Comparison of Consensus Scoring Strategies for Evaluating Computational Models of Protein−Ligand Complexes

Akifumi Oda,*,† Keiichi Tsuchida,† Tadakazu Takakura,† Noriyuki Yamaotsu,‡ and Shuichi Hirono‡

Discovery Laboratories, Toyama Chemical Co., Ltd., 2-4-1 Shimookui, Toyama 930-8508, Japan, and School of Pharmaceutical Sciences, Kitasato University, 5-9-1 Shirokane, Minato-ku, Tokyo 108-8641, Japan

Received July 12, 2005

Here, the comparisons of performance of nine consensus scoring strategies, in which multiple scoring functions were used simultaneously to evaluate candidate structures for a protein−ligand complex, in combination with nine scoring functions (FlexX score, GOLD score, PMF score, DOCK score, ChemScore, DrugScore, PLP, ScreenScore, and X-Score), were carried out. The systematic naming of consensus scoring strategies was also proposed. Our results demonstrate that choosing the most appropriate type of consensus score is essential for model selection in computational docking; although the vote-by-number strategy was an effective selection method, the number-by-number and rank-by-number strategies were more appropriate when computational tractability was taken into account. By incorporating these consensus scores into the FlexX program, reasonable complex models can be obtained more efficiently than those selected by independent FlexX scores. These strategies might also improve the scoring of other docking programs, and more-effective structure-based drug design should result from these improvements.

1. INTRODUCTION

Drug design that is based on the three-dimensional (3D) structures of biopolymers, which are candidate drug targets, is commonly referred to as structure-based drug design (SBDD). Predictions of the 3D structures of protein−ligand complexes play an important role in SBDD.[1−3] Over the past 15 years, a variety of computational docking programs that predict protein−ligand complex structures have been developed.[4−7] FlexX is one of the most useful docking programs,[4] in which ligand molecules are divided into small fragments and reconstructed in the active sites of target proteins guided by physicochemical interactions. It is widely used for computational docking trials.

Because docking programs generally produce numerous model candidates for one system, it is essential to evaluate the predicted models and determine which are the most suitable. Although free energy calculations of these systems are required for this purpose, accurate calculations of free energies using molecular simulations are very time-consuming. Therefore, various scoring functions that approximately estimate the binding free energies of protein−ligand systems using simple functions without molecular simulations have been developed.[4−13] Scoring functions can be classified into three groups: empirical, knowledge-based, and force-field-based scoring functions.[3,14] Empirical scoring functions are fit to reproduce experimental data, such as experimentally obtained binding energies and conformations, as a sum of several parametrized functions and are the most widely employed. Knowledge-based scoring functions are derived from experimental structures and are represented by relatively simple atomic interaction-pair potentials. Force-field-based scoring functions are derived from molecular mechanics force-fields and are represented by physicochemical-interaction terms, such as van der Waals potentials and Coulombic interactions. These scoring functions rank the complex structure candidates, and the most highly ranked models are adopted.

As well as their role in model selection, scoring functions have three essential functions in the computational docking process.[1,15] First, during the steps of model construction, the scores of the docking models that are under construction are calculated using scoring functions. The obtained values are then utilized in the next construction step. Thus, scoring functions are required not only in the selection of constructed models but also in the model-construction steps themselves. Second, one or a few predicted models are selected from a large number of model candidates using scoring functions (as mentioned above). Third, in virtual screening trials, in which numerous ligands are docked into one target protein, scoring functions can identify those that potentially represent favorable drug candidates. These three roles are all important for SBDD.

Scoring functions play significant roles in SBDD, and various functions have been proposed. However, building a scoring function that can make use of every protein−ligand system remains the "final frontier" of computational docking studies.[16] Every existing scoring function has specific advantages and disadvantages, and there is no de facto standard. Although this might suggest that different scoring functions are required for different protein−ligand systems, discussions about which scoring function is most suitable for a particular system are often difficult when the complex structures are not experimentally observed. Recently, the concept of a consensus score was proposed for model selection in computational docking and virtual screening

* Corresponding author phone: +81 76 431 8218; fax: +81 76 431 8208; e-mail: AKIFUMI_ODA@toyama-chemical.co.jp.
† Toyama Chemical Co., Ltd.
‡ Kitasato University.

10.1021/ci050283k CCC: $33.50
Published on Web 11/24/2005

trials.[17−23] According to this approach, multiple scoring functions are simultaneously used for model selection or virtual screening, and improvements can be achieved by compensating for the deficiencies of each function. Reasonable complex models that reproduced experimental structures were reported to be more efficiently selected by consensus scores than by independent scoring functions.[21,22] However, few previous studies have directly compared the performance of the various consensus scoring strategies in complex model selection. Although the potential use of consensus scores for compound selection in virtual screening (the "third role") has been discussed both for idealized systems[19] and for real complexes,[17,23] their capacity for model selection (the "second role") has not been adequately discussed. Combinations of scoring functions for one type of consensus scoring strategy, which is referred to as rank-by-rank in this study, were previously investigated.[22] However, systematic discussions of the use of such combinations for other strategies have not yet been performed. Furthermore, although three types of thresholds for model selections can be considered to average-based consensus scores, such as rank-by-number, rank-by-rank,[19] average rank, and linear combination,[23] exhaustive studies about thresholds also have not been carried out. In fact, different definitions of rank-by-rank were applied in refs 19 and 22 (that is, the thresholds were "top 2%" and "best model", respectively), which might contribute to the lack of clarity, as percentages and ranks perform differently in model selection. Furthermore, although a majority-vote-based consensus score that is known as CScore has sometimes been used together with FlexX,[21] there have been no detailed discussions of this approach.

In the present study, the abilities of various consensus scores were compared for model candidate selection in FlexX. For this purpose, nine strategies and 511 (that is, $2^9 - 1$) combinations of nine scoring functions (4599 consensus scores in total) were considered. This is the first study to systematically compare various consensus scores for model selections in computational docking trials, and the results are expected to lead to improvements in FlexX. Moreover, our findings should facilitate protein−ligand complex-structure predictions not only by FlexX but also by other docking programs.

2. METHODS

2.1. The Protein−Ligand Complex Test Set. The test set that was used in this study was constructed from 220 protein−ligand complexes with known structures. These formed part of the FlexX200,[24,25] Glide,[7] and Wang[22] test sets, and only complexes that had at least one docking model whose root-mean-square deviation (RMSD) between calculated and experimental ligand structures is less than or equal to 2.0 Å were selected (these complexes are shown in Table S1 in the Supporting Information). The experimental 3D structures of the protein−ligand complexes were obtained from the Brookhaven Protein Data Bank (PDB).[26] The structures of the target proteins and the ligand molecules were then extracted from the PDB data.

The sets of predicted complex model candidates were obtained using FlexX version 1.11.1 in SYBYL 6.9.[27] The "Num.Answers" parameter, which was the maximum number of output models, was set to 500. Default values of the other parameters and default settings were used. Although the numbers of output models generated by FlexX were not quite equal for all test complexes, the variation of the number of candidates was limited to at most 500 by the "Num.Answers" parameter. The models whose scores were worse than the default threshold and the models which overlapped with target protein atoms were rejected, because they might be outliers. The clustering of generated models was carried out both for base placement and for adding fragment steps (this is the default setting of FlexX). For the docking calculations, the water molecules were removed from the PDB data and the active sites were defined as the collection of amino acids for which at least one atom was closer than 6.5 Å to any non-hydrogen atom of the bound ligand. The ligand molecules that were extracted from the PDB data were assigned appropriate atom and bond types and were subsequently minimized using the MMFF94 force field[28] after filling the valences. Using the FlexX calculations, between 6 and 500 complex model candidates were obtained for each protein−ligand system. The numbers (n) of obtained models are also shown in Table S1 (Supporting Information). In this study, along with the complete test set, a group of complexes with high-affinity ligands only were tested in order to investigate the abilities of the consensus scores. The binding affinities were derived from refs 7 and 22, and the affinity of *N*-phosphoryl-L-leucinamide to thermolysin (PDB ID: 2TMN) was used as the threshold for "high binding affinity"; that is, if the binding affinity of a complex was equal to, or greater than, that of 2TMN, it was classified as having high affinity. The complexes with high-affinity ligands are shown in bold in Table S1 (Supporting Information). Furthermore, test sets including only complexes whose number of generated candidates were more than or equal to 250 were also investigated.

To evaluate the accuracies of the calculated models, the RMSDs between the experimental and computational ligand structures were computed. The experimental structure that was obtained from the PDB was regarded as the "correct" answer, and calculated models with heavy-atom RMSDs that were less than, or equal to, 2.0 Å were defined as "reasonable" models. As it was difficult to obtain an "exact match" with any of the docking programs and scoring functions, "reasonable" model-selection ability was used as the effectiveness criterion for the consensus scores.

2.2. Scoring Functions. Our study employed nine scoring functions: FlexX score,[4] GOLD score,[5] PMF score,[8] DOCK score,[6] ChemScore,[9,10] DrugScore,[11] PLP,[12] ScreenScore,[20] and X-Score.[13] As the consensus scores were built from combinations of scoring functions, we tested all 511 (that is, $2^9 - 1$) permutations, from single scoring functions to a combination of all nine functions. Thus, the best combination was identified from among the 511 possibilities.

Of the nine scoring functions, five (FlexX score, ChemScore, PLP, ScreenScore, and X-Score) were empirical, two (PMF score and DrugScore) were knowledge-based, and two (GOLD score and DOCK score) were force-field-based. FlexX score, GOLD score, PMF score, DOCK score, and ChemScore were calculated using the CScore module in the SYBYL 6.9 package. DrugScore, PLP, and ScreenScore were computed, according to the method described in ref 20, using FlexX. The X-Score was obtained using the stand-alone X-Score 1.1 program.[29] X-Score uses three types of hydro-

Model	Score		Rank
1	−10.2		4
2	−11.1		3
3	−15.4	→	1
4	−5.5		6
5	−9.8		5
6	−12.4		2

Figure 1. Ranking the scores. The complex models were arranged according to their scores, and rankings were assigned.

phobic terms, of which users can select one or more. HS is the hydrophobic term calculated by using a solvent-accessible surface, HP is the pairwise hydrophobic atom contact potential, and HM is calculated by using microscopic matching of hydrophobic ligand atoms to the hydrophobic part of the binding pocket. In this study, the averages of HS, HP, and HM were used for the hydrophobic terms of the X-Score. For scoring functions, a lower value generally indicates a better model; however, the opposite is true for X-Score (that is, a higher value represents a better model). Therefore, the calculated X-Score values were multiplied by −1 before use in the current study.

2.3. Treatment of the Score Values. Because each scoring function evaluates protein−ligand complex models from its own perspective, the absolute values of the scores generally differ from one another. Therefore, the raw scoring function values might be inappropriate for deriving consensus scores. For example, if one of the scores is much bigger than the others, it will strongly influence the average-based consensus score, regardless of the other values. Two preprocessing methods can be used to overcome this problem: ranking the computational models and scaling the scores.

2.3.1. Ranking Method. Using this method, the computational models are arranged according to the score values and the obtained ranks are employed, rather than the raw scores. As shown in Figure 1, the model candidate with the highest score is ranked as number 1, the second highest is ranked as number 2, and so on. In the subsequent consensus scoring process, the rank numbers are used in a similar way to the score values. For example, if there are 500 models for one protein−ligand system, a rank number from 1 to 500 is assigned to each model candidate. This ranking is repeated for all nine scoring functions, and the rank numbers always fall between 1 and 500, even if raw score values are widely spread. Using the rank number, different scoring functions that evaluate the models from different perspectives can be applied simultaneously in consensus scoring procedures. The rank-by-rank method proposed in ref 19 is an example of a consensus scoring method that uses rank numbers.

2.3.2. Scaling Method. Scaling is the second method by which different scoring functions can be used in consensus scoring procedures. In the current study, autoscaling was employed for this purpose. The score of each model was scaled to a number between 0 and 1 using the following formula:

$$x_{\text{scaled}} = (x - x_{\text{min}})/(x_{\text{max}} - x_{\text{min}}) \quad (1)$$

Here, x_{scaled} is the scaled value of the i_Score (for example, the FlexX score), x is the raw value of the i_Score, x_{min} is the smallest i_Score value of all of the model candidates, and x_{max} is the largest i_Score value of all of the models. As shown in Figure 2, all of the scores can be scaled to values in [0, 1] for each type of scoring function using autoscaled score values, x_{scaled}, rather than raw values. Therefore, all of the scoring functions can be simultaneously employed in consensus scoring procedures, regardless of the differences between them. We previously reported this method, which is an improved consensus score system based on the rank-by-number approach mentioned in ref 19 using autoscaled score values. This strategy is known as the average of autoscaled scores (AASS), as shown in eq 2.[30]

$$x^{\text{AASS}} = \frac{\sum_{\text{i\_Score}} (x^{\text{i\_Score}} - x_{\text{min}}^{\text{i\_Score}})/(x_{\text{max}}^{\text{i\_Score}} - x_{\text{min}}^{\text{i\_Score}})}{n} \quad (2)$$

Here, x^{AASS} is the AASS value, i_Score is the scoring function included in the AASS, $x^{\text{i\_Score}}$ is the raw value of the i_Score, $x_{\text{min}}^{\text{i\_Score}}$ and $x_{\text{max}}^{\text{i\_Score}}$ are the smallest and largest values of i_Score, respectively, and n is the number of scoring functions including in the AASS.

2.4. Consensus Scoring Strategies. In the present study, the names of the consensus scoring strategies were clarified in order to clearly distinguish between the different approaches, as diverse techniques have previously been given the same name despite the different types of threshold involved. Recently, several consensus scoring strategies have been proposed. In ref 19, three types of consensus scores were introduced and referred to as "rank-by-number", "rank-by-rank", and "rank-by-vote". Although they were originally developed for virtual screening trials, these techniques can also be used for model selection. The rank-by-number and rank-by-rank consensus scores were obtained using the averages of the score values and the rank numbers, respectively. These average-based consensus scores can be used

	Non-scaled				Auto-scaled		
model	A score	B score	C score		A score	B score	C score
1	−10.2	−145.2	−290.1		0.525	0.564	0.671
2	−11.1	−200.1	−288.2	auto-scaling	0.434	0.000	0.911
3	−15.4	−177.5	−295.4	→	0.000	0.232	0.000
4	−5.5	−102.8	−287.5		1.000	1.000	1.000
5	−9.8	−124.5	−290.5		0.566	0.777	0.620
6	−12.4	−160.0	−292.2		0.303	0.412	0.405

Figure 2. Autoscaling.

Table 1. Consensus Scores

	consensus scoring method according to ref 19[a]	
threshold	rank-by-number	rank-by-rank
number-by	number-by-number	number-by-rank
rank-by	rank-by-number	rank-by-rank
percent-by	percent-by-number	percent-by-rank

[a] Note that the names of the consensus scoring strategies in the current study differ from those given in ref 19.

for model selections with three types of selection criteria: in the first, models with consensus score values that are less than or equal to $x_{threshold}$ are selected; in the second, the models are ranked according to the consensus scores and the top $y_{threshold}$ models are selected; in the third, the top $z_{threshold}$ percent candidates are selected. No systematic studies of these three criteria have been reported previously. Hence, in the present study, we compared these criteria for both rank-by-number and rank-by-rank consensus scores. A total of six types of average-based consensus score were compared in terms of their effectiveness for model selection. The first criterion, under which the consensus score values themselves are used for model selection, was denoted using the prefix "number-by-" in this study, whereas the second and third criteria were indicated by the prefixes "rank-by-" and "percent-by-", respectively. Because names such as number-by-rank-by-number (referring to the rank-by-number consensus score with the number-by criterion) were considered to be too long, we chose to describe this type of consensus scoring strategy as number-by-number. In the same way, the terms number-by-rank, rank-by-number, rank-by-rank, percent-by-number, and percent-by-rank were used. Note that the meanings of rank-by-number and rank-by-rank in this study differ from those employed in ref 19. The AASS corresponded to the number-by-number, rank-by-number, and percent-by-number approaches.[30] The six types of average-based consensus score are shown in Table 1. In this study, autoscaled scores are used for the by-number strategies. Optimizations of the thresholds, $x_{threshold}$, $y_{threshold}$, and $z_{threshold}$, were also carried out.

By contrast, the rank-by-vote approach uses a majority-vote-based consensus score. According to this strategy, if the score value of a model meets the standard that is set for a vote, the model is awarded one vote. This procedure is repeated for all of the scoring functions that are included in the consensus score, and the models that have many votes are eventually selected. Some previous studies[17−20] employed the rule that "models whose scores are within the top $z_{threshold}$ percent win one vote", and the rule that "models whose scores are less than, or equal to, $x_{threshold}$ win one vote" was used in CScore.[21] In addition to these criteria, the idea that the "top $y_{threshold}$ models obtain one vote" could also be applied. No previous comparative studies have examined these three standards, and they have all been referred to using the same term—rank-by-vote. In the current study, we proposed different names for these three approaches: the majority-vote-based consensus scoring strategy that awarded votes to "models whose scores are within the top $z_{threshold}$ percent" was referred to as vote-by-percent, the strategy that awarded votes to "models whose scores are less than, or equal to, $x_{threshold}$" was referred to as vote-by-number, and the consensus scoring strategy in which the "top $y_{threshold}$ models" were awarded votes was referred to as vote-by-rank. Thus, the rank-by-vote strategy described in ref 19 was referred to as the vote-by-percent strategy in the current study, and the CScore strategy was referred to as the vote-by-number strategy. In the present study, these three types of majority-vote-based consensus scores were investigated together with six types of average-based consensus score. In contrast to the average-based consensus scores, in which only one threshold is used, the vote-by strategies require two types of threshold: one for the voting standard and another for the number of votes. For example, in the vote-by-percent strategy, the model candidate with a score within the top $z_{threshold}$ percent obtains one vote for one scoring function (the $z_{threshold}$ is the threshold for the voting standard), and a candidate with a total number of votes for all of the scoring functions that is greater than, or equal to, the $w_{threshold}$ will eventually be selected (the $w_{threshold}$ is the threshold for the number of votes). Figure 3 illustrates the vote-by strategy. Both thresholds of the vote-by strategies were optimized in our study.

2.5. Evaluations of Consensus Scores. To evaluate the consensus scores, we initially determined the thresholds that enabled reasonable solutions to be obtained for all of the systems without exception. Using these thresholds, the number of model candidates could be reduced, and we investigated how many remained in each candidate group. For example, in the number-by strategies, the $x_{threshold}$ values were optimized in order to reduce the number of models as much as possible while at least one reasonable model remained in the filtered sets of candidates, the consensus score values of which were less than, or equal to, the $x_{threshold}$ for all 220 protein−ligand systems without exceptions. Similarly, the $y_{threshold}$ and $z_{threshold}$ values were optimized for the rank-by and percent-by strategies, respectively. These thresholds are illustrated in Figure 4. As shown in the figure, the model with the best score from all of the reasonable models was identified, and its score value, rank, and percentage were investigated. These values were obtained for all 220 protein−ligand systems, and the highest values

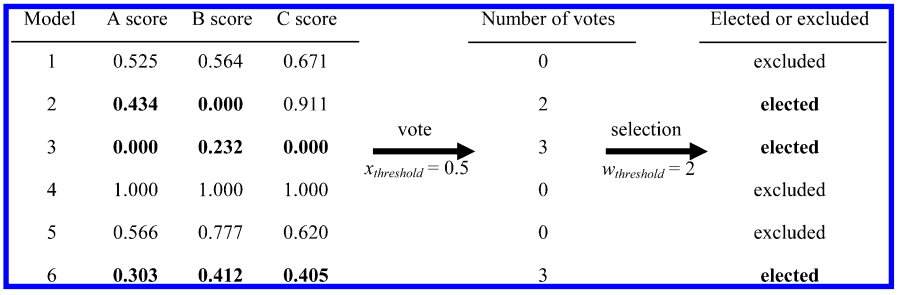

Figure 3. Vote-by strategies.

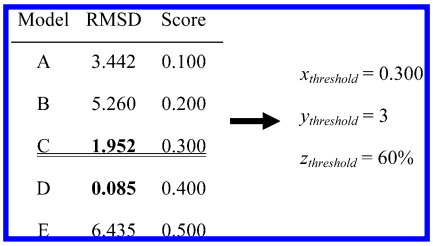

Figure 4. Thresholds. The solution with a RMSD ≤ 2.0 Å was selected as a focus, and its score value ($x_{threshold}$), rank order ($y_{threshold}$), and top % ($z_{threshold}$) were adopted as the thresholds.

among the 220 scores, ranks, and percentages were set as the $x_{threshold}$, $y_{threshold}$, and $z_{threshold}$ values, respectively. Using these thresholds, at least one reasonable model could be obtained for all 220 systems. In contrast to average-based consensus scores, not only $x_{threshold}$, $y_{threshold}$, $z_{threshold}$ but also $w_{threshold}$ need to be optimized for majority-vote-based strategies. When the thresholds for voting standards ($x_{threshold}$, $y_{threshold}$, or $z_{threshold}$) were defined, each model obtained a certain number of votes by vote-by strategies. For example, the situation in which model 1 wins three votes, model 2 wins one vote, and model 3 wins five votes and only models 1 and 2 are "reasonable models", is considered. For this situation, the appropriate $w_{threshold}$ is three, because when the $w_{threshold}$ is greater than three, no reasonable models are selected. When $w_{threshold}$ is less than three, it is possible to select reasonable models, but a higher threshold is more appropriate. In this way, an appropriate $w_{threshold}$ was investigated for each vote-by strategy and each $x_{threshold}$, $y_{threshold}$, and $z_{threshold}$. The ratio of the remaining models to all of the candidates was referred to as the compression ratio and was calculated using the following formula:

$$p_{compress} = \sum_{220\ systems} n_{remain} / \sum_{220\ systems} n_{all} \qquad (3)$$

Here, $p_{compress}$ is the compression ratio, n_{remain} is the number of remaining models using the threshold, and n_{all} is the total number of models. The compression ratio was used as an indicator of the ability of each consensus score strategy.

Using conditions under which reasonable models could be obtained for all 220 systems without exception, the compression ratio was generally relatively large (that is, not well-compressed). Therefore, the numbers of protein−ligand systems for which reasonable models could be selected using several predefined threshold values were also elucidated, to investigate the scenario in which reasonable models could be obtained, not for all 220 systems, but for the majority of the systems, with few exceptions. Using these predefined thresholds, a good compression ratio was expected at the expense of the accuracy of the modeling of a few of the protein−ligand systems. For this investigation, not only the compression ratio but also the ratio of accurate modeling (that is, the ratio of accurately modeled systems to the total number of systems) was evaluated using the following formula:

$$p_{accurate} = n_{accurate}/n_{all} \qquad (4)$$

Here, $p_{accurate}$ is the ratio of accurate modeling, $n_{accurate}$ is the number of protein−ligand systems in which reasonable models can be obtained using the predefined threshold, and n_{all} is the total number of systems (in this study, n_{all} = 220 for the complete test set, n_{all} = 57 for those complexes with high-affinity ligands only, and n_{all} = 122 for groups including complexes with $n \geq 250$). The tradeoff between accuracy and efficiency was investigated using the $p_{compress}$ and $p_{accurate}$ values for each consensus score.

In the current study, the combinations of scoring functions with the best $p_{accurate}$ and $p_{compress}$ values were investigated for each strategy and threshold. However, the computational cost of evaluating all 511 combinations would have been extremely high. Therefore, from a practical standpoint, the results produced using all nine functions could be discussed for the simple consensus scoring methods without exploring the combinations; thus, the computational cost that was involved in searching for the best combination was reduced. To further simplify the procedure, we also investigated the consensus scores using only five of the scoring functions: FlexX score, GOLD score, PMF score, DOCK score, and ChemScore. These functions were included in the CScore module of the SYBYL 6.9 program. Because all of these scores could be calculated simultaneously in a single CScore trial, using the consensus scores with these five functions was the simplest method for our study. Not only exhaustive investigations of 511 combinations but also tests of these simplified consensus scores were carried out in this study.

Consensus scores for the experimental structures of protein−ligand complexes were also calculated in order to consider the wider applications of the consensus scoring strategies. As the experimental structures were regarded as the "correct" answers, the consensus scores for these structures were expected to reflect the ability of the scoring strategies in applications with real complexes. The number-by-number strategy was used for this purpose, because it was one of the most useful consensus scoring strategies and could easily be compared among different protein−ligand complex systems. The models that were calculated by FlexX were used as parent populations for the consensus scoring. The experimental structure was added to n models obtained by FlexX, so the parent population for the consensus scoring included $n + 1$ models.

The ranks and percentages of models might be biased by differences of the numbers of model candidates between test complexes. For example, although we considered that the "top three models of six candidates" means the same as the "top three models of 500 candidates" in rank-by strategies, the latter three models appeared to be more highly selected than the former models. Because the flexibilities of ligands and the sizes of active sites were very different from each other for 220 test systems, it is not a practical setting that FlexX generates completely the same number of candidates for all systems. Thus, we extracted the 122 complex systems that have more than or equal to 250 candidates, and we compared the abilities of consensus scoring strategies for these 122 systems. By using this test set, the bias caused by differences of the numbers of candidates was reduced, and the dependencies of abilities of consensus scores on n were investigated.

3. RESULTS AND DISCUSSION

3.1. Selection of Reasonable Models for the 220 Protein−Ligand Systems. To determine the thresholds under which at least one reasonable model could be selected

Table 2. Comparison of the Nine Types of Consensus Score

(a) Number-by, Rank-by, and Percent-by Strategies

	threshold	$p_{compress}$ of best combination
number-by-number	$x_{threshold} = 0.452$	0.399
number-by-rank	$x_{threshold} = 151$	0.519
rank-by-number	$y_{threshold} = 151$	0.519
rank-by-rank	$y_{threshold} = 151$	0.519
percent-by-number	$z_{threshold} = 62.0\%$	0.617
percent-by-rank	$z_{threshold} = 62.4\%$	0.623

(b) Vote-by Strategies

	threshold for vote or not	threshold for number of votes ($w_{threshold}$)/number of functions included in best combination	$p_{compress}$ of best combination
vote-by-number	$x_{threshold} = 0.5$	5 votes/7 voters	0.371
vote-by-rank	$y_{threshold} = 150$	2 votes/3 voters	0.510
vote-by-percent	$z_{threshold} = 70\%$	3 votes/3 voters	0.502

for all 220 protein−ligand systems without exception, nine consensus scoring strategies and 511 combinations of scoring functions (that is, a total of 4599 types of consensus scores) were examined. The compression ratios that were produced using these thresholds were also investigated, that is, the compression ratios under the condition $p_{accurate} = 1.0$. Although all 511 combinations for each strategy were systematically evaluated, the combination that gave the best (smallest) compression ratio is focused on here (this was defined as the "best combination"). The best combinations are summarized in Table 2, in which both the thresholds and the compression ratios are described. For example, in the case of the number-by-number strategy, the best combination included the FlexX score, GOLD score, PMF score, DOCK score, PLP, and ScreenScore (as shown in Table S2 of the Supporting Information, and see below for discussion). After the autoscaling of the six scores, the average value—that is, the number-by-number consensus score—was calculated for each model. The $x_{threshold}$ of this combination for the number-by-number strategy was 0.452, which meant that at least one reasonable model was produced for each of the 220 protein−ligand systems by selecting the models with number-by-number consensus scores that were less than, or equal to, 0.452 (Table 2a). Using this strategy and threshold value, the number of model candidates to be explored was expected to be reduced to about 40% of the current size, because the compression ratio was equal to 0.399 (Table 2a). The vote-by-number strategy gave the best compression ratio for selecting suitable models for all 220 systems (Table 2a). In the best combination of vote-by-number, when the auto scaled value of one score was less than, or equal to, 0.5 for one model, it obtained one vote because $x_{threshold} = 0.5$ (Table 2b). When one model received more than, or equal to, five votes from the seven voters, it was selected because $w_{threshold} = 5$. Although the vote-by-percent strategy has been discussed for use in virtual screening trials in some previous reports,[17−20] our result suggests that a vote-by-number strategy, such as CScore, is a more appropriate majority-vote-based consensus score system for model selections. The second-best was the number-by-number strategy, which gave the best compression ratio of all average-based consensus scores.

The combinations that gave the top 10 compression ratios for all nine consensus scoring strategies are shown in Table S2 of the Supporting Information. As shown in Table S2a and g, for number-by-number and vote-by-number strategies, which had compression ratios that were superior to those of the other approaches, FlexX score, PLP, and ScreenScore were included in many of the top 10 consensus scores. It has been reported previously that these scoring functions work well when they are used independently.[20,22] Our results suggest that they are appropriate not only for independent scoring but also for consensus scoring. However, although it was reported that the GOLD score and DOCK score implemented in the CScore module were less successful when used alone, they appear high up in the lists of successful consensus scoring methods shown in Table S2a and g (Supporting Information). These force-field-based scoring functions are based on a different concept from that used to develop the empirical and knowledge-based scoring functions, and they seem to compensate for the shortcomings of the empirical scoring functions (such as FlexX score, PLP, and ScreenScore). By contrast, ChemScore and X-Score appeared in relatively few of the top 10 combinations. These are both empirical scoring functions, which are similar in form to FlexX score and ScreenScore, so mutual complementarity between them might not be expected. Therefore, ChemScore and X-Score do not seem to be appropriate for use together with FlexX score or ScreenScore, although they might perform well independently.

3.2. Efficient Selection of Reasonable Models with Some Exceptions. The investigations of the abilities of consensus scores discussed in the previous section were carried out under conditions in which reasonable models were selected for all protein−ligand test sets without exception (that is, $p_{accurate} = 1.0$). In this scenario, the values of the compression ratios tend to become overly large, because the thresholds are set to a high value if there are only a few systems for which reasonable models are difficult to search; this impairs the effectiveness of the consensus scores, even if they work well for most protein−ligand systems. In fact, the compression ratio was around 40%, even for the vote-by-number strategy, which was the best approach for this purpose. Therefore, the tradeoff between the $p_{accurate}$ and $p_{compress}$ values was investigated using several threshold values in order to make effective selections of reasonable models for as many protein−ligand systems as possible. In the current section, we discuss number-by, rank-by, and percent-by strategies, in which only one threshold is used. In the following section, consensus scores that are easy to use without exploring the combinations of scoring functions are described, and the tradeoff between the $p_{accurate}$ and $p_{compress}$ values is also discussed for vote-by strategies, in which two types of thresholds (that is, thresholds for the voting standard and for the number of votes) are required.

Figure S1 of the Supporting Information shows the ratios of the accurate modeling ($p_{accurate}$) for the number-by, rank-by, and percent-by strategies using several threshold values. In this section, the combination with the best $p_{accurate}$ among all of the 511 combinations is defined as the "best combination". The best combination for each threshold is summarized in Figure S1. Note that the meaning of best combination in this section is different from that in the previous section (where the term referred to the best compression ratio, rather than the best ratio of accurate modeling). The best combination of scoring functions for each threshold is shown in Table S3 of the Supporting Information together with the $p_{accurate}$

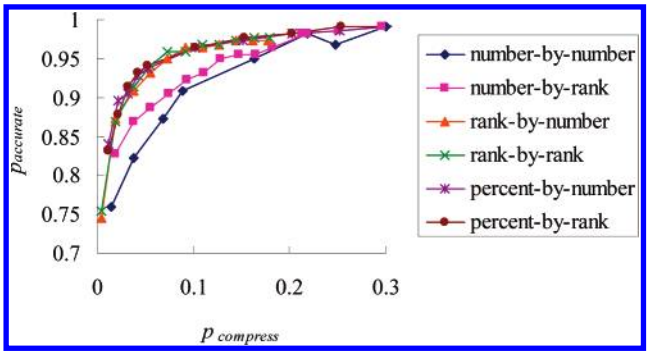

Figure 5. Compression ratios versus the ratios of accurate modeling when several threshold values were used.

and $p_{compress}$ values. In the rank-by and percent-by strategies, the behaviors of the by-number and by-rank consensus scores depending on the thresholds were similar to one another (Figures S1c and d). This indicates that, for these strategies, the results do not depend on whether by-number or by-rank approaches are used when the same thresholds are adopted.

Although Figure S1 (Supporting Information) illustrates the dependency of the $p_{accurate}$ values on thresholds, the tradeoff between the $p_{accurate}$ and $p_{compress}$ values cannot be discussed on the basis of this figure. Thus, the relationships between these parameters are explored in Figure 5, on the basis of the $p_{accurate}$ and $p_{compress}$ values presented in Table S3 (Supporting Information). Figure 5 shows that the $p_{compress}$ values of both of the number-by strategies were much worse than those of the rank-by and percent-by strategies when they were compared at the same values of $p_{accurate}$. This suggests that the rank-by and percent-by strategies are more appropriate for the effective selection of model candidates than the number-by strategies, which differs from the result we obtained for model selection with $p_{accurate} = 1.0$ (in which the number-by-number strategy was more appropriate). Therefore, different strategies should be adopted for different purposes, for example, "$p_{accurate} = 1.0$ is indispensable" or "a good balance between $p_{accurate}$ and $p_{compress}$ is desired".

For the number-by-number and rank-by-number strategies, the top 10 combinations of scoring functions in terms of $p_{accurate}$ are shown in Table 3. The $x_{threshold}$ and $y_{threshold}$ values were as small as possible while maintaining $p_{accurate} \geq 0.9$ for the best combinations (the top 10 combinations for the other strategies are shown in Table S4 in the Supporting Information). Table 3 demonstrates that while all of the top 10 combinations (excluding the number 1 combination) included more than, or equal to, four scoring functions in the rank-by-number strategy, all of the top 10 combinations in the number-by-number strategy included less than, or equal to, three functions; in particular, four of the top 10 combinations included only one scoring function and, thus, were not consensus scoring approaches. This suggests that the concept of consensus scoring does not work well for the number-by-number strategy, in contrast to the rank-by-number strategy, for the purpose of highly effective model selection at the expense of the accurate modeling of a few exceptional systems.

For the fast selection of model candidates, only the top model is frequently investigated. This is the situation with $y_{threshold} = 1$ for rank-by strategies. In Figure 6, the $p_{accurate}$ values of rank-by-rank and rank-by-number with $y_{threshold} = 1$ are illustrated. The results of independent scoring by nine

Table 3. Combinations of Scores That Gave the Top 10 Ratios of Accurate Modeling When $x_{threshold}$ and $y_{threshold}$ Were Small

(a) Number-by-Number ($x_{threshold} = 0.2$)

	scoring functions										
	F	G	PM	DO	C	Dr	PL	S	X	$p_{accurate}$	$p_{compress}$
1						√	√	√		0.9091	0.0882
2							√	√		0.9091	0.0957
3								√		0.9091	0.1179
4							√			0.9045	0.1183
5						√				0.9045	0.1186
6						√	√			0.9000	0.1050
7						√		√		0.8955	0.0909
8							√	√	√	0.8818	0.0706
9	√						√	√		0.8818	0.0856
10									√	0.8818	0.1099

(b) Rank-by-Number ($y_{threshold} = 10$)

	scoring functions										
	F	G	PM	DO	C	Dr	PL	S	X	$p_{accurate}$	$p_{compress}$
1					√		√			0.9091	0.0372
2	√	√		√	√	√	√			0.9045	0.0371
2	√	√		√		√		√		0.9045	0.0371
2					√	√	√	√		0.9045	0.0371
2	√	√		√	√	√	√	√		0.9045	0.0371
2	√	√			√	√	√			0.9045	0.0371
2	√	√		√		√	√	√		0.9045	0.0371
2		√			√	√	√		√	0.9045	0.0371
2		√			√	√	√	√	√	0.9045	0.0371
10					√		√	√	√	0.9045	0.0372

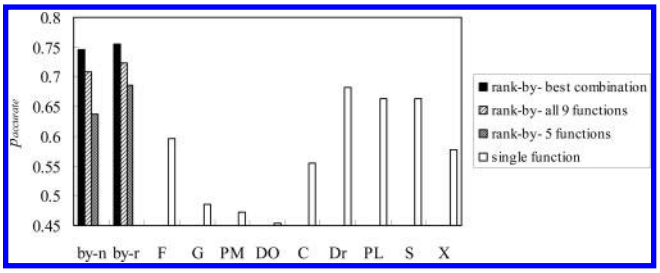

Figure 6. Ratios of accurate modeling for rank-by-number, rank-by-rank, and nine independent scoring functions when $y_{threshold} = 1$ was used. The "by-n" and "by-r" mean rank-by-number and rank-by-rank, respectively.

scores are also shown in the figure. For rank-by-number and rank-by-rank, not only the results of the best combination but also those of combinations including all nine functions and five CScore functions, that is, FlexX score, GOLD score, PMF score, DOCK score, and ChemScore, are illustrated. As shown in this figure, $p_{accurate}$ is around 0.75 for two types of rank-by strategies, and they were superior to all nine independent scoring functions. The $p_{accurate}$ values of the best combinations of two rank-by strategies were similar to one another, and this result was consistent with those of $y_{threshod} > 1$. However, for the combinations including all nine functions and five CScore functions, the results of rank-by-rank were better than those of rank-by-number. In fact, for rank-by-number, although the best combination and the combination including all nine functions were superior to all independent scoring functions, the result of the combination including five CScore functions was worse than the results of the independent scoring of DrugScore, PLP, and ScreenScore. On the other hand, for rank-by-rank, $p_{accurate}$ values obtained by not only best combination and the combination including all nine functions but also the combination including five CScore functions were better than

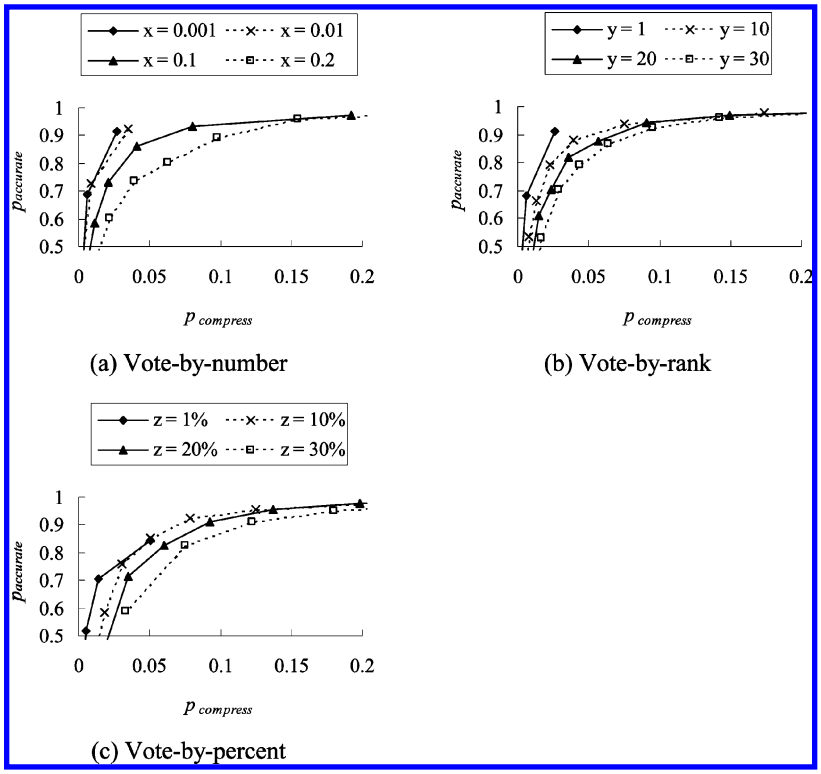

Figure 7. Results obtained using all nine functions depending on the thresholds.

those of all independent scoring functions. These results suggest that rank-by-rank is more robust in terms of combination of scoring functions than rank-by-number when $y_{threshold} = 1$.

3.3. Easy-to-Use Consensus Scores without Exploring the Combinations of Scoring Functions. For practical applications of consensus scores, exploring the best combination of scoring functions is highly expensive in terms of computational cost. Although thorough investigations of the combinations are desirable to allow detailed discussions, the computational cost is frequently as important as accuracy in drug design trials. Thus, in this section, consensus scores that are accurate and easy to use without the need to exhaustively explore the best combinations are discussed. One of these systems includes all nine functions and another includes the five CScore functions. In particular, the latter can be calculated using only a single CScore trial and is, therefore, the most simple consensus score considered in this study.

A comparison of the $p_{compress}$ values of the nine consensus scoring strategies under the condition $p_{accurate} = 1.0$ was carried out. The results are shown in Figure S2 of the Supporting Information, and they were similar to those of the best combinations mentioned in Section 3.1. These results support the finding that the two strategies, that is, vote-by-number and number-by-number, are appropriate for model selection of all systems without exception. Furthermore, for these two strategies, the $p_{compress}$ values calculated by both the combinations including all nine functions and five CScore functions were not much worse than those of the best combination. It suggests that these strategies also have advantages in terms of robustness of the combination of scoring functions.

The $p_{accurate}$ values obtained by the number-by, rank-by, and percent-by strategies depending on the thresholds are shown in Figure S3 of the Supporting Information (this is similar to the investigations mentioned in Section 3.2). The $p_{accurate}$ values of the rank-by and percent-by strategies using all nine functions or the five CScore functions were not much worse than those of the best combinations, in contrast to the number-by strategies. This means that, for the rank-by and percent-by strategies, the combinations using all nine functions or the five CScore functions, without exploring the best combinations, were effective in saving computational costs. Although the $p_{accurate}$ values of the rank-by-number and rank-by-rank strategies were similar to each other by using the best combinations, the rank-by-number strategy gave better $p_{accurate}$ values than the rank-by-rank strategy using all nine functions and the five CScore functions when $y_{threshold}$ was between 5 and 20. These results were different from those of $y_{threshold} = 1$, mentioned in Section 3.2, in which rank-by-rank was more robust than rank-by-number in terms of combinations of scoring functions.

In addition to the number-by, rank-by, and percent-by consensus scoring strategies, the dependencies of the $p_{accurate}$ and $p_{compress}$ values on the $w_{threshold}$ (the thresholds for the number of votes) were discussed for the vote-by strategies with the combinations including all nine functions. These discussions are useful for model selection by vote-by strategies with good balances between the $p_{accurate}$ and $p_{compress}$ values.

Figure 7 compares the balances between the $p_{accurate}$ and $p_{compress}$ values depending on the thresholds for the vote-by consensus scores including all nine functions, to determine which of the threshold values for the voting standard ($x_{threshold}$, $y_{threshold}$, and $z_{threshold}$) are desirable. For each of the thresholds for the voting standard, the threshold values for the number of votes ($w_{threshold}$) were set from 1 to 9. Figure 7 illustrates the results depending on both of the two thresholds (the complete data are shown in Table S5 in the Supporting

Information). As shown in the figure, the tradeoff between the $p_{accurate}$ and $p_{compress}$ values obtained by the vote-by-percent strategy was the worst of all the vote-by approaches. For the vote-by-number and vote-by-rank strategies, smaller $x_{threshold}$ and $y_{threshold}$ values gave smaller (better) compression ratios when similar $p_{accurate}$ values were obtained. For example, if $p_{accurate} \geq 0.9$ was required, the best $p_{compress}$ for the vote-by-rank was around 2.6% and was obtained using $y_{threshold} = 1$ and $w_{threshold} = 1$. These thresholds mean that each scoring function votes for only the top model, and the models that win one or more votes are selected. The same results can be obtained using the vote-by-number strategy when the $x_{threshold}$ is small enough, because the autoscaled score value of the top model is always 0. The compression ratio produced by this threshold was the best of all the consensus scores, including not only the vote-by but also the number-by, rank-by, and percent-by strategies for the condition of $p_{accurate} \geq 0.9$. By contrast, it is difficult to adjust the parameters for vote-by strategies because of the requirement for two thresholds; thus, other strategies, such as the rank-by-number approach, might be more suitable for easy-to-use consensus scoring.

3.4. Consensus Scores for Complexes with High Binding Affinities. Sections 3.1−3.3 discussed our investigations of all 220 protein−ligand complex systems. The current section describes the analysis of the 57 complexes with high affinities that are marked in bold in Table S1 of the Supporting Information.

Figure S4 in the Supporting Information shows the compression ratios produced by the nine consensus scoring strategies under the condition that gave reasonable models for all 57 systems without exception (that is, $p_{accurate} = 1.0$). The results that were obtained using the best combinations, all nine functions, and five CScore functions are illustrated. The number-by-number and vote-by-number strategies were appropriate for the 57 complexes, which was consistent with the results obtained for all 220 test complexes. These findings suggest that the results for the 220 complexes will also be useful for designing high-affinity ligands, which could play important roles in drug design trials.

In addition to the calculations under the condition $p_{accurate} = 1.0$, the tradeoffs between the $p_{accurate}$ and $p_{compress}$ values were investigated for the 57 complexes with high-affinity ligands using several threshold values. The results are shown in Figure 8. The balances between the $p_{accurate}$ and $p_{compress}$ values obtained using the best combinations for number-by, rank-by, and percent-by strategies are shown in Figure 8a (similar to those presented in Figure 5). Figure 8b, which corresponds to Figure 7b, illustrates the tradeoffs between the $p_{accurate}$ and $p_{compress}$ values for the vote-by-rank strategy using all nine functions. According to these calculations, the results of the rank-by and percent-by strategies were better than those of the number-by strategies. For the vote-by-rank, $y_{threshold} = 1$ was the best voting standard threshold. These findings were consistent with the earlier discussions of all 220 complexes.

3.5. Consensus Scores for Experimental Complex Structures. In this section, we discuss the abilities of consensus scores for use with experimentally observed complex structures. Calculating the consensus score values for various experimental structures (that is, the "correct answer") revealed how small scores could be obtained for

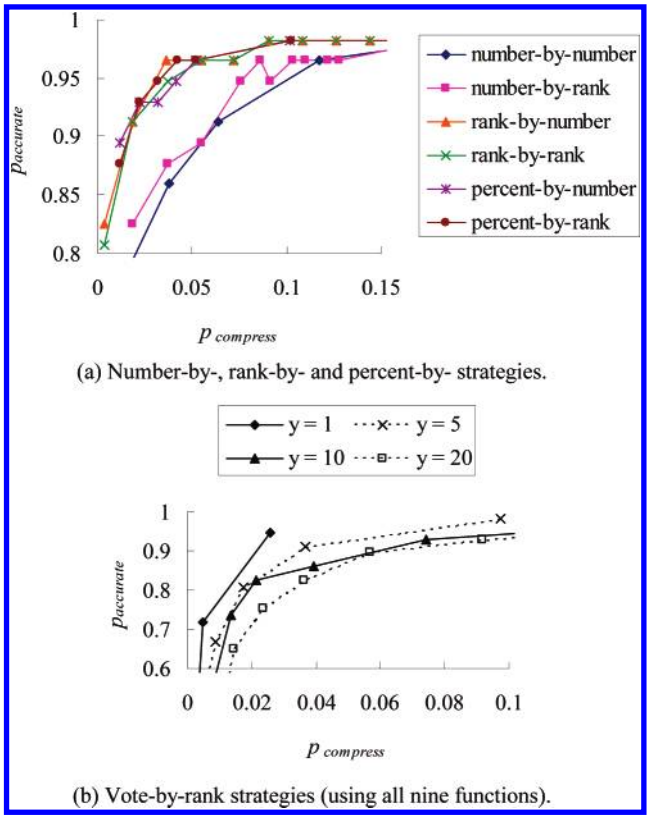

Figure 8. Compression ratios versus the ratios of accurate modeling for the test set including complexes with high-affinity ligands.

Table 4. Thresholds of Number-by-Number Consensus Scores for the Experimental Complex Structures

	$x_{threshold}$	
	all complexes	complexes with high-affinity ligands
best combination	0.386	0.271
using all nine functions	0.689	0.513
using the five CScore functions	0.785	0.676

compounds that experimentally docked into target proteins. In this study, the number-by-number strategy was used for this purpose.

Table 4 and Figure 9 present the computational results for the experimental structures produced using the number-by-number strategy. The thresholds under which reasonable models can be selected (that is, the worst value for the number-by-number strategy) are described for all 220 experimental structures and for the 57 complexes with high-affinity ligands. The score value for the test set including complexes with high-affinity ligands was better than that for the whole test set. This suggests that, when the ligand with a better score is selected, a higher binding affinity can be expected using the number-by-number strategy. In addition, the frequency distributions of the consensus score values of the experimental structures are illustrated in Figure 9. For practical reasons, we used the combinations including all nine scoring functions and the five CScore functions. For the high-affinity ligands, the number-by-number values were less than 0.3 for 90% of the 57 experimental structures produced using all nine functions.

3.6. Consensus Scores for Complexes with More Than or Equal to 250 Candidates Generated by Computational

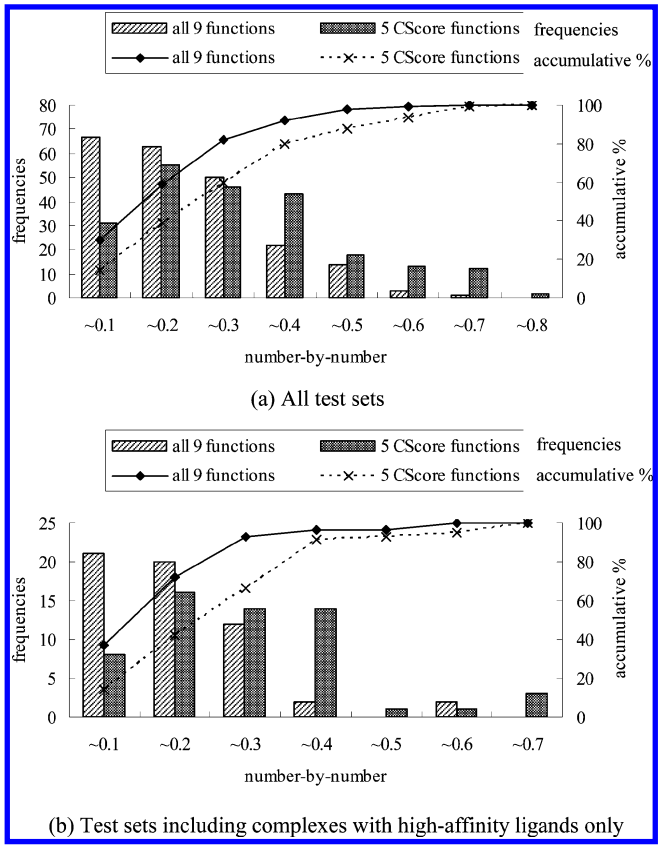

Figure 9. Frequency distributions and accumulative percentages of number-by-number values for the experimental complex structures.

Docking. In computational docking, the numbers of generated candidates depend on flexibilities of ligands and sizes of active sites of target proteins. As shown in Table S1 of the Supporting Information, in this study, the number of candidates was between 6 and 500. To find the consensus scoring strategies which can be widely used, various types of test complexes, for which various numbers of candidates were generated, need to be investigated as mentioned in Sections 3.1−3.5. However, the large differences of the numbers of candidates possibly cause bias for the ranks and percentages. In this section, the comparisons of consensus scores for only 122 test complexes which have more than or equal to 250 model candidates generated by FlexX were carried out in order to reduce the bias.

Figure 10 shows a comparison of the $p_{compress}$ values of the nine consensus scoring strategies under the condition $p_{accurate} = 1.0$. As shown in this figure, the results of the best combinations of percent-by strategies were as good as or better than those of vote-by-number and number-by-number for complexes with $n \geq 250$ although both percent-by strategies gave much worse $p_{compress}$ values than vote-by-number and number-by-number in a test for all 220 complexes, as mentioned in Section 3.1. On the other hand, for the combinations including all nine functions and five CScore functions, results of the percent-by strategies were worse than those of vote-by-number and number-by-number. It indicates that vote-by-number and number-by-number are more appropriate in terms of the robustness of combinations of scoring functions not only for complete test sets but also for complexes with $n \geq 250$. For rank-by-number and vote-by-rank, although $p_{compress}$ values were highly improved in comparison with the results shown in Table 2 in Section 3.1, they remain worse than those of vote-by-number and number-by-number. These results suggest that although ranks and percentages are affected by the number of model candidates (n), vote-by-number and number-by-number are still appropriate under the condition $p_{accurate} = 1.0$ because they workwell regardless of whether n is large or not and they are robust in terms of combinations of scoring functions.

In Figure 11, the tradeoffs between $p_{accurate}$ and $p_{compress}$

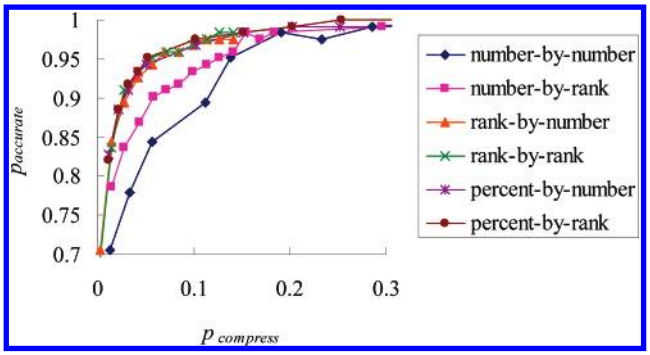

Figure 11. Compression ratios versus the ratios of accurate modeling for the test set including complexes with $n \geq 250$.

for complexes with $n \geq 250$ are illustrated. As shown in this figure, the results were similar to those in Figure 5 in Section 3.2. Thus, to achieve a good balance between $p_{accurate}$ and $p_{compress}$, rank-by and percent-by strategies were appropriate regardless of the number of candidates. Both the results of studies with conditions under $p_{accurate} = 1.0$ and those of tradeoff studies for complexes with $n \geq 250$ were consistent with those for all 220 complexes, and they indicate that the results of this study are useful for various systems regardless of the numbers of generated candidates within the limitation of "Num.Answers" = 500.

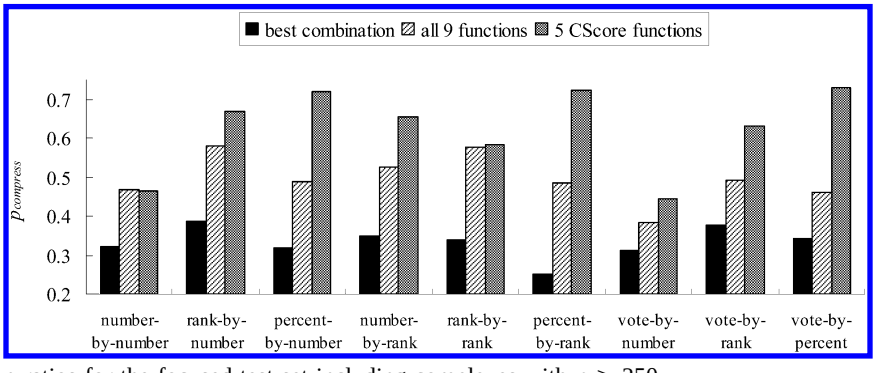

Figure 10. Compression ratios for the focused test set including complexes with $n \geq 250$.

4. CONCLUSIONS

In this study, we investigated the abilities of consensus scores to evaluate docking models constructed using FlexX. We systematically named the nine types of consensus scoring strategies that have been independently proposed (and previously confused with one another) and compared their performance. All 511 types of combinations including the nine scoring functions were investigated for each of the strategies. Consequently, we found that the number-by-number and vote-by-number strategies were appropriate for use in model selection in all of the systems without exception, and the rank-by-number and percent-by-number strategies were useful for model selection with a good tradeoff between accuracy and efficiency. Considering the scoring functions that were utilized, PLP and DrugScore, which were effective for model selection in single scoring systems, were also appropriate for use in consensus scores. In addition, GOLD score and DOCK score, which were not effective in single scoring systems, were also useful in consensus scoring approaches, as they seemed to compensate for the shortcomings of the other scoring functions. Optimizing the combinations of scoring functions is expensive in terms of computational costs, so consensus scores including all nine functions or the five CScore functions without the need for prior optimizations are particularly useful in practice. Although the vote-by strategies were effective for model selection, they require two types of threshold to be defined, and it is difficult to control the numbers of finally selected models. Thus, we recommend the number-by-number strategy (for all systems without exception) or the rank-by-number strategy (for a good balance between accuracy and efficiency), both of which have abilities similar to those of the vote-by strategies.

In previous papers,[17−20,31,32] the vote-by-percent strategy has been used as a representative of vote-by approaches (denoted as rank-by-vote in refs 17 and 20, and the "intersection approach" in ref 31) for compound selection in virtual screening trials. Some of these studies reported that vote-by strategies did not work well in comparison to other approaches or single scoring.[17,19,31] However, as shown in the current study, the vote-by-percent approach is the least appropriate of all the vote-by strategies, at least for model selection, and other techniques should be used for discussions of the abilities of vote-by strategies. Although the scoring scheme of compound selection in virtual screening is different from that of model selection in computational docking, the vote-by-percent strategy might not be appropriate for virtual screening, as is the case in model selections. In our study, the abilities of consensus scores were discussed only for use in model selection in computational docking, and we intend to systematically investigate the performance of consensus scores in virtual screening in future trials.

In the current study, model selection for all systems without exception, and for most systems with few exceptions, were investigated. This is the first comparison of the use of consensus scores in these two situations, and we found that different types of strategies are required. If vote-by strategies are inadequate because of practical problems, number-by-number and rank-by-number approaches should be suitable for the former and latter situations, respectively. We previously described the by-number-type consensus score AASS[30] and argued that both the number-by-AASS and rank-by-AASS approaches should be used according to the demands of the specific situation.

Although this study focused on the selection of docking models produced by FlexX, we expect our results to be relevant to other computational docking programs. We intend to investigate these aspects further in a future study.

Supporting Information Available: Lists of the protein−ligand complexes for the test set, top 10 combinations of the scores, dependencies of p_{accurate} and p_{compress} on $w_{\mathrm{threshold}}$, the figures of dependencies of p_{accurate} on threshold, and the figures of p_{compress} under the condition $p_{\mathrm{accurate}} = 1.0$. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Kroemer, R. T. Molecular Modelling Probes: Docking and Scoring. *Biochem. Soc. Trans*. **2003**, *31*, 980−984.
(2) Leach, A. R. In *Molecular Modelling*, 2nd ed.; Pearson Education Limited: Essex, 2001; Chapter 12, pp 640−726.
(3) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications, *Nat. Rev. Drug Discovery* **2004**, *3*, 935−949.
(4) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.
(5) Jones, G.; Willett, 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) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A Geometric Approach to Macromolecule−Ligand Interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.
(7) 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.
(8) 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.
(9) Eldridge, M. D.; Murray, C. W.; Auton, T. 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.
(10) Murray, C. W.; Auton, T. R.; Eldridge, M. D. Empirical Scoring Functions. II. The Testing of an Empirical Scoring Function for the Prediction of Ligand−Receptor Binding Affinities and the Use of Bayesian Regression to Improve the Quality of the Model. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 503−519.
(11) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based Scoring Function to Predict Protein−Ligand Interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(12) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Molecular Recognition of the Inhibitor AG-1343 by HIV-1 Protease: Conformationally Flexible Docking by Evolutionary Programming. *Chem. Biol.* **1995**, *2*, 317−324.
(13) Wang, R.; Lai, L.; Wang, S. Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11−26.
(14) Stahl, M.; Schulz-Gasch, T. Practical Database Screening with Docking Tools. *Ernst Schering Res. Found. Workshop* **2003**, *42*, 127−151.
(15) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy. *Proteins: Struct., Funct., Bioinf.* **2004**, *57*, 225−242.
(16) Willis, R. C. 2001: A Dock Odyssey. *Mod. Drug Discovery* **2001**, *4*, 26−28.
(17) Verdonk, M. L.; Berdini, V.; Harshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual Screening Using Protein−Ligand Docking: Avoiding Artificial Enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793−806.
(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) Wang, R.; Wang, S. How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422−1426.

(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) Marsden, P. M.; Puvanendrampillai, D.; Mitchell, J. B. O.; Glen, R. C. Predicting Protein−Ligand Binding Affinities: A Low Scoring Game? *Org. Biomol. Chem.* **2004**, *2*, 3267−3273.
(24) Rarey, M.; Kramer, B.; Lengauer, T. Docking of Hydrophobic Ligand with Interaction-Based Matching Algorithms. *Bioinformatics* **1999**, *15*, 243−250.
(25) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FlexX Incremental Construction Algorithm for Protein−Ligand Docking. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 228−241.
(26) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(27) *SYBYL 6.9*; Tripos Inc.: St. Louis, MO, 2002.
(28) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parametrization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(29) *X-Score 1.1*; Department of Internal Medicine, University of Michigan Medical School: Ann Arbor, MI, 2003.
(30) Katsuki, M.; Chuang, V. T. G.; Nishi, K.; Kawahara, K.; Nakayama, H.; Yamaotsu, N.; Hirono, S.; Otagiri, M. Use of Photoaffinity Labeling and Site Directed Mutagenesis for Identification of Key Residue Responsible for Extraordinarily High Affinity Binding of UCN-01 in Human Alpha 1-Acid Glycoprotein. *J. Biol. Chem.* **2005**, *280*, 1384−1391.
(31) 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.
(32) Xiang, L.; Hodgkin, E.; Liu, Q.; Sedlock, D. Evaluation and Application of Multiple Scoring Functions for a Virtual Screening Experiment. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 333−344.

CI050283K