Consensus Scoring with Feature Selection for Structure-Based Virtual Screening

Reiji Teramoto*,† and Hiroaki Fukunishi‡

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

Received July 4, 2007

The evaluation of ligand conformations is a crucial aspect of structure-based virtual screening, and scoring functions play significant roles in it. While consensus scoring (CS) generally improves enrichment by compensating for the deficiencies of each scoring function, the strategy of how individual scoring functions are selected remains a challenging task when few known active compounds are available. To address this problem, we propose feature selection-based consensus scoring (FSCS), which performs supervised feature selection with docked native ligand conformations to select complementary scoring functions. We evaluated the enrichments of five scoring functions (F-Score, D-Score, PMF, G-Score, and ChemScore), FSCS, and RCS (rank-by-rank consensus scoring) for four different target proteins: acetylcholine esterase (AChE), thrombin (thrombin), phosphodiesterase 5 (PDE5), and peroxisome proliferator-activated receptor gamma (PPARγ). The results indicated that FSCS was able to select the complementary scoring functions and enhance ligand enrichments and that it outperformed RCS and the individual scoring functions for all target proteins. They also indicated that the performances of the single scoring functions were strongly dependent on the target protein. An especially favorable result with implications for practical drug screening is that FSCS performs well even if only one 3D structure of the protein−ligand complex is known. Moreover, we found that one can infer which scoring functions significantly enrich active compounds by using feature selection before actual docking and that the selected scoring functions are complementary.

1. INTRODUCTION

Protein−ligand docking is widely used to discover novel ligands in structure-based virtual screening. Over the past 15 years, various docking programs have been developed and evaluated.[1−4] These docking programs attempt to predict the binding conformations of ligands 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, including ones based on 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 evaluates the protein−ligand affinity. Scoring functions are important because the final predicted conformations are selected according to the scores.

There are three kinds of scoring function: force-field-based methods, empirical scoring functions, and knowledge-based potentials. Force-field-based scoring functions apply molecular mechanics energy functions. 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 using a distance-dependent dielectric function, although there are also solvent models based on continuum electrostatics.[5,6] Hydrophobic 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 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, whereby 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 enough information to determine 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] They differ in their definitions of the reference state, the protein and ligand atom types, and the lists of protein−ligand complexes from which they were extracted.

Many reports assessing the performance of docking programs have been published and concluded that docking algorithms reproduce binding modes very well and that

* 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/ci700239t CCC: $40.75
Published on Web 01/30/2008

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

protein	PDB code	resolution (Å)	no. of ligands	no. of decoys	protein family
AChE	1eve	2.5	105	3848	serine esterase
thrombin	1ba8	1.8	70	2441	serine protease
PDE5	1xp0	1.8	87	1967	metalloenzyme
PPARγ	1fm9	2.1	84	3006	nuclear hormone receptor

scoring functions are less successful at identifying binding modes and enhancing enrichments.[11−16]

Recently, it has been reported that consensus scoring (CS) improves enrichments by compensating for the deficiencies of individual scoring functions.[17−28] There are several CS strategies, such as rank-by-rank, rank-by-number, average rank, and linear combinations of scoring functions. The drawback is that CS requires many known active compounds in order to determine the best combination of scoring functions. Moreover, the potential value of consensus scoring might be limited, if terms in different scoring functions are significantly correlated, which could amplify errors rather than balance them and compensate for the deficiencies of each scoring function.

Although the prior studies on consensus scoring are significant,[17−28] the important issue of how individual scoring functions should be selected remains a challenging task when few known active compounds are available.

To address this problem, we proposed feature selection-based consensus scoring (FSCS), which performs supervised feature selection with docked native ligand conformations to select complementary scoring functions. FSCS employs a rough linear correlation between the binding free energy and the RMSD of a native ligand and incorporates a protein−ligand binding process.[25,29] We evaluated the enrichments of FSCS with five scoring functions (F-Score, D-Score, PMF, G-Score, and ChemScore) for four different target proteins: acetylcholine esterase (AChE), thrombin (thrombin), phosphodiesterase 5 (PDE5), and peroxisome proliferator-activated receptor gamma (PPARγ). We compared FSCS, rank-by-rank consensus scoring (RCS) without feature selection, and individual scoring functions on the basis of enrichments with benchmarking sets for molecular docking.

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 that avoid the bias of decoys. Moreover, since DUD is freely available, it is an easy reference by which to compare scoring methods. We collected 3D structures of four target proteins, i.e., AChE, thrombin, PDE5, and PPARγ, their ligands, and decoys from DUD, to evaluate ligand enrichment.[30] These target proteins were chosen as the representative proteins from different protein families: AChE, serine esterase; thrombin, serine protease; PDE5, metalloenzyme; and PPARγ, nuclear hormone receptor. DUD stores many decoys that physically resemble ligands, so that enrichment will not simply be a separation of gross features and will be chemically distinct from them. Table 1 summarizes the test data sets. A detailed description 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 native ligand conformations
AChE	E2020(aricept)	299
thrombin	tripeptidylaldehydes	393
PDE5	vardenafil	293
PPARγ	GI262570	329

2.2. Docking Procedure and Scoring Functions. FlexSIS implemented in Sybyl7.1J was employed to generate an ensemble of docked conformations for each ligand.[31] 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. We generated 999 docked conformations for native ligands at a maximum. Note that we can only set the maximum number of docked conformations and not set the number of docked conformations generated directly. Table 2 summarizes the number of docked conformations of native ligands. Figure 1 shows the RMSD distributions of each native ligand conformation generated by FlexSIS. We generated 100 docked conformations for other compounds, i.e., known ligands and decoys, at a maximum. All docked conformations generated by FlexSIS were scored using F-Score,[32] D-Score,[33] PMF,[34−36] G-Score,[33] and ChemScore[37] implemented in CScore.[38] To represent the correlations between the each scoring functions and RMSD, we show the scatter plots between them in Figures S1−S5 (Supporting Information). 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 for docking and structure-based virtual screening.[16,22]

2.3. Consensus Scoring (CS). CS combines multiple scoring functions and improves hit rates. We used the rank-by-rank strategy, which is a representative and superior CS approach.[22,25] The rank-by-rank strategy is illustrated in Figure 2, and RCS is defined as

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

where R_i is the rank of the top-ranked ligand conformation for the ith scoring function in the screening database, and N is the number of scoring functions. We used all scoring functions and evaluated the enrichments in ascending order of RCS.

2.4. Feature Selection-Based Consensus Scoring (FSCS). Figure 3 illustrates the overall FSCS procedure, and Figure 4 illustrates the binding free energy landscape when RMSD is used as reaction coordinate. We assume that protein−ligand binding has a funnel-shaped landscape, as discussed by Camacho and Vajda.[29] FSCS employs this rough linear correlation relationship between binding free energy and the RMSD of a native ligand.[25] The binding energy prediction of a native ligand is formulated in terms of supervised learning in which explanatory attributes and an objective variable are the scores of five scoring functions (F-Score, D-Score, PMF, G-Score, and ChemScore) and the RMSD

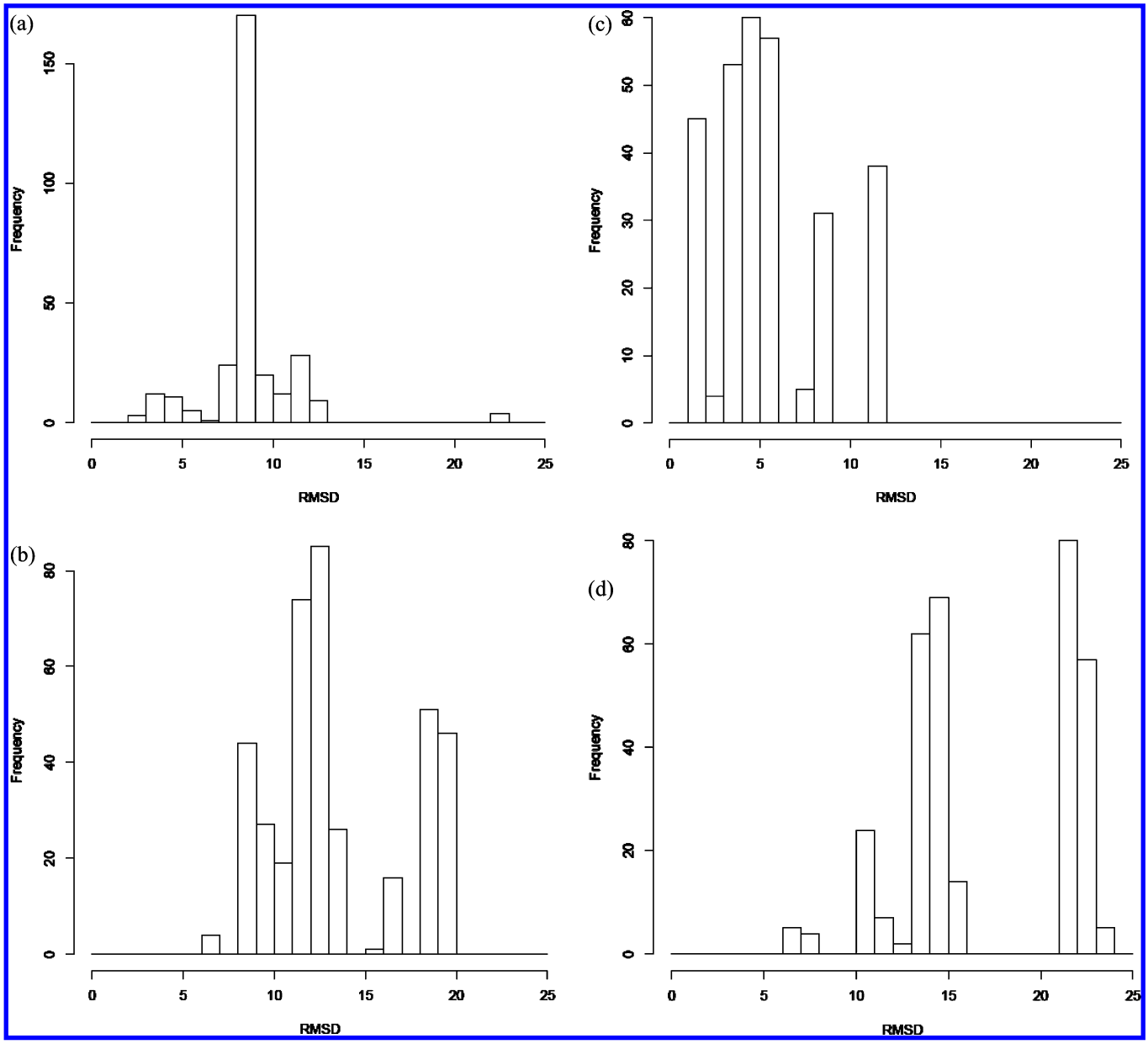

Figure 1. RMSD distributions of native ligand conformations. (a) AChE (PDB code 1eve), (b) thrombin (PDB code 1ba8), (c) PDE5 (PDB code 1xp0), and (d) PPARγ (PDB code 1fm9).

of a native ligand. This formulation intends that FSCS estimates the contributions of scoring functions to the binding energy via RMSD. If the coefficient of a scoring function is positive and large, then it significantly contributes to the binding energy, and if it is negative, then it does not contribute to the binding energy. Consequently, FSCS can infer which scoring functions contribute to the binding energy from their coefficient. This procedure is called *feature selection* in machine learning. We use support vector regression (SVR) with a linear kernel as the supervised learning algorithm, because it explicitly provides the coefficient of a scoring function of a linear regression model and is based on statistical learning theory.[39,40] Other learning machines, such as multiple regression, may also be used. However, since SVR does not require the assumption of normal distribution for data for learning and multiple linear regression requires it, we used SVR as a learning machine. We used the default parameters of SVR in WEKA. The WEKA package is available at http://www.cs.waikato.ac.nz/ml/weka/.

After feature selection, we employ RCS with the selected scoring functions and evaluate the enrichments in ascending order of RCS.

2.5. Support Vector Regression (SVR). SVR is a development of support vector machines that introduces an ϵ-insensitive loss function.[39,40] In SVR, with the input data set $\{(x_i,y_i)_1^n\}$ (where x_i is the input vector (scoring functions), y_i is the desired real value (RMSD of a native ligand), and n is the number of input records) SVR approximates the following linear function

$$y = f(x) = w \cdot x + b$$

where b and w are constants. w is estimated by minimizing the following equation

$$R(C) = \frac{1}{2}||w||^2 + C\cdot\frac{1}{n}\sum_{i=1}^{n} L_\epsilon(y_i, f(x_i))$$

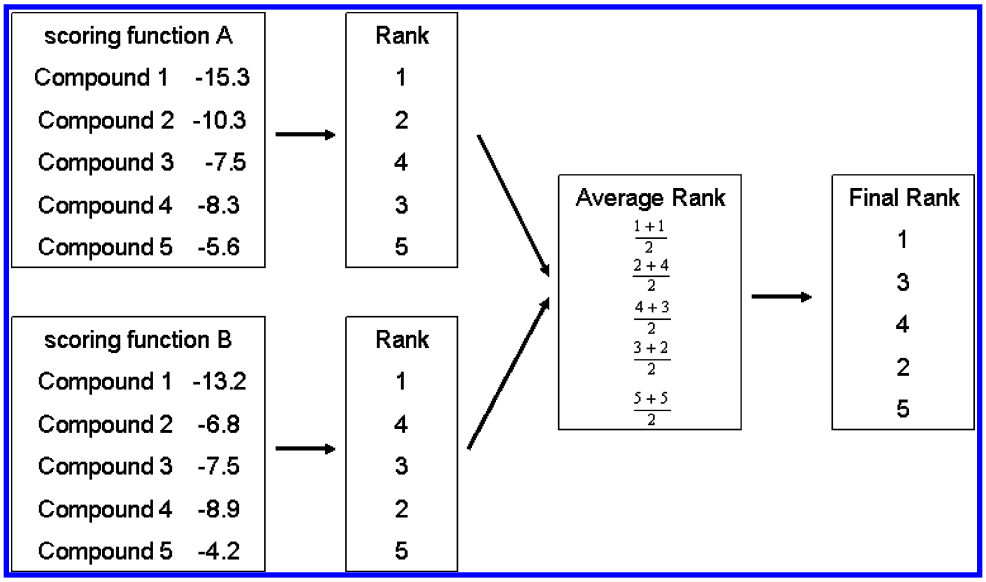

Figure 2. Overview of the rank-by-rank strategy for two scoring functions.

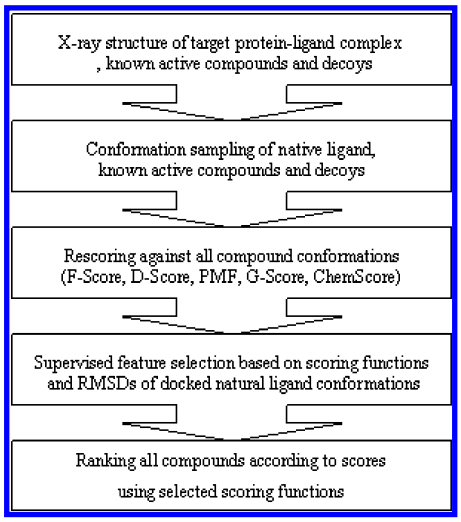

Figure 3. Overview of FSCS procedure.

where $L_\epsilon(y, f(x))$ is the empirical error measured by an ϵ-insensitive loss function

$$L_\epsilon(y, f(x)) = \max\{0, |y - f(x)| - \epsilon\}$$

The constant C is specified empirically, and it determines the tradeoff between the empirical risk and the regularization term. ϵ is an empirical quantity, and it is equivalent to the approximation accuracy of the training data.

w and b are obtained by solving the following optimization problem.

$$R(C) = \frac{1}{2}||w||^2 + C\cdot\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

$$\text{subject to} \begin{cases} y_i - w\cdot x - b_i \le \epsilon + \xi_i \\ w\cdot x + b_i - y_i \le \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases}$$

By introducing Lagrange multipliers, the primal function can be transformed into a quadratic programming problem. The solution takes the following form:

$$f(x, a_i, a_i^*) = \sum_{i=1}^{n}(a_i - a_i^*)x\cdot x_i + b$$

Thus, one can explicitly obtain the coefficients of the linear regression function. Nonlinear regression could be performed by substituting kernel function $K(x\cdot x_i)$ instead of the inner product $x\cdot x_i$. However, since our purpose is to explicitly infer the contributions of scoring functions to the binding energy from the coefficients of a linear regression function and nonlinear regression does not provide interpretability for the contribution of scoring functions, nonlinear regression is not appropriate for feature selection in this study. We used the default parameters of SVR in WEKA ($C = 1.0$, $\epsilon = 0.001$).

3. RESULTS AND DISCUSSION

3.1. RMSD Distributions of Native Ligand Conformations. As shown in Figure 1(b),(d), conformations of less than 5 Å were not sampled for thrombin and PPARγ. As can be seen in Figure 1(a), the conformations less than 5 Å were relatively small in number but were sampled over a broad range for AChE. In contrast, there were many conformations less than 5 Å for PDE5 in Figure 1(c). Note that FSCS works well even though docked conformations of native ligands are not sampled near X-ray structure enough, because SVR approximates a rough linear relationship between RMSDs and scoring functions near X-ray structure by docked conformations of a native ligand in the process of learning them.

3.2. Contribution of Scoring Functions to Binding Free Energy. Since FSCS estimates the contribution of scoring functions to the binding energy via RMSD, SVR with a linear kernel requires strong correlation between RMSD and predicted RMSD of native ligands. To ensure this, we evaluated the correlation between the value predicted by SVR

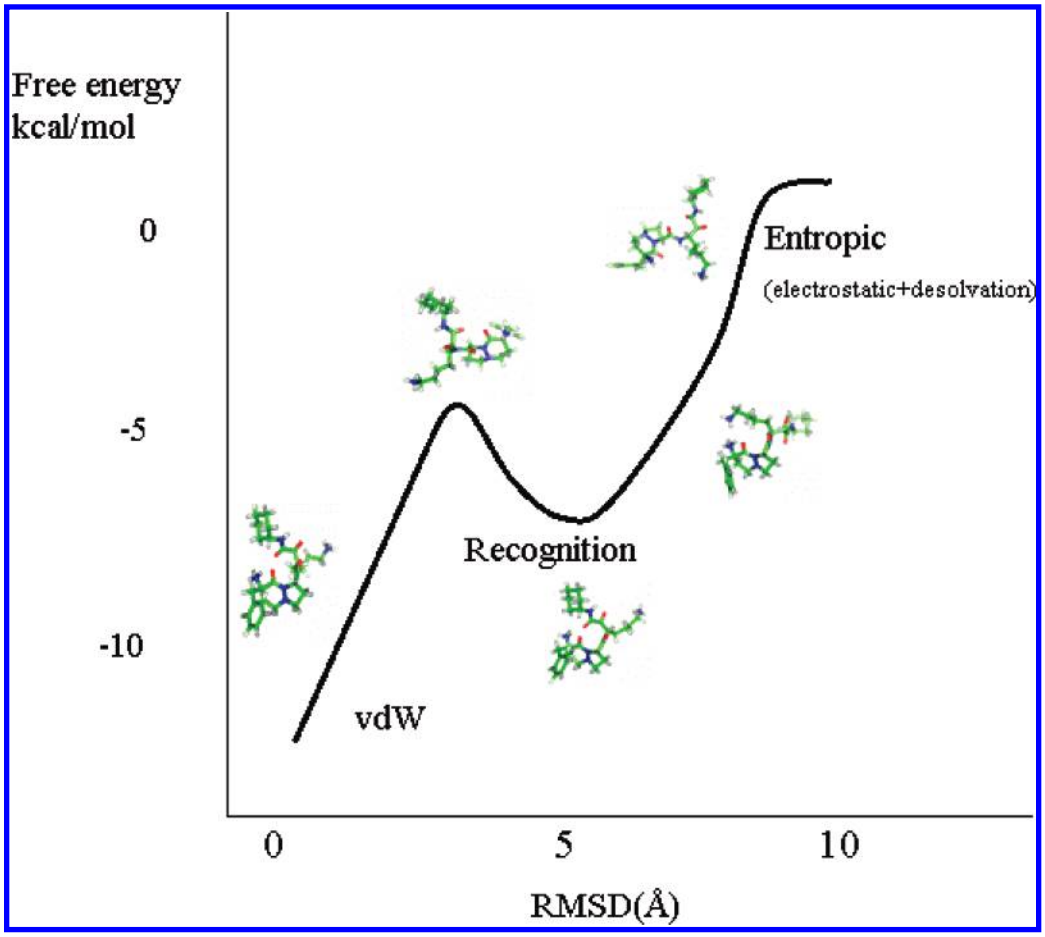

Figure 4. Binding free energy landscape when the RMSD is taken as the reaction coordinate. FSCS employs a rough linear correlation between binding free energy and the RMSD of a native ligand to predict the binding energy. FSCS estimates the contribution of scoring functions to the binding energy via RMSD.

Table 3. Correlation Coefficient and RMSE for Each Target Protein

	AChE	thrombin	PDE5	PPARγ
correlation coefficient	0.71	0.89	0.92	0.96
RMSE	1.79	1.7	1.23	1.32

and the RMSD of native ligand conformations by 10-fold cross-validation. We employed the Pearson correlation coefficient (r_{xy}), which gives the correlation between two sets of variables $\{x\},\{y\}$ as an evaluation measure. r_{xy} is defined as

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \cdot \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

where n is the number of docked ligand conformations, and x and y are the actual RMSD value and the predicted one, respectively. By definition, r_{xy} ranges from -1 to 1. The larger r_{xy} becomes, the more strongly the two sets correlate. We normalized the scoring functions and RMSD to avoid bias from their magnitudes.

Table 3 lists the Pearson correlation coefficients between the value predicted by SVR and RMSD of native ligand conformations and the root mean squared error (RMSE) by 10-fold cross-validation. It appears that the RMSD and the value predicted by SVR are well correlated and that the correlation coefficient and RMSE are inversely correlated overall. Although the correlation coefficient of AChE is lower than the others, its RMSE is not so large and almost the same as thrombin. The low correlation coefficient originates from the difference in the distributions of docked ligand conformations generated. These results suggest that linear regression models built by SVR adequately represent the protein−ligand binding process. Thus, we can confirm that feature selection can be performed using valid models to estimate the contributions of scoring functions to the binding energy via RMSD. Since correlation coefficients became high enough using the default parameters for SVR in this study, we used them. However, if correlation coefficient is low, e.g., less than 0.6, then one should explore the parameters for SVR in order that correlation coefficient becomes high.

The linear regression function is concretely defined as

$$\text{RMSD} = a\cdot\text{F-Score} + b\cdot\text{D-Score} + c\cdot\text{PMF} + d\cdot\text{G-Score} + e\cdot\text{ChemScore} + f$$

where a, b, c, d, and e are the coefficients of each scoring function, and f is a regression constant. Table 4 shows the coefficients of each scoring function obtained from linear regression models for four target proteins. As discussed in section 2.4, if the coefficient of a scoring function is positive and large, then it significantly contributes to the binding

Table 4. Coefficients of Scoring Functions of the Linear Regression Model

scoring functions	AChE	thrombin	PDE5	PPARγ
(a) F-Score	−0.34	0.65	0.17	−0.28
(b) D-Score	0.23	0.04	−0.03	1.01
(c) PMF	0.07	−0.05	0.67	0.23
(d) G-Score	0.07	0.41	−0.1	−0.44
(e) ChemScore	0.15	0.67	0.31	0.56

energy, and if it is negative, then it does not contribute the binding energy.

3.3. Performance Evaluation of Virtual Screening. Our main objective is to compare the five scoring functions (F-Score, D-Score, PMF, G-Score, and ChemScore), FSCS, and RCS when they are applied to the same target proteins. At a coarse level, virtual screening tests the ability of scoring methods to differentiate between active and inactive compounds. Figure 5 shows the overall profiles of percentage of ligands found (*y*-axis) plotted as a function of the percentage of the ranked docked database (*x*-axis) for AChE, thrombin, PDE5, and PPARγ by the five scoring functions, FSCS, and RCS. As another indicator of performance, we use the enrichment factor (EF) 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. By definition, 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), EF_5 (enrichment factor at 5% of the ranked database), EF_{10} (enrichment factor at 10% of the ranked database), and EF_{20} (enrichment factor at 20% of the ranked database) (Table 5). In Figure 5 and Table 5, FSCS (single) indicates RCS with a single selected scoring function, FSCS (double) indicates RCS with two selected scoring functions, and FSCS (triple) indicates RCS with three selected scoring functions.

Although the performance of FSCS depends on the target proteins, Table 5 shows that FSCS (double) exhibits high enrichments for all target proteins and outperforms RCS and individual scoring functions overall. FSCS performs especially well at the top ranks of the screened compounds; this result should be of interest to those involved in practical drug screening.

For AChE and PDE5, there is a great difference in enrichments between FSCS (double) and RCS. For thrombin, FSCS and RCS are competitive at EF_1 and EF_2. For PPARγ, although FSCS (single) is D-Score and enriches active compounds most at EF_1 and EF_2, FSCS (double) is competitive and enriches active compounds most at EF_5 and EF_{10}.

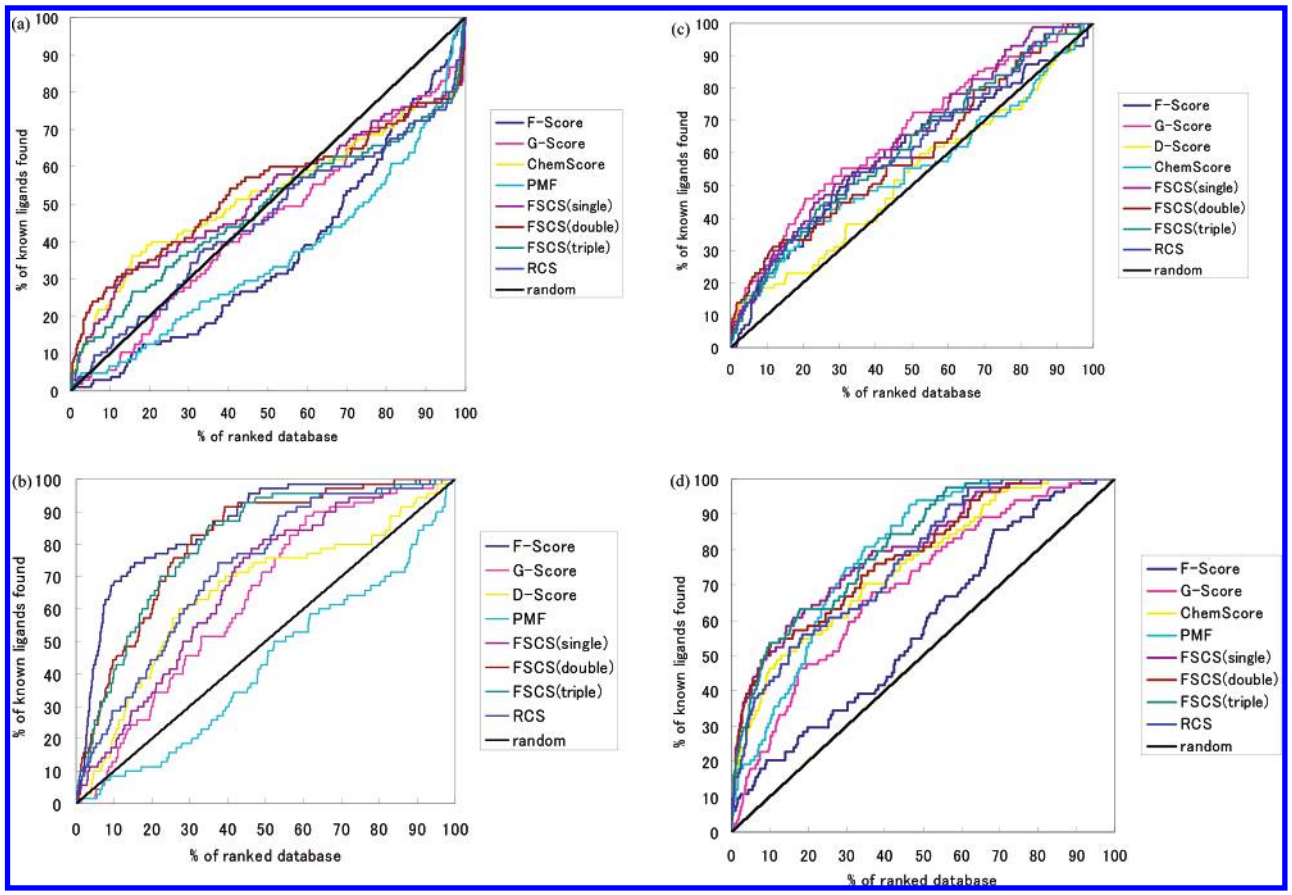

Figure 5. Docking enrichment plots for AChE, thrombin, PDE5, and PPARg using five scoring functions (F-Score, D-Score, PMF, G-Score, and ChemScore), FSCS and RCS. (a) AChE, (b) thrombin, (c) PDE5, and (d) PPARγ. The docking ranked database (*x*-axis) is plotted against the percentage of known ligands found by each scoring method (*y*-axis) at any given percentage of the ranked database.

Table 5. Enrichment Factors of Five Scoring Functions (F-Score, D-Score, PMF, G-Score, and ChemScore), FSCS, and RCS

scoring method	EF_1	EF_2	EF_5	EF_{10}	EF_{20}
(a) AChE					
F-Score	0.95	0.48	0.19	0.29	0.57
G-Score	0.95	0.95	0.57	0.48	0.76
ChemScore	4.75	3.8	2.85	2.19	**1.86**
PMF	0	1.43	0.95	0.48	0.57
FSCS(single)(D-Score)	2.85	3.8	2.85	2	1.67
FSCS(double)(D-Score+ ChemScore)	**8.55**	**6.18**	**4.18**	**2.76**	1.71
FSCS(triple)(D-Score+ ChemScore+PMF)	4.75	3.33	2.66	1.72	1.33
RCS	1.9	1.43	0.95	1.05	1
(b) Thrombin					
F-Score	4.3	5.74	**8.54**	**6.72**	**3.79**
G-Score	0	0.72	0.28	1.29	1.5
D-Score	0	0.72	1.99	2.14	2.07
PMF	0	0.72	0.28	0.86	0.57
FSCS(single)(ChemScore)	2.87	2.87	2.28	1.71	1.71
FSCS(double)(ChemScore+ F-Score)	**8.61**	**7.17**	4.84	4.43	2.93
FSCS(triple)(ChemScore+ F-Score+G-Score)	**8.61**	4.3	5.12	4	3.22
RCS	**8.61**	5.02	3.42	2.86	2.22
(c) PDE5					
F-Score	0	1.15	1.38	2.42	1.55
G-Score	5.62	5.18	**3.67**	2.53	**2.13**
D-Score	5.62	5.76	2.98	1.84	1.15
ChemScore	4.5	4.03	3.21	2.19	1.84
FSCS(single)(PMF)	6.75	4.03	2.75	2.3	1.9
FSCS(double)(PMF+ ChemScore)	**10.12**	**6.91**	3.44	**2.76**	1.67
FSCS(triple)(PMF+ ChemScore+F-Score)	4.5	3.46	2.75	2.07	1.78
RCS	5.62	4.03	2.75	2.42	1.84
(d) PPARγ					
F-Score	5.93	4.75	2.37	2.02	1.49
G-Score	2.37	1.78	3.56	2.5	2.38
ChemScore	10.68	10.09	5.93	4.64	2.74
PMF	9.49	8.9	4.03	3.1	2.62
FSCS(single)(D-Score)	**18.99**	**14.24**	7.83	5	**3.15**
FSCS(double)(D-Score+ ChemScore)	16.61	13.65	**8.31**	**5.24**	2.92
FSCS(triple)(D-Score+ ChemScore+PMF)	17.8	11.87	7.59	**5.24**	**3.15**
RCS	15.43	8.31	6.65	4.17	2.8

No single scoring function performed well for all target proteins.[16] Moreover, the performances of the scoring functions strongly depended on the target protein.

3.4. Relationship between the Coefficients of Scoring Functions and the Enrichments. From Table 4, Figures 5, and Tables 5, it appears that there is a correlation between the coefficients of scoring functions and enrichments. For AChE, the coefficient of F-Score is negative, and the enrichment of F-Score is inferior to random screening. In contrast, the coefficients of D-Score and ChemScore are positive and large, and their enrichments are high. Although, this relationship holds for thrombin, it does not hold for PDE5 or PPARγ. However, there is a relatively strong correlation between the coefficients of scoring functions and enrichments for PDE5 and PPARγ. These differences depend on the protein−ligand complexes used or the simplistic nature of the scoring function. However, regardless of these limitations, it is very interesting that one can infer which scoring functions enrich active compounds a great deal with only one 3D structure of the protein−ligand complex before actual docking. It appears that these successful estimations prove that FSCS works well. Thus, it seems that FSCS is also able to infer which scoring functions contribute to binding energy via the coefficients of the scoring functions in the linear regression functions.

In this study, we applied FSCS to FlexSIS, which implements the incremental construction algorithm for conformation sampling, and employed no other conformation sampling because of our limitation on available docking programs. However, since conformation sampling algorithms are only used to generate training data for FSCS, it appears that FSCS will perform well when it is applied to other conformation sampling algorithms, e.g., Monte Carlo simulation, or simulated annealing.

3.5. Combination of Scoring Functions. Table 5 indicates that the combination of scoring functions differs according to the target proteins. For AChE, FSCS (double) employs D-Score and ChemScore. D-Score is a force-field-based scoring function, while ChemScore is an empirical scoring function. Thus, different types of scoring functions are combined for CS, and this results in higher enrichment through complementary effects. They suggested that the enrichment can be improved if the individual scoring functions are distinctive, and our result is consistent with their suggestion. Furthermore, FSCS (triple) employs D-Score, ChemScore, and PMF. PMF is a knowledge-based scoring function and is different from D-Score and ChemScore. This clear complementation of scoring functions hold true for PDE5 and PPARγ. For thrombin, FSCS (double) employs ChemScore and F-Score. Both scoring functions are empirical scoring functions, but they differ in their accounting for neural and ionic hydrogen bonds.[22,32,37] This difference may significantly affect the enrichment for thrombin. Since AChE, thrombin, PDE5, and PPARγ belong to quite different protein families, it is likely that a different combination of scoring functions works better for different families. Although we used only five scoring functions in this study, if more scoring functions were used, then one would be able to enrich active ligands to a higher level.

4. CONCLUSION

We proposed feature selection-based consensus scoring (FSCS) that performs supervised feature selection with docked native ligand conformations to select complementary scoring functions. Our results show that FSCS can select the combination of scoring functions that enhances enrichments and that it outperforms RCS and individual scoring functions for all target proteins overall. They also show that the performances of single scoring functions are strongly dependent on the target protein and are unstable. It is quite favorable for practical drug screening that FSCS performs well even when only one 3D structure of the protein−ligand complex is known. Moreover, we demonstrated that one can infer which scoring functions significantly enrich active compounds by feature selection before actual docking, and the selected scoring functions are complementary. Thus, supervised feature selection is quite useful for selection of scoring functions to get higher enrichments. Our results also suggested that there is a correlation between the coefficients of scoring functions and enrichments. Since FSCS is easily applied to other docking programs and scoring functions, one can determine the best combination of scoring functions

for a target protein. Moreover, the feature selection used in FSCS is applicable to the selection of terms of a scoring function for providing a target-specific scoring function.

Since FSCS is only applicable to single active compound per target, the extension method to be applied to multiple active compounds per target should be developed in the future. In addition, since FSCS is based on linear regression, the feature selection methods without assuming any functional relation between energy and RMSD should be also developed.

Regardless of the simplistic nature of the scoring functions, e.g., lack of protein flexibility and inadequate treatment of solvation, we demonstrated that FSCS performs well. If more appropriate scoring functions including such effects are used, then FSCS will perform even better.

ACKNOWLEDGMENT

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

Supporting Information Available: Scatter plots between the five scoring functions (F-Score, D-Score, PMF, G-Score, and ChemScore) and RMSDs (Figures S1−S5). 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 phopphatase-1B. *J. Med. Chem.* **2002**, *45*, 2213−2221.

(2) Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discovery 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 salvation. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 88−105.

(6) Zou, X.; Sun, Y.; Kuntz, I. D. *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) 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.

(18) 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.

(19) 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.

(20) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.

(21) 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.

(22) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.

(23) 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.

(24) 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.

(25) Teramoto, R; Fukunishi, H. Supervised consensus scoring for docking and virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 526−534.

(26) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Model.* **2001**, *41*, 1422−1426.

(27) Fukunishi, Y.; Hojo, S.; Nakamura, H. An efficient in silico screening method based on the protein-compound affinity matrix and its application to the design of a focused library for cytochrome P450 (CYP) ligands. *J. Chem. Inf. Model.* **2006**, *46*, 2610−2622.

(28) Betzi, S.; Suhre, K.; Chetrit, B.; Guerlesquin, F.; Morelli, X. GFscore: a general nonlinear consensus scoring function for high-throughput docking. *J. Chem. Inf. Model.* **2006**, *46*, 1704−1712.

(29) Camacho, C. J.; Vajda, S. Protein docking along smooth association pathways. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10636−10641.

(30) Huang, N.; Schoichet, K. B.; Irwin, J. J.; Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.

(31) *FlexSIS, Sybyl7.1J*; BioSolveIT GmbH: Sankt Augustin, Germany, 2005.

(32) 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.

(33) Rarey, M.; Kramer, B.; Lengauer, T. Evaluation of the FlexX incremental construction algorithm for protein-ligand docking. *Proteins* **1999**, *37*, 228−241.

(34) Muegge, I. A knowledge-based scoring function or protein-ligand interactions: probing the reference state. *Perspect. Drug. Discovery Des.* **2000**, *20*, 99−114.

(35) Muegge, I. Effect of ligand volume correction on PMF scoring. *J. Comput. Chem.* **2001**, *22*, 418−425.

(36) 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.

(37) 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.

(38) 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.

(39) Hastie, T.; Tibshirani, R.; Friedman, J. H. *The Elements of Statistical Learning;* Springer-Verlag: New York, 2001.

(40) Vapnik, V. *Statistical learning theory:* Wiley: New York, 1998.

CI700239T