependent dielectric function, although solvent models based on continuum electrostatics have been developed.[21,22] Nonpolar contributions are usually assumed to be proportional to the solvent-accessible surface area. A drawback is that the energy landscapes associated with force-field potentials are generally rugged, and, therefore, minimization is required prior to any energy evaluation.

Empirical scoring functions estimate the binding free energy by summing interaction terms derived form weighted structural parameters. The weights are determined by fitting the scoring function to experimental binding constants of a training set of protein−ligand complexes. The main drawback is that it is unclear whether they are able to predict the binding affinity of ligand structurally different from those used in the training sets.

Knowledge-based scoring functions represent the binding affinity as a sum of protein−ligand atom pair interactions. Those potentials are derived from the protein−ligand complexes with known structures, where probability distributions of interatomic distances are converted into distance-dependent interaction free energies of protein−ligand atom pairs. However, 3D structures of protein−ligand complexes do not provide a thermodynamic ensemble at equilibrium, and, therefore, a knowledge-based potential should be considered as s statistical preference rather than a potential of mean force. A key ingredient of a knowledge-based potential is the reference state, which determines the weights between the various probability distributions. Before now, several approaches to derive these potentials have been proposed.[23−26] They differ in their definition of the reference state, the protein and ligand atom types, and the list of protein−ligand complexes from which they were extracted.

Although there has been some success in designing scoring functions that can describe protein−ligand interactions, there is the limitation in current docking and scoring, such as a

* Corresponding author phone: +81 298 850 1410; fax: +81 298 856 6136; e-mail: r-teramoto@bq.jp.nec.com.

10.1021/ci6004993 CCC: $37.00
Published on Web 02/13/2007

lack of protein flexibility, inadequate treatment of salvation, and the simplistic nature of the energy function used. In particular, it has been pointed out that the major weakness of docking programs lies in the scoring functions.[27,28]

Recently, it has been reported that consensus scoring (CS) improves performance by compensating for the deficiencies of each scoring function.[28−34] However, CS strategies, such as rank-by-rank, rank-by-number, average rank, and the linear combination of multiple scoring functions, do not perform better than the best scoring function among those used for CS.[28] In addition, several known active compounds are required in order to determine the best combination of scoring functions. Furthermore, CS and individual scoring functions have the same problems which we mentioned above.

Although there are many problems in current scoring functions, we focus attention on the incorporation of unbound ligand conformations. To address this problem, we propose supervised CS method (SCS), which takes into account the protein−ligand binding process using unbounded ligand conformations with supervised learning. The concept of SCS is based on the free energy landscape of the protein−ligand binding process, and supervised learning is used for learning the protein−ligand binding process using unbound ligand conformations. To demonstrate the effectiveness of SCS, we tested 100 diverse protein−ligand complexes. The performance of scoring methods and dependency to binding affinity was evaluated on the basis of the reproducibility of the pose of X-ray structures, i.e., the success rate and average root-mean-square deviation (rmsd). Moreover, we also introduce a procedure for determining the appropriate criterion for judging active or inactive compounds for virtual screening of large-scale compound databases.

2. METHODS

2.1. Preparation of Data Sets. The 100 protein−ligand complexes used in this study are listed in Table 1. In each complex, 100 decoys ware generated via AutoDock.[28] All decoys of each ligand were scored using 11 scoring functions: AutoDock, G-Score, D-Score, LigScore, PLP, LUDI, F-Score, ChemScore and X-Score, PMF, and DrugScore. These scoring functions can be grouped into three types: (1) force field (AutoDock, G-Score, and D-Score); (2) empirical scoring functions (LigScore, PLP, LUDI, F-Score, ChemScore, and X-Score); and (3) knowledge-based potentials (PMF and DrugScore). Detailed descriptions of the data-generation procedure and all of the data sets are available online at http://sw16.im.med.umich.edu/software/xtool.

2.2. Supervised Consensus Scoring (SCS). The overall SCS procedure is illustrated in Figure 1, and the basic concept is illustrated in Figure 2. In SCS, the binding pose prediction problem is formulated as supervised learning in which explanatory attributes and an objective variable are the scores and the rmsd between decoys and X-ray structures for ligands, respectively. The binding pose prediction problem is also formulated as a classification problem by defining ligand conformations in the range of the rmsd less than the threshold as the bound state and other ligand conformations as the unbound state. For the regression and classification problems, we use Random Forests as a supervised learning algorithm, because it can handle both types of problems easily.[35] Other learning machines, such as support vector machines, decision trees, and artificial neural networks, are also available. However, they generally require time-consuming parameter tuning through trial and

Table 1. 100 Protein−Ligand Complexes Used in This Study

PDB code	$-\log K_d$	resolution (Å)	PDB code	$-\log K_d$	resolution (Å)
1bbz	5.82	1.65	4xia	1.54	2.3
8xia	2.95	1.9	2xim	2.28	2.3
1fkf	9.4	1.7	1fkb	9.7	1.7
1hvr	9.51	1.8	1tet	6.2	2.3
2cgr	7.27	2.2	1abf	5.42	1.9
1apb	5.82	1.76	7abp	5.54	1.67
5abp	6.64	1.8	8abp	4	1.49
9abp	8	1.97	1abe	6.52	1.7
1bap	6.85	1.75	6abp	5.64	1.67
1e96	5.22	2.4	1add	6.74	2.4
2ak3	3.86	1.9	1adb	8.4	2.4
9aat	8.22	2.2	1bzm	6.03	2
1cbx	6.35	2	2ctc	3.89	1.4
3cpa	4	2	1cla	5.28	2.34
3cla	4.94	1.75	4cla	5.47	2
2csc	3.36	1.7	5cna	2	2
1af2	3.1	2.3	1dr1	5.57	2.2
1dhf	7.4	2.3	1drf	7.44	2
1ela	6.35	1.8	7est	7.6	1.8
3fx2	9.3	1.9	2gbp	7.4	1.9
1hsl	7.3	1.89	2qwd	4.85	2
2qwe	7.48	2	2qwf	5.67	1.9
2qwg	8.4	1.8	2qwc	3.55	1.6
2qwb	2.74	2	1mnc	9	2.1
1exw	3.9	2.4	1apw	8	1.8
1apt	9.4	1.8	1bxo	10	0.95
1fmo	8.64	2.2	2pk4	4.32	2.25
1inc	8	1.94	4sga	3.27	1.8
5sga	2.85	1.8	5p21	5.32	1.35
1rbp	6.72	2	1rgk	4.31	1.87
6rnt	2.37	1.8	1rgl	4.43	2
1rnt	5.18	1.9	1zzz	5.13	1.9
1yyy	5.09	2.1	1b5g	8	2.07
1ba8	9	1.8	1bb0	8.36	2.1
2sns	6.7	1.5	1sre	4	1.78
7tln	2.47	2.3	4tln	3.72	2.3
1tmn	7.47	1.9	2tmn	5.89	1.6
3tmn	5.9	1.7	5tln	6.37	2.3
1tlp	7.56	2.3	1etr	7.41	2.2
1ets	8.22	2.3	1d3d	9.09	2.04
1d3p	7.39	2.1	1a46	5.7	2.12
1a5g	10.15	2.06	1bcu	5	2
1tha	5.35	2	4tim	2.16	2.4
6tim	3.21	2.2	7tim	5.4	1.9
1bra	1.82	2.2	1tnj	1.96	1.8
1pph	6.22	1.9	1tnk	1.49	1.8
1tnh	3.37	1.8	1tni	1.7	1.9
1ppc	6.16	1.8	1tng	2.93	1.8
3ptb	4.5	1.7	1tnl	1.88	1.9
1bhf	4.38	1.8	2xis	5.82	1.71

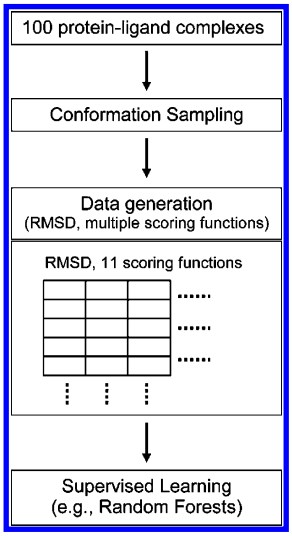

Figure 1. Overview of SCS procedure.

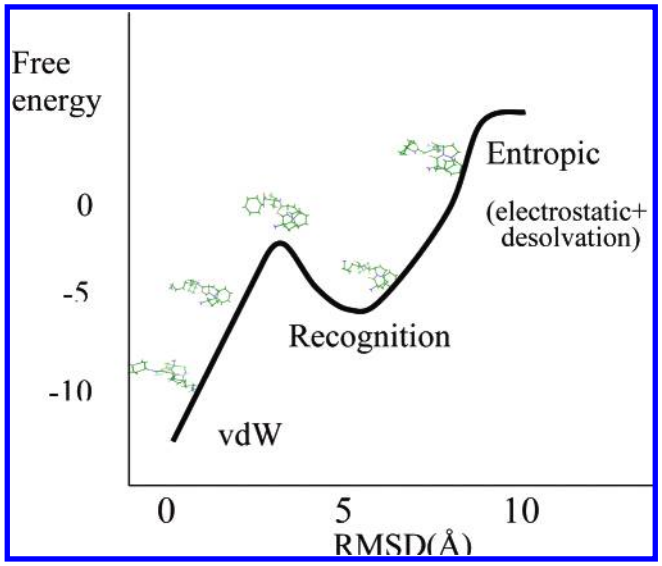

Figure 2. Binding free energy landscape when the rmsd is taken as the reaction coordinate. The concept of SCS is based on the free energy landscape of the protein−ligand binding process, and supervised learning is used for learning the protein−ligand binding process using unbound ligand conformations.

Step 1. Sample with replacement to form N bootstrap samples ($B_1, \ldots, B_N$).

Step 2. Use each sample B_k to construct a tree classifier Tk to predict those samples that are not in B_k (called *out-of-bag* samples). These predictions are called *out-of-bag* estimators.

Step 3. When constructing T_k , at each node splitting we first randomly select *m* variables, then we choose one best split from these *m* variables.

Step 4. Final prediction is the average or majority votes of *out-of-bag* estimators over all bootstrap samples.

Figure 3. Random Forests algorithm.

error. Random Forests combines two machine learning techniques: bagging and random feature subset selection using decision tree and regression tree as base learner.[35] Bagging, which stands for *bootstrap aggregating*, uses resampling to produce pseudoreplicates to improve predictive accuracy. Random Forests can significantly improve predictive accuracy through random feature subset selection. The Random Forests algorithm is illustrated in Figure 3. We constructed three types of SCS models, i.e., a classification model with the rmsd threshold of 1 Å (SCS1), a classification model with the rmsd of 2 Å (SCS2), and a regression model (SCS3), and investigated the model dependence of performance and the properties of each model in detail. The scores obtained from all 11 scoring functions were used as explanatory attributes. In constructing models SCS1 and SCS2, we found that the number of positive and negative examples was extremely imbalanced. To cope with this problem, we used cost-sensitive learning and the undersampling technique.[36] We used the default parameters of Random Forests, except the number of trees, i.e., 2000.

The performance of SCS was evaluated using *out-of-bag* samples to compare CS and the 11 scoring functions fairly. *Out-of-bag* samples are samples left out in a bootstrap sample, and we used them to estimate the predictive accuracy and reliable range of SCS by applying the predictive model constructed using a bootstrap sample. This procedure is equivalent to *n*-fold cross-validation asymptotically and described in other studies.[35,37−39] The R package used in this study, randomForest, is available at http://cran-r-project.org.

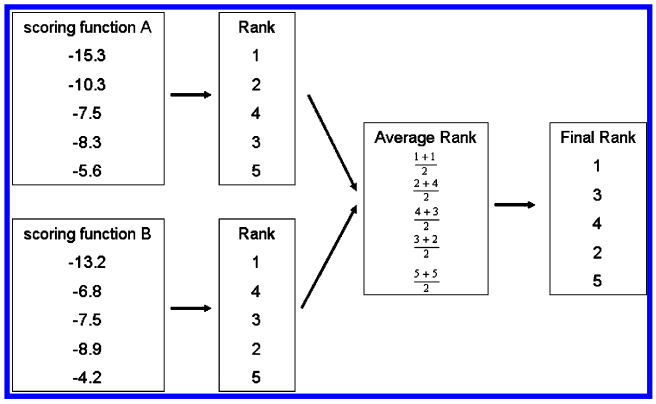

Figure 4. Overview of the rank-by-rank strategy for two scoring functions.

2.3. Consensus Scoring (CS). CS combines multiple scoring functions and improves hit rates. In this study, we used the rank-by-rank strategy, which is a representative and superior CS approach.[28] The rank-by-rank strategy is illustrated in Figure 4. The rank-by-rank-based CS (RCS) is formally defined as

$$\mathrm{RCS} = \frac{\sum_{i=1}^{N} R_i}{N}$$

where R_i is the rank of ligand conformation for the *i*th scoring function, and N is the number of scoring functions. The conformation of minimum score of RCS is evaluated as the predicted binding pose. Since the performance of individual scoring functions has been evaluated already, all possible double and triple combinations of the six relatively successful scoring functions, F-Score, LigScore, PLP, DrugScore, LUDI, and X-Score, were used in the rank-by-rank strategy.[28] Note that CS does not perform better than the best scoring function among those used in it. Therefore, it is reasonable to use only the relatively successful individual scoring function. Since we calculated a tie in the ranking as the average rank, the results of RCS are slightly different from those in a previous study.[28] But, there is little difference in the final results.

2.4. Indicators of Performance for Docking Accuracy. To evaluate the performance of scoring methods in terms of docking accuracy, we examined how closely the top-scoring conformation resembles the ligand conformation in the X-ray structure. For this purpose, we defined a successful prediction as one where the rmsd of the top-scoring conformation is less than or equal to the threshold and the success rate as the percentage of successful predictions over all complexes. Although the general threshold was 2 Å, we set several thresholds, i.e., 1, 2, and 3 Å, to investigate the performance and properties of each scoring method in detail.

As another measure of the performance of the scoring methods, we used the average rmsds ($\mathrm{RMSD_{av}}$) of the top-scoring ligand conformations over all complexes in each scoring method, formally defined as

$$\mathrm{RMSD_{av}} = \frac{\sum_{i=1}^{N} \mathrm{RMSD}_i}{N}$$

Table 2. Success Rates of SCS, CS, and 11 Individual Scoring Functions under Different rmsd Thresholds

scoring method	success rate (%)		
	rmsd < 1.0 Å	rmsd < 2.0 Å	rmsd < 3.0 Å
SCS1	**75**	**91**	94
SCS2	54	90	93
SCS3	63	89	**95**
triple scoring(LigScore+DrugScore+F-Score)	71	85	87
triple scoring(LigScore+DrugScore+PLP)	71	84	86
triple scoring(LigScore+DrugScore+LUDI)	62	82	84
double scoring(DrugScore+LigScore)	68	80	83
double scoring(DrugScore+F-Score)	70	81	84
double scoring(DrugScore+LUDI)	63	80	81
PLP	63	76	80
F-Score	56	74	77
LigScore	64	74	76
DrugScore	63	72	74
LUDI	43	67	67
X-Score	37	66	74
AutoDock	34	62	72
PMF	40	52	57
G-Score	24	42	56
ChemScore	12	35	40
D-Score	8	26	41

where $RMSD_i$ is the rmsd of top-scoring conformation of the *i*th conformation of a ligand, and *N* is the number of protein−ligand complexes. $RMSD_{av}$ is an indictor of how close the top-scoring docked conformation is to the X-ray structure on average and explains other aspects of the performance of the scoring methods.

3. RESULTS AND DISCUSSION

3.1. Success Rate. First, we compare the performance of SCS, CS, and the 11 individual scoring functions. Double and triple combinations of the rank-by-rank strategy were calculated. The best three combinations of scoring functions shown by Wang et al. were used.[28] Success rates for all individual scoring functions, CS, and SCS, are summarized in Table 2, and cumulative success rates are shown in Figure 5 for a visual interpretation.

These results show that SCS1 outperforms other scoring methods at the all rmsd threshold, i.e., 1, 2, and 3 Å. SCS2 and SCS3 outperform all other methods in the range of rmsd < 2 Å and rmsd < 3 Å. However, the performance of SCS2 and SCS3 in the range of rmsd < 1 Å is not always high. In the range of rmsd < 1 Å, SCS2 and SCS3 are inferior or equal to even individual scoring functions, i.e., PLP, F-Score, LigScore, and DrugScore. The poor performance of SCS2 in the range of rmsd < 1 Å results from the difference of the rmsd threshold, whereas that of SCS3 may reflect the effects of scoring functions that do not perform very well, since we used all scores of the11 scoring functions as explanatory attributes. These results suggest that the rmsd threshold is very important and that the overall performance can be improved by learning the protein−ligand binding process from the decoy structures and scoring functions. They also suggest that the rmsd threshold should be 1 Å rather than 2 Å in the classification model.

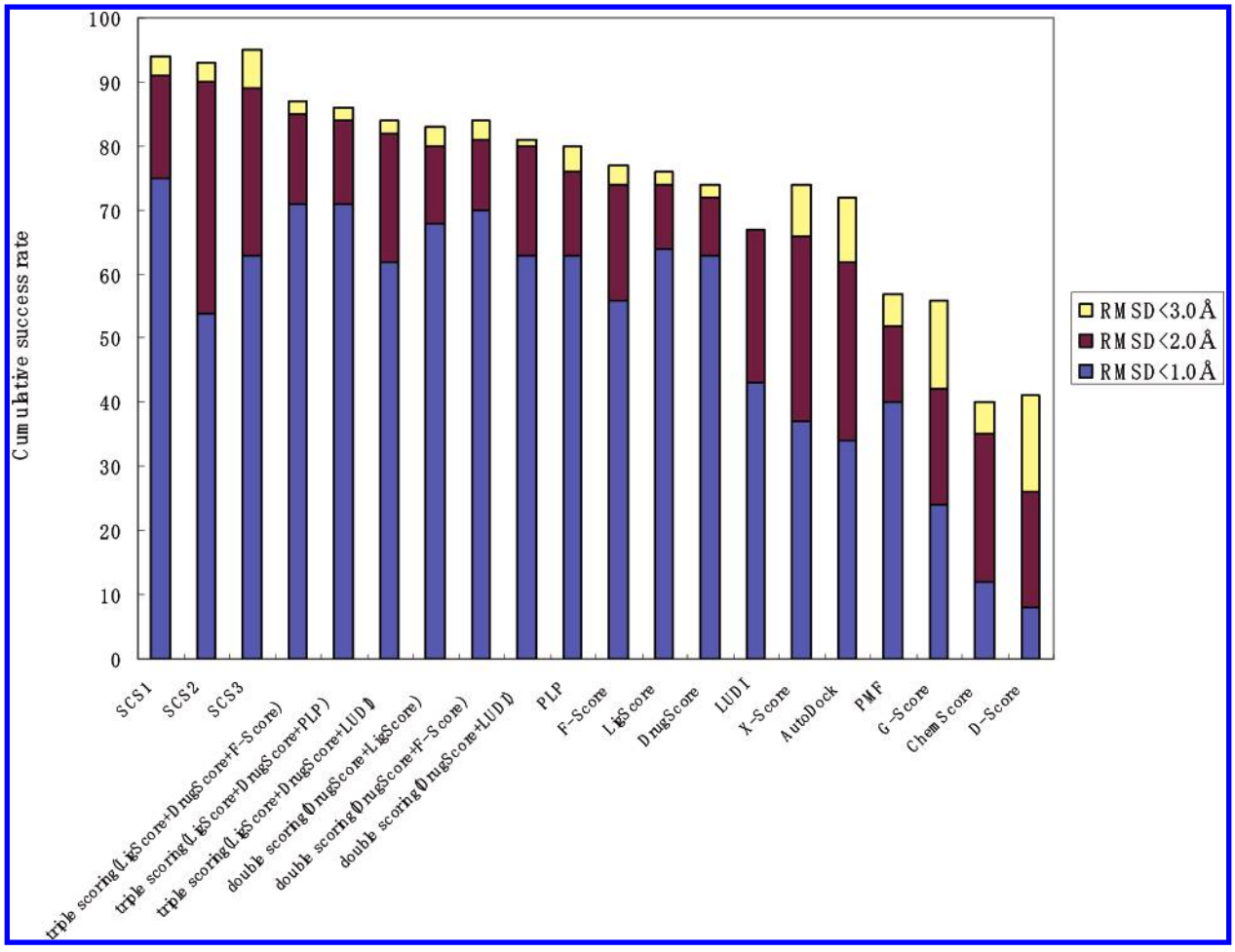

Figure 5. Cumulative success rates of SCS, CS, and 11 individual scoring functions.

Table 3. Average RMSD of the Best-Scored Ligand Conformations over All Complexes ($RMSD_{av}$) for SCS, CS, and 11 Individual Scoring Functions

consensus scoring methods	average rmsd (Å)
SCS1	0.77
SCS2	1.32
SCS3	1.1
triple scoring(LigScore+DrugScore+F-Score)	1.58
Triple scoring(LigScore+DrugScore+PLP)	1.56
Triple scoring(LigScore+DrugScore+LUDI)	2.06
Double scoring(DrugScore+LigScore)	1.94
Double scoring(DrugScore+F-Score)	1.75
Double scoring(DrugScore+LUDI)	2.09
PLP	2.38
F-Score	2.1
LigScore	2.36
DrugScore	2.69
LUDI	3.12
X-Score	2.98
AutoDock	2.94
PMF	4.54
G-Score	4.43
ChemScore	6.07
D-Score	4.83

The maximum difference between SCS and other scoring methods can be seen in the range of rmsd < 3 Å. Success rates in the range of rmsd < 3 Å for all SCS models exceed 90%, while other scoring methods hardly improve success rates at all in this range. We discuss this in the next section.

When we compare the performance among the 11 individual scoring function, the success rates of empirical scoring functions, i.e., PLP, F-Score, and LigScore, and knowledge-based potentials, i.e., DrugScore, in the range of rmsd < 2 Å exceed 70%.

3.2. Average rmsd of the Top-Scoring Ligand Conformations over All Complexes. The success rate is one aspect of performance evaluation in molecular docking and scoring methods. Since the success rate depends on the RMSDs of top-scoring ligand conformations, we examined average rmsds ($RMSD_{av}$) of the top-scoring ligand conformations over all complexes to evaluate the performance of the scoring methods in detail. When $RMSD_{av}$ is small, the top-scoring ligand conformation of the scoring method is close to the X-ray structure on average and vice versa. Therefore, $RMSD_{av}$ can be another indicator of the performance of scoring methods for different aspects of docking accuracy. Table 3 shows the $RMSD_{av}$ for each scoring method. SCS outperforms other scoring methods. In particular, $RMSD_{av}$ of SCS1 is less than 1 Å and shows the best performance as well as the best success rate. $RMSD_{av}$ of SCS3 is smaller than that of SCS2, which is consistent with the success rates in the range of rmsd < 1 Å. From Tables 2 and 3, there is relatively strong correlation between $RMSD_{av}$ and success rate in CS, and $RMSD_{av}$ of CS is always smaller than individual scoring functions. There is relatively weak correlation between $RMSD_{av}$ and success rate in the individual scoring functions. These differences among SCS, CS, and individual scoring functions may originate from multiple utilizations of scoring functions. When we compare the $RMSD_{av}$ of SCS and CS, SCS is always superior to CS. This explains the difference in the success rates in the range of rmsd < 3 Å very well.

As discussed above, the top-scoring ligand conformations in SCS and CS are relatively close to the X-ray ligand structures on average, and the performance of SCS and CS is robust in comparison with individual scoring functions.

Table 4. Success Rates of SCS, CS, and 11 Individual Scoring Functions for Two Binding Affinity Groups: (a) $1 < -\log K_d < 6$ (Group 1) and (b) $7 < -\log K_d < 11$ (Group 2)

	success rate (%)		
scoring method	rmsd < 1.0 Å	rmsd < 2.0 Å	rmsd < 3.0 Å
(a) $1 < -\log K_d < 6$ (Group 1)			
SCS1	**71.4**	**89.3**	**92.9**
SCS2	48.2	85.7	89.3
SCS3	55.4	83.9	**92.9**
triple scoring(LigScore+DrugScore+F-Score)	64.3	82.1	85.7
triple scoring(LigScore+DrugScore+PLP)	64.3	80.4	83.9
triple scoring(LigScore+DrugScore+LUDI)	60.7	80.4	83.9
double scoring(DrugScore+LigScore)	60.7	76.8	80.4
double scoring(DrugScore+F-Score)	62.5	73.2	78.6
double scoring(DrugScore+LUDI)	62.5	76.8	78.6
PLP	57.1	71.4	76.8
F-Score	48.2	66.1	71.4
LigScore	57.1	69.6	75
DrugScore	57.1	62.5	66.1
LUDI	37.5	60.7	60.7
X-Score	35.7	57.1	64.3
AutoDock	26.8	55.4	69.6
PMF	35.7	48.2	51.8
G-Score	23.2	33.9	55.4
ChemScore	5.4	32.1	35.7
D-Score	5.4	17.9	35.7
(b) $7 < -\log K_d < 11$ (Group 2)			
SCS1	**90.3**	**96.8**	96.8
SCS2	67.7	**96.8**	**100**
SCS3	80.6	**96.8**	**100**
triple scoring(LigScore+DrugScore+F-Score)	87.1	**96.8**	96.8
triple scoring(LigScore+DrugScore+PLP)	87.1	**96.8**	96.8
triple scoring(LigScore+DrugScore+LUDI)	64.5	90.3	90.3
double scoring(DrugScore+LigScore)	83.9	93.5	96.8
double scoring(DrugScore+F-Score)	**90.3**	**96.8**	96.8
double scoring(DrugScore+LUDI)	77.4	90.3	90.3
PLP	80.6	90.3	90.3
F-Score	83.9	90.3	93.5
LigScore	77.4	83.9	83.9
DrugScore	77.4	87.1	87.1
LUDI	58.1	80.6	80.6
X-Score	45.2	77.4	87.1
AutoDock	41.9	71	77.4
PMF	48.4	58.1	64.5
G-Score	32.3	61.3	67.7
ChemScore	25.8	45.2	51.6
D-Score	9.7	41.9	54.8

3.3. Dependency of Docking Accuracy to Binding Affinity. We investigated the dependency of docking accuracy to binding affinity. To achieve this, we prepared two groups of protein−ligand complexes according to $-\log K_d$ values from 100 diverse protein−ligand complexes. The $-\log K_d$ value of one group (group 1) is in the range of 1−6, and another group (group 2) is in the range of 7−11. The number of group 1 and group 2 is 56 and 31, respectively. Success rates of each group for individual scoring functions, CS, and SCS are summarized in Table 4(parts (a) and (b), respectively). Table 5(a),(b) shows the $RMSD_{av}$ for each group. From Table 4(a),(b), all scoring methods perform better in group 2 than group 1. SCS drastically outperforms CS and individual scoring functions in group 1. In contrast, there is little difference between SCS and CS in group 2. From Tables 2 and 4, success rates in group 1 are lower than ones in all protein−ligand complexes, and success rates in group 2 are higher than ones in all protein−ligand complexes. From Table 5(a),(b), all scoring methods also perform better in group 2 than group 1. This result is

Table 5. Average rmsd of the Best-Scored Ligand Conformations over All Complexes ($RMSD_{av}$) for Two Binding Affinity Groups: (a) $1 < -\log K_d < 6$ (Group 1) and (b) $7 < -\log K_d < 11$ (Group 2)

consensus scoring methods	average rmsd (Å)
(a) $1 < -\log K_d < 6$ (Group 1)	
SCS1	1.04
SCS2	1.65
SCS3	1.41
triple scoring(LigScore+DrugScore+F-Score)	1.79
triple scoring(LigScore+DrugScore+PLP)	1.92
triple scoring(LigScore+DrugScore+LUDI)	2.19
double scoring(DrugScore+LigScore)	2.26
double scoring(DrugScore+F-Score)	2.32
double scoring(DrugScore+LUDI)	2.58
PLP	2.94
F-Score	2.55
LigScore	2.5
DrugScore	3.4
LUDI	4.01
X-Score	3.52
AutoDock	3.25
PMF	5.45
G-Score	5.24
ChemScore	6.9
D-Score	5.31
(b) $7 < -\log K_d < 11$ (Group 2)	
SCS1	0.32
SCS2	0.91
SCS3	0.68
triple scoring(LigScore+DrugScore+F-Score)	0.58
triple scoring(LigScore+DrugScore+PLP)	0.59
triple scoring(LigScore+DrugScore+LUDI)	1.23
double scoring(DrugScore+LigScore)	0.67
double scoring(DrugScore+F-Score)	0.54
double scoring(DrugScore+LUDI)	0.88
PLP	1.25
F-Score	0.94
LigScore	1.83
DrugScore	1.57
LUDI	1.77
X-Score	2.01
AutoDock	2.2
PMF	3.01
G-Score	2.94
ChemScore	4.59
D-Score	3.56

consistent with success rates. However, the $RMSD_{av}$ of SCS1 is smallest in all scoring methods, and the difference of $RMSD_{av}$ is more clear than success rates. From Tables 3 and 5, the $RMSD_{av}$ in group 1 are larger than ones in all protein−ligand complexes, and the $RMSD_{av}$ in group 2 are smaller than ones in all protein−ligand complexes.

To summarize, although the performance of all scoring methods depends on binding affinity, SCS also performs quite well in the range of weak binding affinities. These results demonstrate that SCS considerably outperforms other scoring methods under hard situations for docking and is a robust scoring method.

3.4. Criterion Enriching Virtual Screening. In practical situations, i.e., the virtual screening of a large-scale compound database, the value, or rank of a compound is employed to find novel ligands, and we need a criterion to judge whether a compound is active or inactive. Since CS and individual scoring functions depend on the number of conformation ensembles for each target protein and it is difficult to determine a criterion clearly and uniquely, we here focus on SCS only. It is appropriate to determine the criterion for ligand conformations within a rmsd by estimating reliable range of the scores of SCS using *out-of-bag* samples. Scatter plots of rmsd and the score of each SCS are shown in Figure 6(a)−(c). From Figure 6(a), a ligand conformation of rmsd < 2 Å is expected if the score of SCS1 is less than 0.5, and rmsd < 1 Å is expected if the score of SCS1 is less than 0.2. However, we cannot estimate the rmsd of ligand conformation for SCS2 from Figure 6(b) because there is no correlation or trend between rmsd and SCS2. Therefore, SCS2 is not appropriate for virtual screening. From Figure 6(c), a ligand conformation of rmsd < 2 Å is expected if the score of SCS3 is less than 1.5. Thus, we can obtain a criterion to judge whether a compound is active or

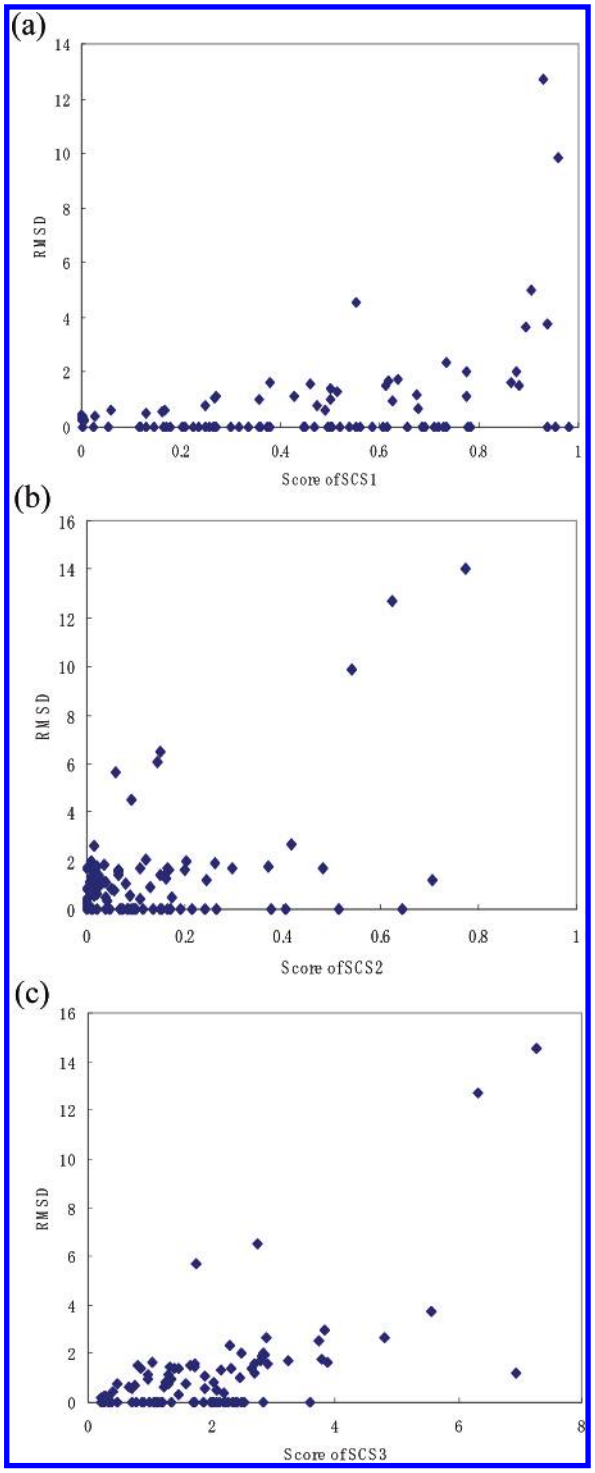

Figure 6. Scatter plots between rmsd (Å) and score of SCS: (a) SCS1, (b) SCS2, and (c) SCS3.

inactive from the score distribution of SCS. Although the number of docked conformations is 101 in this study, the higher the number is, the more precise the SCS may become. To summarize, SCS1 is the most appropriate method for virtual screening, it is expected that novel ligands in the range of rmsd < 1 Å can be found if SCS1 is less than 0.2, and SCS may become more precise if a larger number of docked conformations, i.e., training data, is available. Even if only one protein−ligand complex is available, SCS with supervised learning will improve docking accuracy when many conformation ensembles are generated by conformation sampling. This is the major advantage of SCS, compared with other scoring methods.

Since SCS does not aim at predicting binding affinity, SCS cannot estimate it. Therefore, it is necessary that the binding affinity should be predicted by calculating it by other methods. In practice, for predicting binding affinity, SCS would presumably be useful for screening a large-scale compound database before the binding affinity is calculated precisely by MD simulation and so on. This concept is similar to hierarchical virtual screening.[40,41]

3.5. Free Energy Landscape of Protein−Ligand Complexes. Here, we discuss the relation between the free energy landscape and SCS. Since SCS takes into account the protein−ligand binding process by unbound ligand conformations, SCS is expected to be associated with the free energy landscape. In this study, we hypothesize the 101 docked conformations of each ligand as spots on the protein−ligand complexation energy landscape. With regard to the energy landscape of protein−ligand interaction, Camacho and Vajda discussed the funnel-shaped feature of binding free energy landscape when the rmsd is taken as the reaction coordinate.[42] The funnel-shaped energy landscape is well-known for protein folding.[43]

In this section, we examine the correlation between rmsd and the score of SCS3, because SCS3 is a regression model for rmsd and appropriate for discussing this issue. Of course, correlation between rmsd and score alone may not sufficiently define a funnel-shaped energy landscape. However, such correlation is an important feature of a funnel-shaped energy landscape. We evaluate the correlation between the rmsd and score using the Spearman correlation coefficient (R_s), which calculates the correlation between two sets of ranking. R_s is formally defined as

$$R_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n-1)}$$

where n is the number of sets, and d_i is the ranking difference of the ith complex under two criteria, i.e., the rmsd and score of SCS3. By definition, R_s ranges from −1 to 1. The larger R_s becomes, the more strongly two sets correlate.

Three examples, i.e., PDB code 1bbz, 2qwg, and 2sns, are shown in Figure 7. By the definition of previous study,[28] 1bbz includes protein−ligand interaction dominated by a hydrophilic factor, 2qwg includes protein−ligand interaction dominated by a hydrophobic factor, and 2sns includes protein−ligand interaction having hydrophilic and hydrophobic factors. There are relatively strong correlations between rmsd and score, i.e., the R_s values of 1bbz, 2qwe, and 2sns are 0.684, 0.883, and 0.660, respectively. Assuming that the funnel bottom corresponds to the X-ray complex structure, we would expect that a lower score is associated with a smaller rmsd and vice versa. These results may suggest that SCS3 reflects the concept of a funnel-shaped free energy landscape for receptor−ligand interaction.

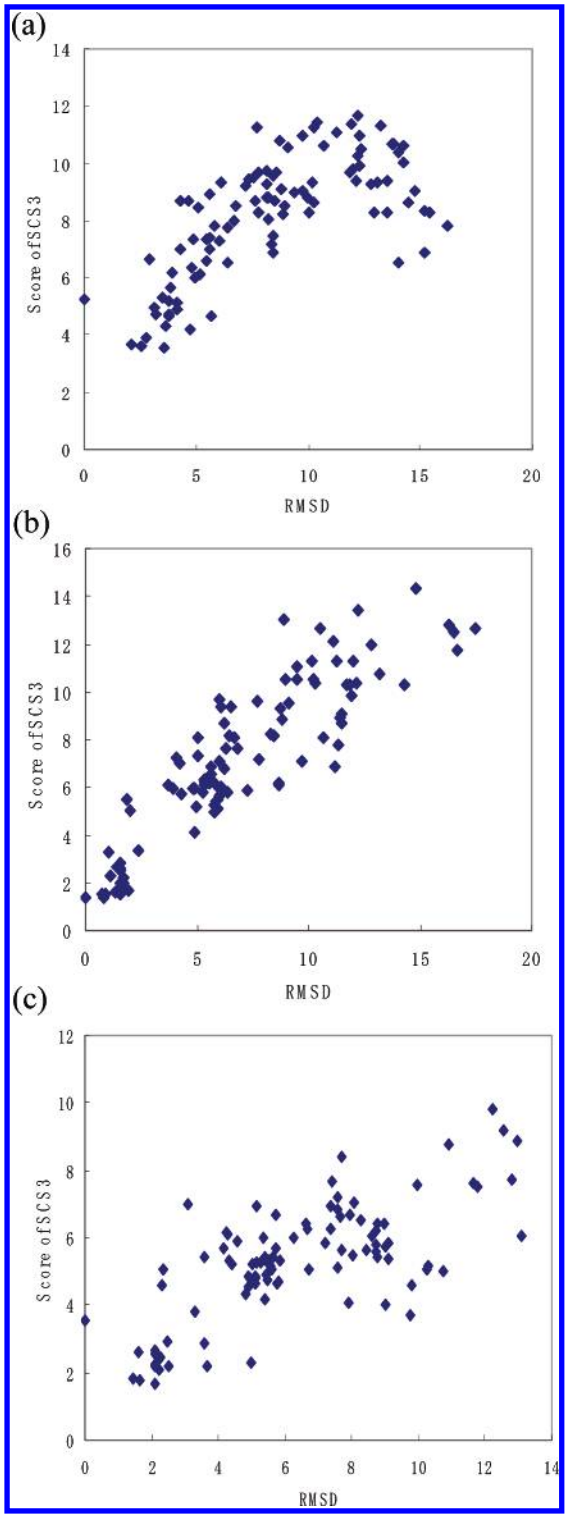

Figure 7. Correlations between rmsd (Å) and score of SCS3: (a) 1bbz (R_s = 0.684), (b) 2qwg (R_s = 0.883), and (c) 2sns (R_s = 0.660).

3.6. Discussion of Outliers. We can see that there are several outliers in Figure 7 (a)−(c). As examples of outliers, we demonstrate the correlation between the rmsd and SCS3

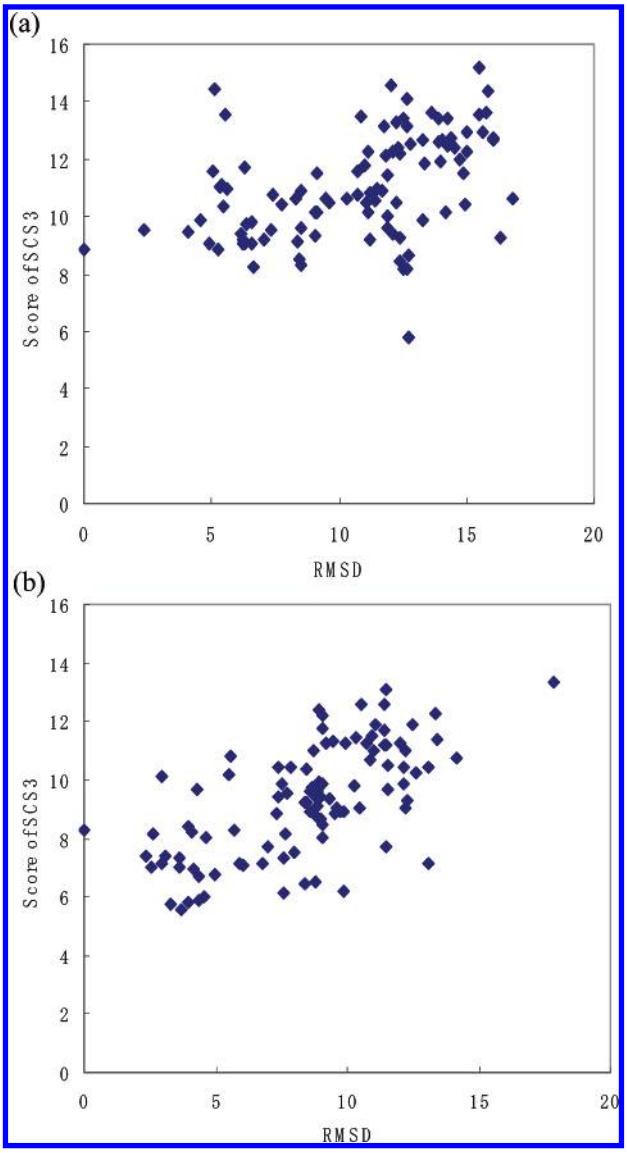

Figure 8. Scatter plots of rmsd (Å) and score of SCS3: (a) 3cla (R_s = 0.468) and (b) 1tha (R_s = 0.463).

of PDB code 3cla and 1tha in Figure 8. We can see that the conformation sampling is insufficient in the range of rmsd < 2 Å. However, there is a correlation between rmsd and SCS3, i.e., R_s of 3cla and 1tha are 0.468 and 0.463, respectively. The 3cla is chloramphenicol acetyltransfrase with chloramphenicol. rmsd of the top-scoring docked conformation is 12.72 Å in SCS3. In this case, layers of water molecules exist on the protein−ligand binding site, and conformation sampling is performed without those water molecules. In this case, there is limitation in the current docking approach. The 1tha is transthyretin/3,3′-diiodo-L-thyronine. In this case, the ligand exists along a shallow groove but not in a well-defined pocket. The rmsd of top-scoring docked conformation is 3.71 Å in SCS3, while that of the top-scoring docked conformation in CS is 9.86 Å. This improvement may originate from the incorporation of unbound ligand conformations.

4. CONCLUSION

We have proposed supervised the consensus scoring method (SCS), which takes into account the protein−ligand binding process using unbound ligand structures with supervised learning. To confirm the effectiveness of SCS, we tested 100 diverse protein−ligand complexes. The docking accuracy of scoring methods was evaluated on the basis of the reproducibility of the pose in the X-ray structure. To compare the performance of conventional CS methods and 11 individual scoring functions fairly, we separated the docking procedure and scoring procedure. In this way, we could evaluate scoring methods on the same ground without dependency on any particular docking program.

We showed that SCS outperforms both CS and individual scoring function in predicting the binding pose, i.e., the success rate and average rmsd of the top-scoring docked conformation over all complexes ($RMSD_{av}$). In particular, SCS1 achieved 75% in success rate and 0.769 in $RMSD_{av}$, while other scoring methods achieved 71% and 1.56 at most. However, in the classification model of SCS, the performance largely depends on the rmsd threshold. Therefore, it is very important to determine the threshold appropriately, and these results suggest that the threshold should be 1 Å.

To estimate the effectiveness of learning the protein−ligand binding process, we examined the correlation between rmsd and SCS3. There are relatively strong correlations between the rmsd and score of SCS3 in all three interaction types, i.e., protein−ligand interaction dominated by the hydrophilic factor, that dominated by the hydrophobic factor, and that having hydrophilic and hydrophobic factors.

Moreover, we also described a procedure for determining appropriate criterion for judging whether a compound is active or inactive in virtual screening. It is expected that this procedure will enable us to find novel ligands for target proteins with high probability. The results of the performance evaluation suggest that even if only one protein−ligand complex is available, SCS may improve docking accuracy using supervised learning with many conformation ensembles generated by conformation sampling. This is the major advantage of SCS, compared with other scoring methods.

Since SCS does not aim at predicting binding affinity, SCS cannot estimate it. Therefore, to predict binding affinity, it should be calculated by other methods. In practice, for predicting binding affinity, SCS is presumably useful for screening a large-scale compound database before binding affinity is calculated precisely by MD simulation and so on.

Nevertheless, there are several problems in scoring functions, such as how to take into account the water-mediated protein−ligand interactions or particular binding sites, such as a ligand existing along a shallow groove but not in a well-defined pocket. An ideal scoring function for docking should be good at both conformation sampling and scoring. Although SCS is still the poor representation of thermodynamics and incomplete exploration of the molecular degrees that limit docking predictions, our study sheds light on how to design the desirable scoring functions.

ACKNOWLEDGMENT

The authors thank our colleagues at NEC Corp. for fruitful discussions.

REFERENCES AND NOTES

(1) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Sheridan, R. P.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.

(2) Abagyan, R. A.; Totrov, M. M.; Kuznetsov, D. A. ICM: a new method for structure modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488−506.
(3) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−89.
(4) Welch, W.; Ruppert, J.; Jain, A. N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.* **1996**, *3*, 449−462.
(5) Jones, G.; Wilett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(6) McMartin, C.; Bohacek, R. QXP: Powerful, rapid computer algorithms for structure-based drug design. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 333−344.
(7) Morris, G. M.; Goodsell, D. S.; Halliday, R.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(8) Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins* **1998**, *33*, 367−382.
(9) Hou, T.; J., W.; Chen, L.; Xu, X. Automated docking of peptides and proteins by using a genetic algorithm combined with a tabu search. *Protein Eng.* **1999**, *12*, 639−647.
(10) Liu, M.; Wang, S. MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 435−451.
(11) Perola, E.; Xu, K.; Kollmeyer, T. M.; Kaufmann, S. H.; Prendergast, F. G.; Pang, Y. P. Successful virtual screening of a chemical database for farnesyltransferase inhibitor leads. *J. Med. Chem.* **2000**, *43*, 401−408.
(12) Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided. Mol. Des.* **2001**, *15*, 411−428.
(13) Zavodszky, M. I.; Sanschagrin, P. C.; Korde, R. S.; Kuhn, L. A. Distilling the essential features of a protein surface for improving protein. ligand docking, scoring, and virtual screening. *J. Comput.-Aided. Mol. Des.* **2002**, *16*, 883−902.
(14) Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499−511.
(15) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.
(16) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, *57*, 225−242.
(17) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **2004**, *57*, 225−242.
(18) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of docking performance: Comparative data on docking algorithms. *J. Med. Chem.* **2004**, *56*, 558−565.
(19) Kroemer, R. T.; Vulpetti, A.; McDonald, J. J.; Rohrer, D. C.; Trosset, J. Y.; Giordanetto, F.; Cotesta, S.; McMartin, C.; Kihlen, M.; Stouten, P. F. W. Assessment of docking poses: Interactions-based accuracy classification (IBAC) versus crystal structure deviations. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 871−881.
(20) Kontoyianni, M.; Sokol, G. S.; McClellan, L. M. Evaluation of library ranking efficacy in virtual screening. *J. Comput. Chem.* **2005**, *26*, 11−22.
(21) Majeux, N.; Scarsi, M.; Apostolakis, J.; Caisch, A. Exhaustive docking of molecular fragments on protein binding sites with electrostatic salvation. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 88−105.
(22) Zou, X.; Sun, Y.; Kuntz, I. D. *J. Am. Chem. Soc.* **1999**, *121*, 8033−8043.
(23) DeWitte, R.; Shakhnovich, E. SMoG: de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.* **1996**, *118*, 11733−11744.
(24) Mitchell, J. B. O.; Laskowski, R. A.; Alexander, A.; Thornton, J. M. BLEEP-Poteintial of mean force describing protein-ligand interactions: I. Generating potential. *J. Comput. Chem.* **1999**, *20*, 1165−1176.
(25) Muegge, I.; Martin, Y. C.; A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.
(26) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(27) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; Lalonde, J.; Lambert, M. H.; Lindvalee, M.; Nevins, N.; Semus, S. F.; Senquer, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912−5931.
(28) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.
(29) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.
(30) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases: 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.
(31) Stahl, M.; Rarey, M.; Detailed, analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.
(32) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graphics Modell.* **2002**, *20*, 281−295.
(33) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, T. D.; Watson, P. Virtual screening using protein-ligand docking: Avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793−806.
(34) Yang, J. M.; Chen, Y. F.; Shen, T. W.; Kristal, B. S; Hsu, D. F. Consensus scoring criteria for improving enrichment in virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 1134−1146.
(35) Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5−32.
(36) Chen, C.; Liaw, L.; Breiman, L. *Using random forest to learn imbalanced data*; Technical Report 666; Statistics Department, University of California at Berkeley: 2004.
(37) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, C.; Sheridan, R.; Feuston, B. Random Forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Model.* **2003**, *43*, 1947−1958.
(38) Breiman, L. *Out-of-estimator*; Technical Report; Department of Statistics, UC Berkeley: 1996.
(39) Hastie, T.; Tibshirani, R.; Friedman, J. *The element of statistical learning*; Springer: 2001.
(40) Wely, F.; Vaidehi, N.; Zamanakos, G.; Goddard, W.; HierVLS hierarchical docking protocol for virtual ligand screening of large-molecule databases. *J. Med. Chem.* **2004**, *47*, 56−71.
(41) Gruneberg, S.; Stubbs, M.; Klebe, G.; Successful, virtual screening for novel inhibitors of human carbonic anhydrase: strategy and experimental confirmation. *J. Med. Chem.* **2002**, *45*, 3588−3602.
(42) Camacho, C. J.; Vajda, S. Protein docking along smooth association pathways. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10636−10641.
(43) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, Pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **1995**, *21*, 167−195.

CI6004993