Structure-Based Virtual Screening with Supervised Consensus Scoring: Evaluation of Pose Prediction and Enrichment Factors

Reiji Teramoto*,† and Hiroaki Fukunishi‡

Bio-IT Center and Nano Electronics Research Laboratories, NEC Corporation, 34, Miyukigaoka, Tsukuba, Ibaraki 305-8501, Japan

Received December 18, 2007

Since the evaluation of ligand conformations is a crucial aspect of structure-based virtual screening, scoring functions play significant roles in it. However, it is known that a scoring function does not always work well for all target proteins. When one cannot know which scoring function works best against a target protein a priori, there is no standard scoring method to know it even if 3D structure of a target protein–ligand complex is available. Therefore, development of the method to achieve high enrichments from given scoring functions and 3D structure of protein–ligand complex is a crucial and challenging task. To address this problem, we applied SCS (supervised consensus scoring), which employs a rough linear correlation between the binding free energy and the root-mean-square deviation (rmsd) of a native ligand conformations and incorporates protein–ligand binding process with docked ligand conformations using supervised learning, to virtual screening. We evaluated both the docking poses and enrichments of SCS and five scoring functions (F-Score, G-Score, D-Score, ChemScore, and PMF) for three different target proteins: thymidine kinase (TK), thrombin (thrombin), and peroxisome proliferator-activated receptor gamma (PPARγ). Our enrichment studies show that SCS is competitive or superior to a best single scoring function at the top ranks of screened database. We found that the enrichments of SCS could be limited by a best scoring function, because SCS is obtained on the basis of the five individual scoring functions. Therefore, it is concluded that SCS works very successfully from our results. Moreover, from docking pose analysis, we revealed the connection between enrichment and average centroid distance of top-scored docking poses. Since SCS requires only one 3D structure of protein–ligand complex, SCS will be useful for identifying new ligands.

1. INTRODUCTION

Recently, structure-based virtual screening is widely used to discover novel ligands efficiently. Various docking programs have been developed and evaluated in detail.[1–4] 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 docked conformations are generated. In the scoring step, a scoring function is used to evaluate the protein–ligand affinity. Since the final predicted conformations are selected according to the scores, the scoring functions are crucially important.

There are three kinds of scoring functions: force-field-based, empirical, and knowledge-based potentials. Force-field-based scoring functions are based on the energy functions of classical molecular mechanics. They approximate the binding free energy of protein–ligand complexes by summing the van der Waals and electrostatic interactions. Solvation is usually taken into account by using a distance-dependent dielectric function, although solvent models based on continuum electrostatics have also been developed.[5,6] Empirical scoring functions estimate the binding free energy by summing interaction terms derived from 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 ligands 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. These potentials are derived from the protein–ligand complexes with known structures, where the probability distributions of interatomic distances are converted into distance-dependent interaction free energies of protein–ligand atom pairs. However, the 3D structures of protein–ligand complexes do not provide a thermodynamic ensemble at equilibrium, and therefore, a knowledge-based potential should be considered as a 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. Several approaches to derive these potentials have been proposed.[7–10]

While many reports assessing the docking poses and enrichments of docking programs have been published, many of these reports concluded that docking algorithms reproduce binding modes very well and that a scoring function does not always work well for all target proteins.[11–16] Therefore, a scoring function should provide a target specific scoring function depending on a target protein and work well independently against target proteins. Moreover, when one

* Corresponding author. Phone: +81 298 850 1410. Fax: +81 298 856 6136. E-mail: r-teramoto@bq.jp.nec.com.
† Bio-IT Center.
‡ Nano Electronics Research Laboratories.

10.1021/ci700464x CCC: $40.75
Published on Web 03/05/2008

Table 1. Protein–Ligand Complexes, Ligands, and Decoys Used in This Study

protein	PDB code	resolution (Å)	no. of ligands	no. of rotatable bonds of native ligand	no. of decoys
TK	1kim	2.1	21	2	887
thrombin	1ba8	1.8	70	14	2441
PPARγ	1fm9	2.1	84	12	3006

can not know which scoring function works best against a target protein a priori, there is no standard scoring method to predict new ligands with high probability.

To address this problem, we apply supervised consensus scoring (SCS), which employs a rough linear correlation between the binding free energy and the root-mean-square deviation (rmsd), incorporates protein–ligand binding process with docked ligand conformations using supervised learning, and provides a target specific scoring model, to virtual screening.[17] A rough linear correlation between the binding free energy and the rmsd was discussed by Camacho, et al. originally in terms of funnel-shaped free energy landscape.[18] They reported that docking based on the above assumption performs very successfully. Although SCS has been successfully applied in docking accuracy in the previous study, it lacked enrichment studies and their discussions that are crucial and quite different task in docking study. It is well-known that docking accuracy and enrichments are quite different aspects in docking study.[15,16] Therefore, only docking accuracy is not enough to evaluate the performance of a scoring method and enrichment studies are crucial and significant. Consequently, in the present study, we evaluate both the docking poses and enrichments of SCS and five scoring functions (F-Score, G-Score, D-Score, ChemScore, and PMF) for three different target proteins: thymidine kinase (TK), thrombin (thrombin), and peroxisome proliferator-activated receptor gamma (PPARγ). The comparison of enrichments of each scoring method is performed on the basis of enrichments with benchmarking sets for molecular docking. Since docking poses analysis is also important for evaluating docking performance, we discuss based on centroid distance between a native ligand conformation and top-scored docking poses.

2. METHODS

2.1. Preparation of Data Sets. Since a directory of useful decoys (DUD) provides a stringent test by which to evaluate the performance of structure-based virtual screening, DUD is appropriate for fair and rigorous evaluations of ligand enrichment to avoid the bias of decoys. Moreover, since DUD is freely available, it is an easy reference by which to compare the performances of scoring methods. We collected 3D structures of three target proteins, i.e., TK, thrombin, and PPARγ, their ligands, and decoys from DUD to do a fair evaluation of ligand enrichment.[19] These target proteins were chosen as the representative proteins from different protein families, TK kinase, thrombin serine protease, and PPARγ nuclear hormone receptor. DUD stores many decoys that physically resemble ligands, so that enrichment is not simply a separation of gross features, but is chemically distinct from them, so that they are unlikely to be binders. The test data sets are summarized in Table 1. Detailed descriptions of DUD and all of the data sets are available online at http://blaster.docking.org/dud/.

Table 2. Number of Docked Conformations of Native Ligands

protein	ligand	no. of ligand conformations
TK	deoxythymidine	236
thrombin	tripeptidylaldehydes	393
PPARγ	GI262570	329

2.2. Docking Procedure and Scoring Functions. FlexSIS implemented in Sybyl7.1J is employed to generate an ensemble of docked conformations for each ligand.[20] Although FlexSIS incorporates ligand flexibility, receptor is treated as rigid body. Since we wanted to dock diverse ligands with large variations in size and possible interactions, all water molecules in the active sites were removed to avoid biasing the docking to one particular binding mode according to the procedure of the previous studies.[21,22] Since it is difficult to know whether conformation sampling is enough to lead to meaningful scoring, we generated 999 docked conformations for native ligands at a maximum to generate them as many as possibleThe number of docked conformations of native ligands is summarized in Table 2. We generated 100 docked conformations for other compounds, i.e., known ligands and decoys at a maximum. Note that we can only set the maximum number of docked conformations and cannot set the number of docked conformations generated directly. All docked conformations generated by FlexSIS were scored using F-Score,[23] D-Score,[24] PMF,[25–27] G-Score,[24] and ChemScore[28] implemented in CScore.[29] D-Score and G-Score are force-field based scoring functions, F-Score and ChemScore are empirical scoring functions, and PMF is a knowledge-based scoring function. These scoring functions are very popular and assessed in docking and structure-based virtual screening.[16,22]

2.3. Supervised Consensus Scoring (SCS) for Virtual Screening. The overall SCS procedure for virtual screening is illustrated in Figure 1, and the binding free energy landscape when rmsd is used as a reaction coordinate is illustrated in Figure 2. As shown in Figure 2, we assumed that protein–ligand binding has a funnel-shaped landscape, as discussed by Camacho and Vajda.[18] SCS employs this rough linear correlation relationship between binding free energy and the rmsd of a native ligand.[17] Note that we do not incorporate ligand symmetry, because SCS employs rough linear relationship between binding free energy and the root mean square deviation (rmsd) of docked native ligand conformations. The screening procedure is as follows (Figure 1):

(Step1) Prepare X-ray structure of target protein–ligand complex and compounds for screening. In this study, we prepared known active compounds and decoys as compounds for screening to evaluate the enrichments as described in section 2.1.

(Step 2) Perfrom conformation sampling for a native ligand and compounds for screening.

(Step 3) Rescore against decoy conformations. In this study, we used five scoring functions: F-Score, G-Score, D-Score, ChemScore, and PMF.

(Step 4) Perform supervised learning with scoring functions and rmsds of a native ligand conformations as training data. We employ the explanatory attributes as the scores of

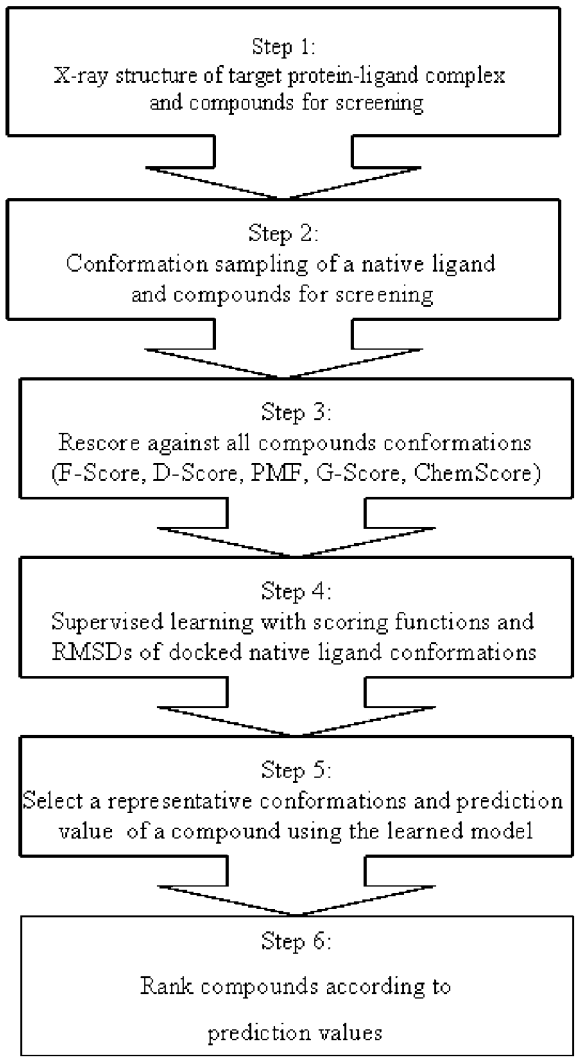

Figure 1. Overview of SCS procedure.

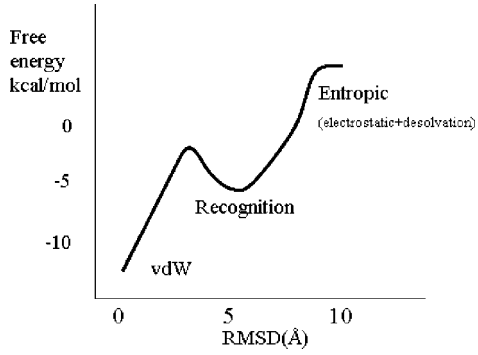

Figure 2. Binding free energy landscape when the rmsd is taken as the reaction coordinate. SCS employs a rough linear correlation between binding free energy, and the rmsd of a native ligand in order to predict the binding energy. SCS takes into account the protein–ligand binding process using docked ligand conformations with supervised learning.

five scoring functions and an objective variable as the rmsd between docked native ligand conformations and X-ray structure for a ligand.

(Step 5) Apply the learned model in step 4 to compounds for screening. First, select a representative conformation of a top-ranked compound conformation. Subsequently, a prediction value of a representative conformation is defined as a representative prediction value of a compound.

(Step 6) Rank compounds for screening according to their representative prediction values. Compounds are ranked in ascending order according to score in regression and are ranked in ascending order according to score of the active class in classification.

Thus, in SCS, the binding energy prediction problem is formulated as supervised learning in which explanatory attributes and an objective variable are the scores of five scoring functions (F-Score, G-Score, D-Score, ChemScore, and PMF) and the rmsd between docked native ligand conformations and X-ray structure for a ligand, respectively. Note that the rmsd of the docked native ligand loses its original physical meaning and give pseudo binding free energy when the trained SCS is applied to virtual screening. The binding energy prediction problem was also formulated as a classification problem by defining ligand conformations in the range of the rmsd less than the threshold as bound states and other ligand conformations as unbound states. For the regression and classification problems, we used random forests as a supervised learning algorithm, because it can handle both problems easily, achieve high generalization performance, and does not overfit to training data.[30,31] Although other learning machines, such as support vector machines, decision trees, and artificial neural networks, are also available, they generally require time-consuming parameter tuning through trial and error. Random forests combines two machine learning techniques: bagging and random feature subset selection using a decision tree and a regression tree as the base learner.[30,31] 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 and the illustration of the random forests algorithm are illustrated in Figure 3. Note that SCS does not provide an explicit expression as a scoring function unlike the existing scoring functions, because SCS construct a scoring model implicitly.

We set the threshold based on the rmsd distributions for each protein–ligand complex for the classification models. Table 3 shows the rmsd thresholds used in constructing classification model for each protein–ligand complex. We set two thresholds in ascending order per 1 Å among the rmsds of generated conformations of a native ligand. We investigated the threshold dependence of performance and the properties of each model in detail.

In the classification, one found that the numbers of positive and negative examples were extremely imbalanced when one defines positive examples as docked conformations with rmsd less than a threshold and negative examples as docked conformations with rmsd more than a threshold. To address this problem, we used cost-sensitive learning.[32] We weighted the active class 100 times against the negative class. We used the default parameters of random forests, except the number of trees, i.e., 5000. We evaluated the enrichments according to the ascending order of scores in regression and the descending order of scores of the active class in classification.

To evaluate whether SCS can predict RMSDs of a native ligand, we employed *out-of-bag*. 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

(a) Random Forests algorithm

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 T_k 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 one chooses 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.

(b) Illustration of Random Forests algorithm.

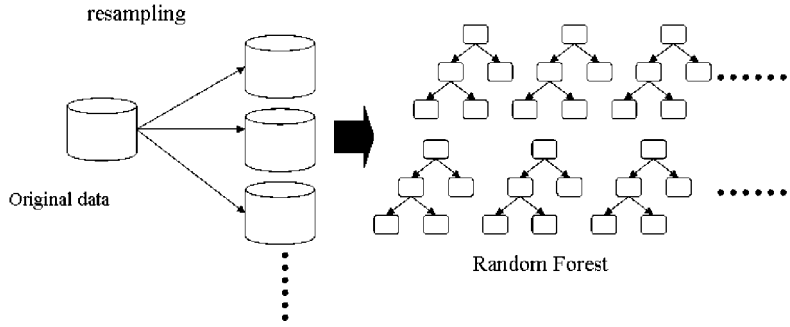

Figure 3. Random forests: (a) random forests algorithm, (b) illustration of random forests algorithm.

Table 3. Root Mean Square Deviation Thresholds for Each Protein–Ligand Complex in Classification Models

protein	ligand	threshold (Å)
TK	deoxythymidine	1.2
thrombin	tripeptidylaldehydes	7.8
PPARγ	GI262570	7.8

bootstrap sample. This procedure is equivalent to *n*-fold cross-validation asymptotically. The R package used in this study, randomForest, is available at http://cran-r-project.org.

3. RESULTS AND DISCUSSION

3.1. Performance Evaluation of Virtual Screening. We compared the performances of SCS and five scoring functions (F-Score, G-Score, D-Score, ChemScore, and PMF) as virtual screening tools when they are applied to the same target proteins. At a coarse level, virtual screening is a test of the ability of scoring methods to differentiate between active and inactive compounds. As the indicator of performance, we used the enrichment factor (EF). EF is defined as

$$\mathrm{EF} = \frac{\mathrm{Hits}_{\mathrm{sampled}}^{x\%}}{N_{\mathrm{sampled}}^{x\%}} \cdot \frac{N_{\mathrm{total}}}{\mathrm{Hits}_{\mathrm{total}}}$$

where $\mathrm{Hits}_{\mathrm{sampled}}^{x\%}$ is the number of hits found at *x*% of the database screened, $N_{\mathrm{sampled}}^{x\%}$ is the number of compounds screened at *x*% of the database, $\mathrm{Hits}_{\mathrm{total}}$ is the number of active compounds in the entire database, and N_{total} is the number of compounds in the entire database. EF is the relative enrichment of active compounds in the set of compounds predicted to be active in relation to the fraction of active compounds in the entire database. From the definition of EF, the EF of random screening is 1. We calculated EF_1 (enrichment factor at 1% of the ranked database), EF_2 (enrichment factor at 2% of the ranked database), and EF_5 (enrichment factor at 5% of the ranked database) of SCS and five scoring functions to evaluate enrichments at the top-rank screened database. These enrichment factors are summarized in Table 4.

From Table 4, one can see that for TK, $\mathrm{SCS}_{\mathrm{reg}}$, $\mathrm{SCS}_{\mathrm{class}}$(1 Å), $\mathrm{SCS}_{\mathrm{class}}$(2 Å), and G-Score work best at EF_1, after which G-Score performs best. For thrombin, $\mathrm{SCS}_{\mathrm{reg}}$ is the best at EF_1, after which F-Score and $\mathrm{SCS}_{\mathrm{class}}$(7 Å) are the best at EF_2, and F-Score performs best at EF_5. For PPARγ, $\mathrm{SCS}_{\mathrm{class}}$(8 Å) is the best at EF_1. D-Score is the best from EF_2 and EF_5. The cutoff thresholds of rmsds for the thrombin and PPARγ cases are 7–8 Å and different from their crystal structures. However, when rmsds of the docked native ligand conformations are large, SCS is trained as binding free energy is high. Therefore, SCS is able to provide a precise scoring model by this bias. It may lead to higher enrichments.

Thus, although the performance of SCS depends on the target proteins, learning models, and thresholds of the classification model, SCS is competitive or superior to a best single scoring function for a target protein at the top ranks of the screened compounds, which is of interest for practical drug screening. On the other hand, no single scoring function performed well for all target proteins likewise the previous study.[16] Our results demonstrate that SCS is competitive or superior to a best single scoring function even if one can not know which scoring function works best against a target protein a priori.

Although apparently SCS does not work very well from Table 4 as compared to a best single scoring function, note that SCS is always competitive or superior to a best single function at the top ranks of screened database. For example, if one chooses G-Score or D-Score in thrombin because of a lack of available active compounds, it results in a disaster. However, if one employs SCS, enrichments are competitive to a best single scoring function, i.e., F-Score in thrombin. Note that the enrichments of SCS could be limited by a best scoring function, because SCS is obtained on the basis of the five individual scoring functions. Therefore, it is concluded that SCS works very successfully from our results.

Table 4. Enrichment Factors of SCS and Five Scoring Functions (F-Score, G-Score, D-Score, ChemScore, and PMF)

TK			
scoring method	EF_1	EF_2	EF_5
F-Score	0	0	0
G-Score	**4.8**	4.8	3.84
D-Score	0	0	1.92
ChemScore	0	0	0
PMF	0	2.4	1.92
SCSreg	0	2.4	0.96
SCS_{class}(1 Å)	**4.8**	2.4	1.92
SCS_{class}(2 Å)	**4.8**	4.8	1.92
Thrombin			
scoring method	EF_1	EF_2	EF_5
F-Score	4.3	**5.74**	8.61
G-Score	0	0.72	0.29
D-Score	0	0.72	2.01
ChemScore	2.87	2.87	2.3
PMF	0	0.72	0.29
SCSreg	**7.17**	4.3	4.59
SCS_{class}(7 Å)	5.74	**5.74**	4.02
SCS_{class}(8 Å)	5.74	5.02	4.02
PPARγ			
scoring method	EF_1	EF_2	EF_5
F-Score	6.13	4.82	2.39
G-Score	2.45	1.81	3.58
D-Score	18.39	**14.47**	7.88
ChemScore	11.04	10.25	5.97
PMF	9.81	9.05	4.06
SCSreg	15.94	11.46	6.93
SCS_{class}(7 Å)	15.94	8.44	6.69
SCS_{class}(8 Å)	**19.62**	12.06	7.64

Table 5. Spearman Correlation Coefficient of Each Scoring Method (SCS, F-Score, G-Score, D-Score, ChemScore, and PMF)

scoring method	TK	thrombin	PPARγ
SCSreg	0.88	0.97	0.95
F-Score	0.38	0.29	0.43
G-Score	0.7	0.64	0.89
D-Score	0.77	0.09	0.92
ChemScore	0.68	0.81	0.93
PMF	0.49	−0.1	−0.12

Note that our goal of this study is not to outperform the five scoring functions and to achieve high enrichments as much as possible based on given scoring functions without any available active compounds except a 3D structure of protein–ligand complex.

Moreover, from above discussions, one can expect that SCS enriches more active compounds if one can use scoring functions that enrich active compounds more than the five scoring functions used in the present study.

Since TK is considered to be a difficult target for docking because of receptor flexibility, a highly exposed binding pocket, the importance of water-bridged interactions, and the low affinity of most known ligands,[19,33] enrichment factors are small overall. However, since our goal of this study is to improve the performance and not to maximize it, our results demonstrate the validness of SCS enough at a conceptual level.

Table 6. Average Centroid Distance for Top-Scored Conformations for Decoys and Ones for Ligands

TK		
scoring method	decoys (Å)	ligands (Å)
F-Score	3.76	2.1
G-Score	1.78	2.13
D-Score	1.72	1.51
ChemScore	2.37	1.83
PMF	2.13	2.64
SCS	1.53	1.85
Thrombin		
scoring method	decoys (Å)	ligands (Å)
F-Score	5.83	2.51
G-Score	5.95	3.02
D-Score	4.91	2.22
ChemScore	4.67	2.15
PMF	5.96	4.02
SCS	5.87	2.37
PPARγ		
scoring method	decoys (Å)	ligands (Å)
F-Score	10.97	7.88
G-Score	10.21	7.27
D-Score	10.19	6.58
ChemScore	9.98	6.19
PMF	11.82	7.59
SCS	10.09	6.18

Although we test SCS by using one protein–ligand structure only for training, it is worthwhile to note how the SCS trained for one protein–ligand structure performs when testing against the same protein but has a different bound structure with different ligands.

3.2. Relationship between Score Distribution of a Native Ligand Conformations and Enrichment. To visualize the correlation between each scoring methods (SCS, F-Score, G-Score, D-Score, ChemScore, and PMF) and the rmsds of native ligands conformations, we show the scatter plots between them in Figure S1–S6 (see the Supporting Information). From Figure S1–S6, it appears that the distribution of rmsd in thrombin and PPARγ is biased. This bias mainly originates from the smallness of binding site, not the flexibility of ligand or docking to multiple regions in the binding site. In SCS, out-of-bag was used to predict rmsds of a native ligand. To quantitatively evaluate the correlation between them, we use the Spearman correlation coefficient (R_s) that gives the correlation between two ranked sets. R_s is 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 ligand conformation under two criteria: the rmsd and score of SCS_{reg}. By definition, R_s ranges from –1 to 1. The larger R_s becomes, the more strongly two sets correlate. Table 5 summarizes the spearman correlation coefficient between five scoring functions and the rmsds of native ligands conformations.

From Tables 4 and 5, the Spearman correlation coefficient between each scoring methods and rmsd of native ligands

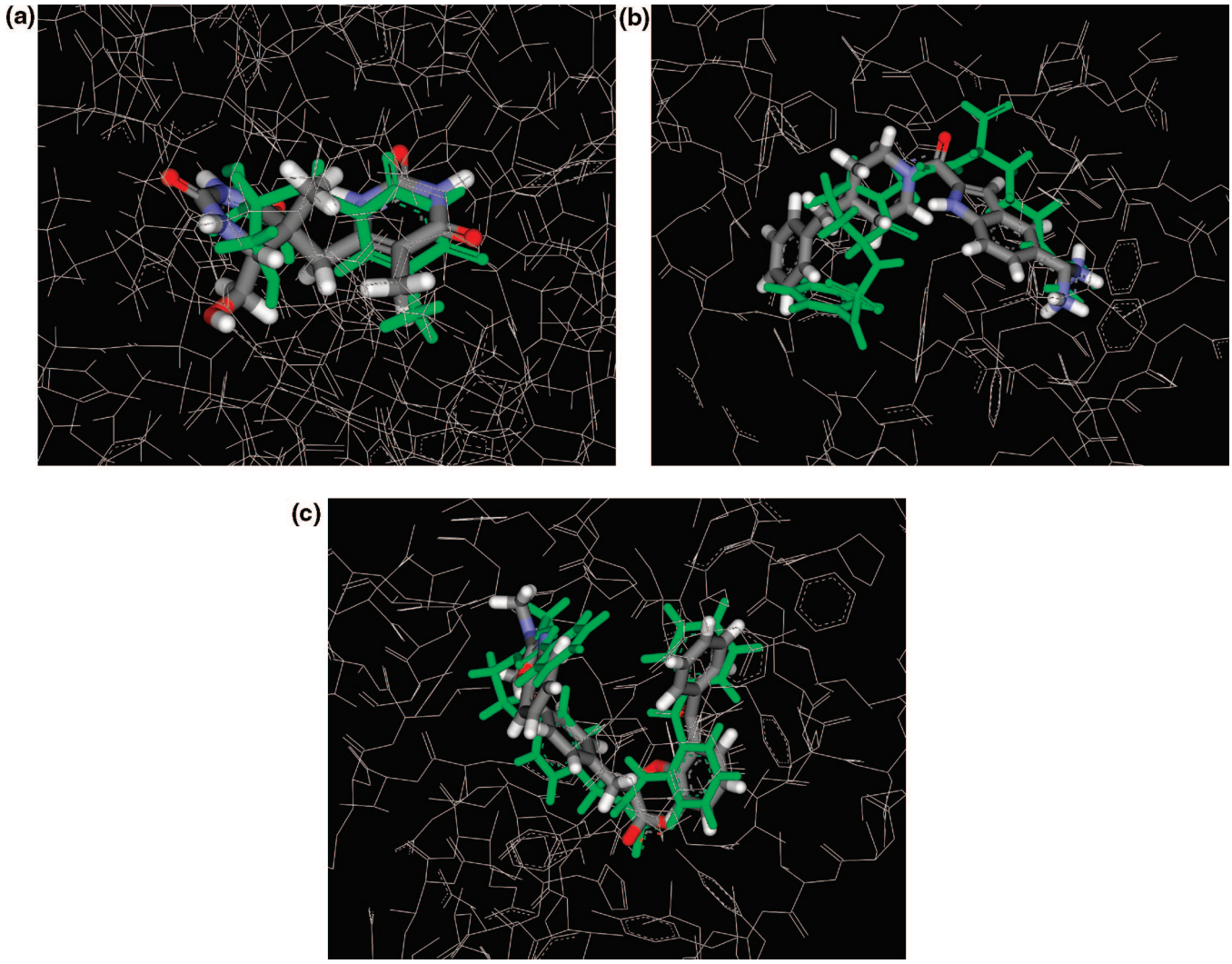

Figure 4. Representative examples of the docking pose close to X-ray structure of native ligand. (a) TK. The molecule in color-by-atom is the docking pose of ZINC03834172, and the green molecule is the X-ray structure of deoxythymidine. The centroid distance is 0.21 Å. (b) Thrombin. The molecule in color-by-atom is the docking pose of ZINC03834114, and the green molecule is the X-ray structure of tripeptidylaldehydes. The centroid distance is 0.23 Å. (c) PPARγ. The molecule in color-by-atom is the docking pose of ZINC03834058, and the green molecule is the X-ray structure of GI262570. The centroid distance is 0.27 Å.

conformations and enrichment factors do not correlate. For TK, although G-Score works best in single scoring functions, the Spearman correlation coefficient is 0.7 and the one of D-Score (0.77) indicates higher correlation. For thrombin, although F-score enriches active compounds, the Spearman correlation coefficient is only 0.29. While the Spearman correlation coefficient of ChemScore is 0.81, ChemScore is largely inferior to F-Score in enrichments. For PPARγ, D-Score works best, but the Spearman correlation coefficients of D-Score, ChemScore and G-Score are high (0.92, 0.93, and 0.89). Thus, these results suggest that it is difficult for single scoring functions to infer which scoring function enriches active compounds a priori even if 3D structure of a protein–ligand complex is available. On the other hand, since SCS and rmsd highly correlate and the enrichments of SCS are competitive or superior to a best single scoring function at the top-ranks screened database, it appear that SCS surely employs a rough linear correlation relationship between binding free energy and the rmsd of a native ligand conformations and 3D structure of a protein–ligand complex is able to provide the useful information for a target specific.

3.3. Evaluation of Docking Poses. To evaluate quality of docking poses, we should define a measure between binding pose of native ligand and docking poses of top-scored compound conformations. Although a docking pose of native ligand is evaluated based on rmsd generally, it is difficult to compare between X-ray structure of a native ligand and top-scored conformations of other compounds. To compare between X-ray structure of a native ligand and ones of other compounds, we used centroid distance (D_{centroid}) between them.

$$D_{\text{centroid}} = \sqrt{(x_i - x_{\text{native}})^2 + (y_i - y_{\text{native}})^2 + (z_i - z_{\text{native}})^2}$$

where x_i, y_i, and z_i are coordinates of centroid of top-scored conformation of compound i, respectively, and x_{native}, y_{native}, and z_{native} are coordinates of centroid of X-ray conformation of native ligand, respectively. Since our purpose is to compare the conformations between different molecules roughly, centroid distance could be appropriate for evaluating scoring methods based on docking poses. Although the centroid of the ligand depends on the shape and size of the ligands, it is not problematic to use the centroid of ligands for statistical evaluation of many docking poses, because we

deal with many ligands with various shape and size and they cancel out the dependency on shape and size each other overall. Table 6 shows average centroid distance for top-scored poses for decoys and ones for ligands. From Table 6, average centroid distance of both decoys and ligands of TK is small overall. This result indicates that most top-scored poses are in the active site and it might become difficult to discriminate ligands and decoys. From above discussion, TK results in poor enrichment overall. For thrombin, while average centroid distance of decoys is large, one of ligands is small overall. This result suggests that most top-scored decoy poses are out of the active site and most ligands poses are in active site. This discussion leads to high enrichments of SCS, F-Score. On the other hand, average centroid distance of D-Score and PMF are relatively small in both decoys and ligands and it might become difficult to discriminate ligands and decoys likewise TK. For PPARγ, while average centroid distance of decoys is large, one of ligands is small overall. This result also suggests that most top-scored decoy poses are out of the active site and that most ligands poses are in active site and leads to high enrichments of SCS, D-Score, and ChemScore.

For three target proteins, the average centroid distance of ligands for SCS is relatively small overall. This result suggests that SCS gives a top score to a ligand conformation near an active site. Since top-scored ligand pose should be near an active site, this characteristic is desirable. Figure 4 shows the representative examples of docking pose close to X-ray structure of native ligands for three target proteins. Since SCS enriches ligands well at the top ranks of the screened compounds only, SCS could not work well far from an active site.

4. CONCLUSION

We applied SCS (supervised consensus scoring), which employs a rough linear correlation between the binding free energy and the rmsd and incorporates protein–ligand binding process with docked ligand conformations using supervised learning, to virtual screening and evaluated the docking poses and the enrichments of SCS and five scoring functions (F-Score, G-Score, D-Score, ChemScore, and PMF) for three different target proteins: TK, thrombin, and PPARγ. Our enrichment studies show that SCS is competitive or superior to a best single scoring function at the top ranks of the screened compounds, which is of interest for practical drug screening. The enrichments of SCS could be limited by a best scoring function, because SCS is obtained based on the five individual scoring functions. Therefore, it is concluded that SCS works very successfully from our results.

Moreover, we confirmed that SCS is able to predict rmsd of a native ligand using out-of-bag likewise the previous study.[17] This means that SCS surely employs a rough linear correlation between the binding free energy and the rmsd for virtual screening.

From docking pose analysis, we made connection between enrichment and average centroid distance for top-scored poses of decoys and ones of ligands and revealed that SCS gives a top score to a ligand conformation near an active site. Since SCS requires only one 3D structure of protein–ligand complex, SCS will be useful for identifying new ligands, when one can not know which scoring function works best against a target protein a priori.

ACKNOWLEDGMENT

The authors thank our colleagues at NEC Corp. for the fruitful discussions we had with them.

Supporting Information Available: Scatter plots for each scoring method and rmsd of native ligand conformations (Figures S1−S6). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Doman, T. N.; McGovern, S. L.; Witherbee, B. J.; Kasten, T. P.; Kurumbail, R; Stallings, W. C.; Connolly, D. T.; Schoichet, B. K. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phophatase-1B. *J. Med. Chem.* **2002**, *45*, 2213–2221.
(2) Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discov. Today* **2002**, *7*, 1047–1055.
(3) Shoichet, B. K.; McGovern, S. L.; Wei, B.; Irwin, J. Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.* **2002**, *6*, 439–446.
(4) Shoichet, B. K. Virtual screening of chemical library. *Nature.* **2004**, *432*, 862–865.
(5) Majeux, N.; Scarsi, M.; Apostolakis, J.; Caisch, A. Exhaustive docking of molecular fragments on protein binding sites with electrostatic solvation. *Proteins* **1999**, *37*, 88–105.
(6) Zou, X.; Yaxiong, S.; Kuntz, I. D. Inclusion of Solvation in Ligand Binding Free Energy Calculations Using the Generalized-Born Model. *J. Am. Chem. Soc.* **1999**, *121*, 8033–8043.
(7) 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.
(8) Mitchell, J. B. O.; Laskowski, R. A.; Alexander, A.; Thornton, J. M. BLEEP-Potential of mean force describing protein-ligand interactions: I. Generating potential. *J. Comput. Chem.* **1999**, *20*, 1165–1176.
(9) 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.
(10) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337–356.
(11) 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.
(12) 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.
(13) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of docking performance: Comparative data on docking algorithms. *J. Med. Chem.* **2004**, *56*, 558–565.
(14) 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.
(15) Kontoyianni, M.; Sokol, G. S.; McClellan, L. M. Evaluation of library ranking efficacy in virtual screening. *J. Comput. Chem.* **2005**, *26*, 11–22.
(16) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, 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.
(17) Teramoto, R; Fukunishi, H. Supervised consensus scoring for docking and virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 526–534.
(18) Camacho, C. J.; Vajda, S. Protein docking along smooth association pathways. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 10636–10641.
(19) Huang, N.; Schoichet, K. B.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.
(20) *FlexSIS, Sybyl7.1J*; BioSolveIT GmbH: Sankt Augustin, Germany, 2005.
(21) Jacobsson, M.; Liden, P.; Stjernschantz, E.; Bostrom, H.; Norinder, U. Improvement structure-based virtual screening by multivariate analysis of scoring data. *J. Med. Chem.* **2003**, *46*, 5781–5789.
(22) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287–2303.
(23) 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.

(24) Rarey, M.; Kramer, B.; Lengauer, T. Evaluation of the FlexX incremental construction algorithm for protein-ligand docking. *Proteins* **1999**, *37*, 228–241.
(25) Muegge, I. A knowledge-based scoring function or protein-ligand interactions: probing the reference state. *Perspect. Drug. Discov. Des.* **2000**, *20*, 99–114.
(26) Muegge, I. Effect of ligand volume correction on PMF scoring. *J. Comput. Chem.* **2001**, *22*, 418–425.
(27) 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.
(28) Eldridge, M. D.; Murray, C. W.; Auton, R. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligand in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425–445.
(29) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graphics. Model.* **2002**, *20*, 281–295.
(30) Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5–32.
(31) 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.
(32) Chen, C.; Liaw, L.; Breiman, L. *Using random forest to learn imbalanced data*, Technical Report 666; Statistics Department, University of California at Berkeley: Berkeley, CA, 2004.
(33) 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.

CI700464X