Bootstrap-Based Consensus Scoring Method for Protein–Ligand Docking

Hiroaki Fukunishi,*,† Reiji Teramoto,‡ Toshikazu Takada,§ and Jiro Shimada†

Nano Electronics Research Laboratories and Bio-IT Center, Central Research Laboratories, NEC Corporation, 34, Miyukigaoka, Tsukuba, Ibaraki 305-8501, Japan, and Riken, Next-Generation Supercomputer R&D Center, sixth Fl., Meiji Seimei Kan, 2-1-1 Marunouchi, Chiyoda-ku, Tokyo 100-0005

Received June 11, 2007

To improve the performance of a single scoring function used in a protein–ligand docking program, we developed a bootstrap-based consensus scoring (BBCS) method, which is based on ensemble learning. BBCS combines multiple scorings, each of which has the same function form but different energy-parameter sets. These multiple energy-parameter sets are generated in two steps: (1) generation of training sets by a bootstrap method and (2) optimization of energy-parameter set by a *Z*-score approach, which is based on energy landscape theory as used in protein folding, against each training set. In this study, we applied BBCS to the FlexX scoring function. Using given 50 complexes, we generated 100 training sets and obtained 100 optimized energy-parameter sets. These parameter sets were tested against 48 complexes different from the training sets. BBCS was shown to be an improvement over single scoring when using a parameter set optimized by the same *Z*-score approach. Comparing BBCS with the original FlexX scoring function, we found that (1) the success rate of recognizing the crystal structure at the top relative to decoys increased from 33.3% to 52.1% and that (2) the rank of the crystal structure improved for 54.2% of the complexes and worsened for none. We also found that BBCS performed better than conventional consensus scoring (CS).

INTRODUCTION

A protein–ligand docking program has been used to efficiently discover lead compounds for a target protein from a huge compound database. Since the pioneering work by Kuntz et al.,[1] numerous docking programs have been developed.[2–15] There are two standpoints for assessing docking programs: conformation search or scoring. For the former, it is assessed whether the pose or binding mode of a crystal structure can be reproduced. For the latter, it is assessed whether the predicted score is correlated with experimentally measured binding affinity. Many reports assessing the performance of a docking program have been published.[16–29] Of particular interest is the recent critical assessment reported by Warren et al.[21] These reports generally reached a conclusion that docking algorithms are highly successful at generating good binding modes, while scoring functions are less successful at correctly identifying a binding mode. More specifically, some of the docking programs can generate a conformation very close to the crystal binding structure, but one cannot know which conformation created by which program is close to the true structure. Under such situation, it is desirable to further improve scoring method to rerank conformations generated. To supplement the weakness of a single scoring function, a consensus scoring (CS) method combining multiple scoring functions was proposed.[22–30] Although CS typically does not perform better than the best of the scoring functions combined, it has the advantage of generally providing stable results. Since which scoring function is the best can only be known after the analysis of results, not before, CS is a reliable method for a blind trial. It has been demonstrated[22–30] that hit rate or false positive is improved by CS.

In this study, we developed a bootstrap-based consensus scoring (BBCS) method. BBCS combines multiple scorings, each of which has the same function form but different energy-parameter set. BBCS is derived from the idea of bootstrap aggregating (bagging),[31] which is one kind of ensemble learning. In general, it is not easy to uniquely determine the energy-parameter set that provides the global minimum of binding free energy over all complexes when the number of data in a training set is not enough. To circumvent this problem, BBCS based on ensemble learning is a rational method.

Multiple energy-parameter sets are determined in two steps: (1) generation of training sets by the bootstrap method and (2) optimization of an energy-parameter set by a *Z*-score approach against each training set. The *Z*-score approach is based on energy landscape theory, which leads to describes a funnel-shaped landscape of protein folding[32] and has been applied to folding simulation, structure prediction[33–38] and de novo design[39] of proteins and protein-peptide docking.[40] We consider that protein–ligand docking also has a funnel-shaped landscape, as discussed by Camacho and Vajda,[41] and the *Z*-score approach works well for protein–ligand docking.

Machine leaning has been applied in several ways to overcome difficulties in structure-based consensus scoring.[27,42,43] For example, Jacobsson et al. used the ensemble method (bagging) to create rule-based prediction models based on the scoring matrices.[27] Another example is in a recent paper of Antes et al., who applied an ensemble method to neural networks and *k*-nearest neighbor models.[42] Al-

* Corresponding author. E-mail address: h-fukunishi@bu.jp.nec.com. Phone: +81 298 856 6155. Fax: +81 298 856 6136.
† Nano Electronics Research Laboratories, NEC Corporation.
‡ Bio-IT Center, NEC Corporation.
§ Riken, Next-Generation Supercomputer R&D Center.

10.1021/ci700204v CCC: $40.75
Published on Web 04/22/2008

Table 1. List of Protein–Ligand Complexes (PDB Codes) Classified According to the Dominant Interactions between Protein and Ligand

type	training data	test data
type 1: electrostatic	1adb, 1af2, 1bap, 1e96, 1fmo, 1hsl, 1rgk, 1rgl, 1zzz, 2csc, 2gbp, 2qwc, 2qwe, 2qwf, 3fx2, 6abp, 6rnt, 6tim, 7abp, 8xia, 9aat, 9abp (TR1) (22)	1abe, 1abf, 1add, 1apb, 1rnt, 1tet, 1yyy, 2ak3, 2qwb, 2qwd, 2sns, 2xim, 3cpa, 3ptb, 4tim, 4xia, 5abp, 5cna, 5p21, 7tim, 8abp (TE1) (21)
type 2: hydrophobic	1bcu, 1d3p, 1exw, 1hvr, 1inc, 1tmn, 1tni, 1tnj, 1tnk, 1tnl, 7est, 7tln (TR2) (12)	1bbz, 1cla, 1d3d, 1ela, 1ets, 1fkb, 1fkf, 1rbp, 2cgr, 3cla, 4cla (TE2) (11)
type 3: mixed	1a46, 1apt, 1bxo, 1drf, 1etr, 1mnc, 1ppc, 1pph, 1tng, 1tnh, 2ctc, 2qwg, 2tmn, 3tmn, 4sga, 5tln (TR3) (16)	1a5g, 1apw, 1b5g, 1ba8, 1bb0, 1bhf, 1bra, 1bzm, 1cbx, 1dhf, 1dr1, 1sre, 1tlp, 2pk4, 4tln, 5sga (TE3) (16)
type 4: all	TR1 + TR2 + TR3 (50)	TE1 + TE2 + TE3 (48)

though the aim of theirs and our work is to find a more plausible binding mode among numerous modes, there are marked differences. They used an ensemble method to optimize a parameter set efficiently, while we use it to improve the rank of the crystal structure relative to decoy structures. In other words, they determined a single parameter set used for a scoring function, not multiple parameter sets as ours does. Their parameter set is optimized so that the binding mode with the lowest rmsd becomes close to the crystal structure, while our parameter set is optimized so that crystal structure is maximally separated from decoys with respect to score.

METHODS

Preparation of training and test sets. In this study, 98 protein–ligand complexes were used. These complexes are classified into three types[26] according to the dominant interactions between the protein and ligand: electrostatic interaction for type 1, hydrophobic for type 2, and mixed electrostatic and hydrophobic for type 3. The complexes in each type were randomly divided into "training set" and "test set," as listed in Table 1. For each complex, the crystal structure and 100 docked structures (decoys) generated by Wang et al.[26] were used. Their docked structures are carefully generated in conformational space of ligand as completely as possible so that their distribution is not biased by the docking algorithm employed. Such structures allow various scoring functions to be compared on the same ground. Wang et al. used AutoDock,[7] because it can generate a wide variety of conformations during long multiple stochastic searches on potential energy surface. Wang's data set is useful for testing the performance of scoring independently of the performance of docking algorithm. Our intention of using Wang's data set is to proceed to refining scoring functions beyond comparing their performances. To show an example of clear advantage of Wang's data set over other data sets such as the data set generated by FlexX program with default protocol, we compared the distributions of conformations in the two data sets, in Supporting Information S-1. There is seen a remarkable difference for some of the complexes; FlexX fails to generate conformations sufficiently thoroughly in a low range of rmsd for the complex shown. If such insufficient data set of conformations is used for test of scoring, the performance of scoring should be coupled with the performance of docking algorithm. On the other hand, Wang's data set is more complete, containing conformations searched in wide range. Therefore, we consider that Wang's data set is useful as benchmark set for scoring.

Wang's data set is published online at http://sw16.im.med.umich.edu/software/xtool. In the published structures, bound water molecules are eliminated but metal atoms are included. In this study, metal atoms are also eliminated.

Scoring Function. In BBCS, any scoring function proposed can be employed. We employ the FlexX scoring function, which is given as[24]

$$\Delta G_{\mathrm{bind}} = \Delta G_{\mathrm{match}} F_{\mathrm{match}} + \Delta G_{\mathrm{lipo}} F_{\mathrm{lipo}} + \Delta G_{\mathrm{ambig}} F_{\mathrm{ambig}} + \Delta G_{\mathrm{clash}} F_{\mathrm{clash}} + \Delta G_{\mathrm{rot}} n_{\mathrm{rot}} + \Delta G_0 \quad (1)$$

Here, F_i are functions of the protein and ligand coordinates. Term F_{match} is the sum of the energy contributions from each hydrogen bond, metal contact, and specific aromatic interaction, where each contribution has been multiplied by two linear penalty functions for angle and distance deviations from predefined ideal values. Terms F_{lipo} and F_{ambig} provide a measure of hydrophobic contact as functions of protein–ligand atom pairs: F_{lipo} involving only pairs of nonpolar atoms, and F_{ambig} involving pairs of one polar and one nonpolar atom. Term F_{clash} is a penalty function for protein–ligand overlap, and n_{rot} is equal to the number of rotatable bonds in the ligand. ΔG_i is the energy-parameter set (coefficients). Most parameters were derived from Böhm type scoring function.[44]

For convenience in optimizing energy-parameter, we rewrite eq 1 in the form

$$\Delta G_{\mathrm{bind}} = (\alpha_1 \Delta G_{\mathrm{match}}) F_{\mathrm{match}} + (\alpha_2 \Delta G_{\mathrm{lipo}}) F_{\mathrm{lipo}} + (\alpha_3 \Delta G_{\mathrm{ambig}}) F_{\mathrm{ambig}} + (\alpha_4 \Delta G_{\mathrm{clash}}) F_{\mathrm{clash}} + (\alpha_5 \Delta G_{\mathrm{rot}}) n_{\mathrm{rot}} + \Delta G_0 \quad (2)$$

where α_i represents an adjustment factor of energy-parameter. In practice, $p = \{\alpha_i\}$ is optimized (see below).

Determining Multiple Energy-Parameter Sets. Multiple energy-parameter sets are determined in two steps, as shown in Figure 1.

Step 1: Generation of multiple training sets.

Each observation in our sampling is taken as one crystal structure plus N_{decoy} decoy structures: $C = \{ r^{\mathrm{crystal}}, r_1^{\mathrm{decoy}}, r_2^{\mathrm{decoy}}, r_3^{\mathrm{decoy}}, \ldots, r_{N_{\mathrm{decoy}}}^{\mathrm{decoy}} \}$, for a given protein–ligand system (complex), where r represents the Cartesian coordinates of the structure. The size of the original sample is the number N_{complex} of complexes. Then, from the original sample, a bootstrap sample $T^{(B)} = \{C^{(1)}, C^{(2)}, C^{(3)}, \ldots, C^{(N_s)}\}$ of size N_s is generated by randomly sampling with replacement, with the superscript on C indicating the sampled complex. This means that some observation C in the original sample may be included several times in the bootstrap sample $T^{(B)}$, i.e., $C^{(i)} = C^{(j)}$ is allowed for $i \neq j$. By repeating the above procedure N_{train} times, multiple training sets denoted by T_{Mul}

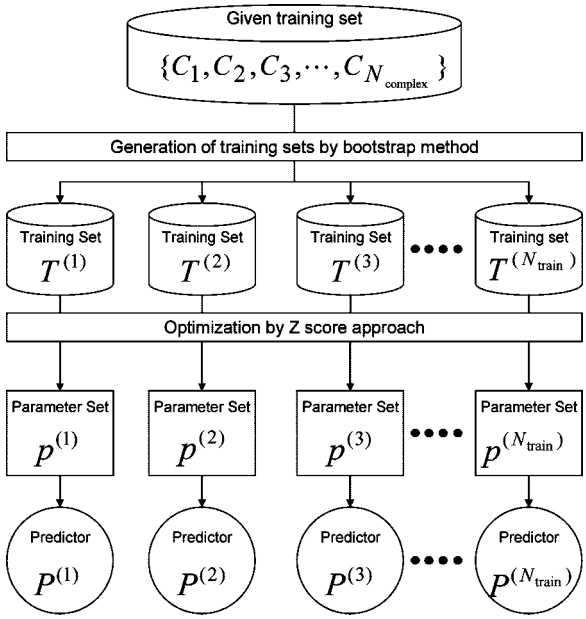

Figure 1. Process for determining multiple parameter sets.

(a)

		Predictor						
		Pred_1	Pred_2	Pred_3	·	·	·	Pred_M
Docking structure	DS_1	1	1	45	·	·	·	21
	DS_2	3	5	19	·	·	·	10
	DS_3	8	17	2	·	·	·	6
	·	·	·	·	·	·	·	·
	·	·	·	·	·	·	·	·
	·	·	·	·	·	·	·	·
	DS_N	45	65	9	·	·	·	33

(b)

		Rank						
		1	2	3	·	·	·	N
Docking structure	DS_1	0.25	0.19	0.12	·	·	·	0.00
	DS_2	0.18	0.30	0.16	·	·	·	0.02
	DS_3	0.00	0.05	0.03	·	·	·	0.14
	·	·	·	·	·	·	·	·
	·	·	·	·	·	·	·	·
	·	·	·	·	·	·	·	·
	DS_N	0.00	0.00	0.01	·	·	·	0.20

Figure 2. Illustration of determining the weight factor of rank-by-wcs. Pred and DS represent the predictor and the docking structure, respectively. Table a of rank is converted into Table b of the population $P_R = n_R/N_{pred}$, where n_R is the number of the predictors that rank the structure as Rth, and N_{pred} is the total number of the predictors.

$= \{T^{(1)}, T^{(2)}, T^{(3)}, \ldots, T^{(N_{train})}\}$ are generated. In this study, N_{decoy}, N_s, and N_{train} are all 100. In various applications of bagging, N_{train} of order 10–10^3 is empirically found to be suitable.[31]

Step 2: Optimization of the parameter set.

We want to determine the parameter set $p^{(B)}$ optimized for the training set $T^{(B)}$. For a set of structures $C = \{r^{crystal}, r_1^{decoy}, r_2^{decoy}, r_3^{decoy}, \ldots, r_{N_{decoy}}^{decoy}\}$ of a given complex, the Z-score is defined as a function of parameter set p as follows,

$$Z(p) = \frac{S(r^{crystal}, p) - \langle S(r^{decoy}, p)\rangle_D}{\sigma_D} \quad (3)$$

where

$$\sigma_D = [\langle S(r,p)^2\rangle_D - \langle S(r,p)\rangle_D{}^2]^{1/2} \quad (4)$$

Here, S and σ_D represent the score and its standard deviation for decoys, respectively, and $\langle\rangle_D$ indicates an ensemble average over decoys. When a p that provides a larger negative $Z(p)$ for the complex is determined, the crystal structure can be clearly discriminated from the decoys. In practice, an optimal parameter set applicable to all complexes in the training set is required. Therefore, a $p^{(B)}$ providing a large negative $\langle Z(p)\rangle_C$ is determined, where $\langle\rangle_C$ represents an average over all complexes C in $T^{(B)}$:

$$p^{(B)} = \arg\max_p [-\langle Z(p)\rangle_C] \quad (5)$$

To explore $p^{(B)}$ numerically, adjustment factors α_i in eq 2 are set at 0.1 intervals between –3.0 and 3.0 and the Z-scores are calculated for 61^5 combinations. Having determined $p^{(B)}$ using $T^{(B)}$, we obtain multiple parameter sets:

$$p_{Mul} = \{p^{(1)}, p^{(2)}, p^{(3)}, \cdots, p^{(N_{train})}\} \quad (6)$$

The prediction made using $p^{(B)}$ yields the predictor $P^{(B)}$, which constitutes multiple predictors P_{Mul}:

$$P_{Mul} = \{P^{(1)}, P^{(2)}, P^{(3)}, \cdots, P^{(N_{train})}\} \quad (7)$$

Hereafter, N_{train} is written as the number of predictors N_{pred}, because $N_{train} = N_{pred}$ in our scheme.

Combining Multiple Predictors (Consensus Ranking Method). Multiple predictors can be combined in various ways to make a final prediction. Here, we consider two methods, rank-by-wcs (weighted consensus score) and rank-by-number. Multiple predictors under consideration are $p_{Mul} = \{p^{(1)}, p^{(2)}, p^{(3)}, \ldots, p^{(N_{pred})}\}$ for BBCS and $P_{cs} = \{$FlexX, DOCK, GOLD, PMF, ChemScore$\}$ for CS.

(i) Rank-by-wcs. We introduce a weighted consensus score as an extension of rank-by-rank.[30] First, the protein–ligand docking structures including crystal structure and decoys are ranked according to the ascending order of scores calculated by each predictor. A matrix of rank is then obtained as shown in Figure 2a. Next, for each structure, population $P_R = n_R/N_{pred}$ is calculated, where n_R is the number of the predictors that rank the structure as Rth among N_{pred} predictors. As a result, a matrix of P_R is obtained, as shown in Figure 2b. For example, DS_1 gets first, second, and third rank by 25%, 19%, and 12% of all predictors, respectively. A weighted consensus score, S_{WCS}, is defined by using P_R as a weight factor:

$$S_{WCS} = \sum_R^{N_{DS}} f(R) P_R \quad \text{(weighted consensus score)} \quad (8)$$

where N_{DS} is the total number of docking structures. In this study, we adopt $f(R)$ defined as

$$f(R) = N_{DS} - R \quad (9)$$

Finally, the consensus rank of each docking structure is obtained by sorting S_{WCS} in descending order. We refer to this rank as rank-by-wcs. It should be noted that consensus rank determined by eqs 8 and 9 is equivalent to rank-by-rank of Wang et al.[30]

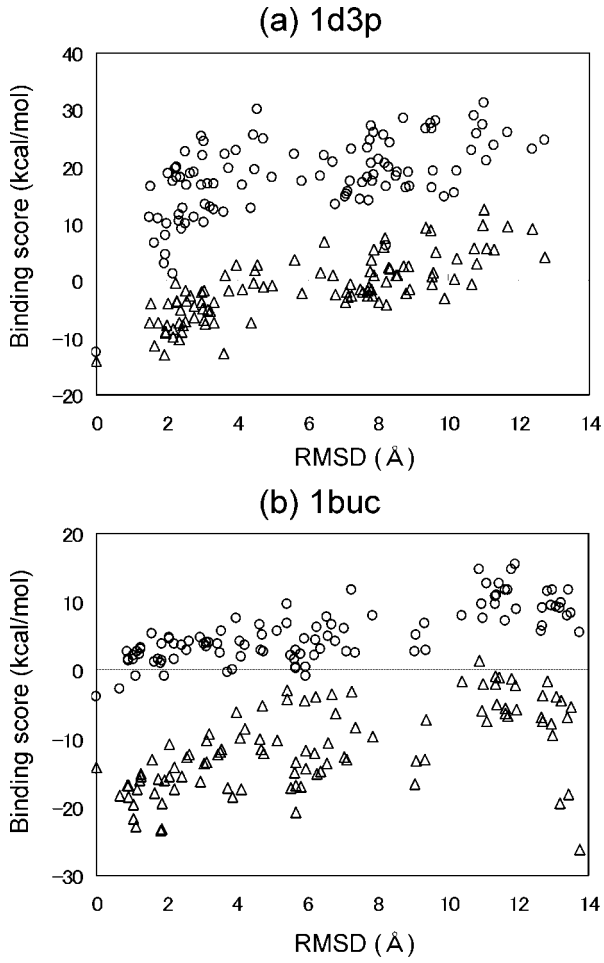

Figure 3. Typical plots of binding score vs rmsd value: (a) 1d3p and (b) 1buc. The triangles represent values obtained from eq 1 by using the original FlexX parameter set and the circles represent values obtained from eq 2 by using a typical optimized parameter set in p_{Mul}.

(ii) Rank-by-Number. As an alternative, consensus rank by rank-by-number[30] is also calculated. First, an average score is calculated for every docking structure as follows:

$$S_{\mathrm{ave}} = \frac{1}{N_{\mathrm{pred}}} \sum_{j}^{N_{\mathrm{pred}}} S^{(j)} \tag{10}$$

where $S^{(j)}$ is the score calculated by the jth predictor $P^{(j)}$ in P_{Mul} Consensus rank of each docking structure is then obtained by sorting S_{ave} in ascending order. This rank is referred to as rank-by-number. In our special case of using eq 2, S_{ave} is just equal to the score calculated with simple averaged parameter set $p_{\mathrm{ave}} = \{\alpha_1^{\mathrm{ave}}, \alpha_2^{\mathrm{ave}}, \alpha_3^{\mathrm{ave}}, \ldots\}$ with

$$\alpha^{\mathrm{ave}} = \frac{1}{N_{\mathrm{pred}}} \sum_{j}^{N_{\mathrm{pred}}} \alpha^{(j)} \tag{11}$$

This simple relation $S_{\mathrm{ave}} = S(p_{\mathrm{ave}})$, however, does not always hold. It holds only when scoring function is represented as a linear combination of various energy terms and their coefficients are optimized.

Why BBCS Works. A semi-complete scoring function can be constructed if all possible conformations of all complexes existing in nature are used for optimizing a parameter set. Note that it is *complete* only within a parameter space (e.g., adjustment factor $p = \{\alpha_i\}$) of a given scoring function. However, because it is impossible to use all complexes as a training set, some error between actually constructed scoring function and semi-complete one is inevitable. The error becomes large, especially when the number of data in the training set is small. In this case, BBCS, whose key idea is based on bootstrap aggregating (bagging) in ensemble learning, reduces the error of scores between the actual scoring function and the semi-complete one. On the basis of bagging theory,[31] we can explain why BBCS works, for example, in case of rank-by-number. Here, e_{A} is defined as the error of the combined predictor:

$$e_{\mathrm{A}} = (E - S_{\mathrm{ave}})^2 \tag{12}$$

where E and S_{ave} are the scores calculated by the semi-complete scoring function and by eq 10, respectively, for a given complex. And e is defined as the average of errors over multiple predictors $P_{\mathrm{Mul}} = \{P^{(1)}, P^{(2)}, P^{(3)}, \ldots, P^{(N_{\mathrm{pred}})}\}$, where each error is obtained individually from a single predictor $P^{(j)}$:

$$\begin{aligned} e &= \frac{1}{N_{\mathrm{pred}}} \sum_{j}^{N_{\mathrm{pred}}} (E - S^{(j)})^2 \\ &= \frac{1}{N_{\mathrm{pred}}} \sum_{j}^{N_{\mathrm{pred}}} \left[(E - S_{\mathrm{ave}}) + (S_{\mathrm{ave}} - S^{(j)})\right]^2 \\ &= \frac{1}{N_{\mathrm{pred}}} \sum_{j}^{N_{\mathrm{pred}}} \left[(E - S_{\mathrm{ave}})^2 + (S_{\mathrm{ave}} - S^{(j)})^2\right] + \\ &\quad \frac{2}{N_{\mathrm{pred}}} (E - S_{\mathrm{ave}}) \sum_{j}^{N_{\mathrm{pred}}} (S_{\mathrm{ave}} - S^{(j)}) \end{aligned} \tag{13}$$

where we recall that $S^{(j)}$ is the score calculated by the single predictor $P^{(j)}$. Because the last term of eq 13 vanishes,

$$e = e_{\mathrm{A}} + \frac{1}{N_{\mathrm{pred}}} \sum_{j}^{N_{\mathrm{pred}}} (S_{\mathrm{ave}} - S^{(j)})^2 \tag{14}$$

Because the second term in eq 14 is always positive, i.e., $e \geq e_{\mathrm{A}}$, the combined predictor is statistically more reliable than the single predictor. It should be noted that although some single predictors in P_{Mul} may provide less error than the combined predictor, one cannot know them before blind trial.

RESULTS AND DISCUSSION

Validation of *Z*-score Approach (Optimization Method). Figure 3 shows the correlation between binding scores and rmsd values of the heavy atom positions of a ligand against the crystal structure as a reference, where scores are obtained with the original FlexX parameter set[45] and a typical optimized parameter set in p_{Mul}. For the plot of 1d3p (Figure 3a), the original parameter set succeeds in ranking the crystal structure at the top but leaves only a marginal score gap between the crystal structure and decoys. In contrast, the optimized parameter set makes the score gap quite clear. For the plot of 1buc (Figure 3b), the optimized parameter set makes the crystal structure top-ranked (low score), but the original parameter set does not. These results are examples of how the *Z*-score approach successfully discriminates crystal structure from decoys. The plots with the optimized parameter set appear to describe a funnel shaped energy landscape.

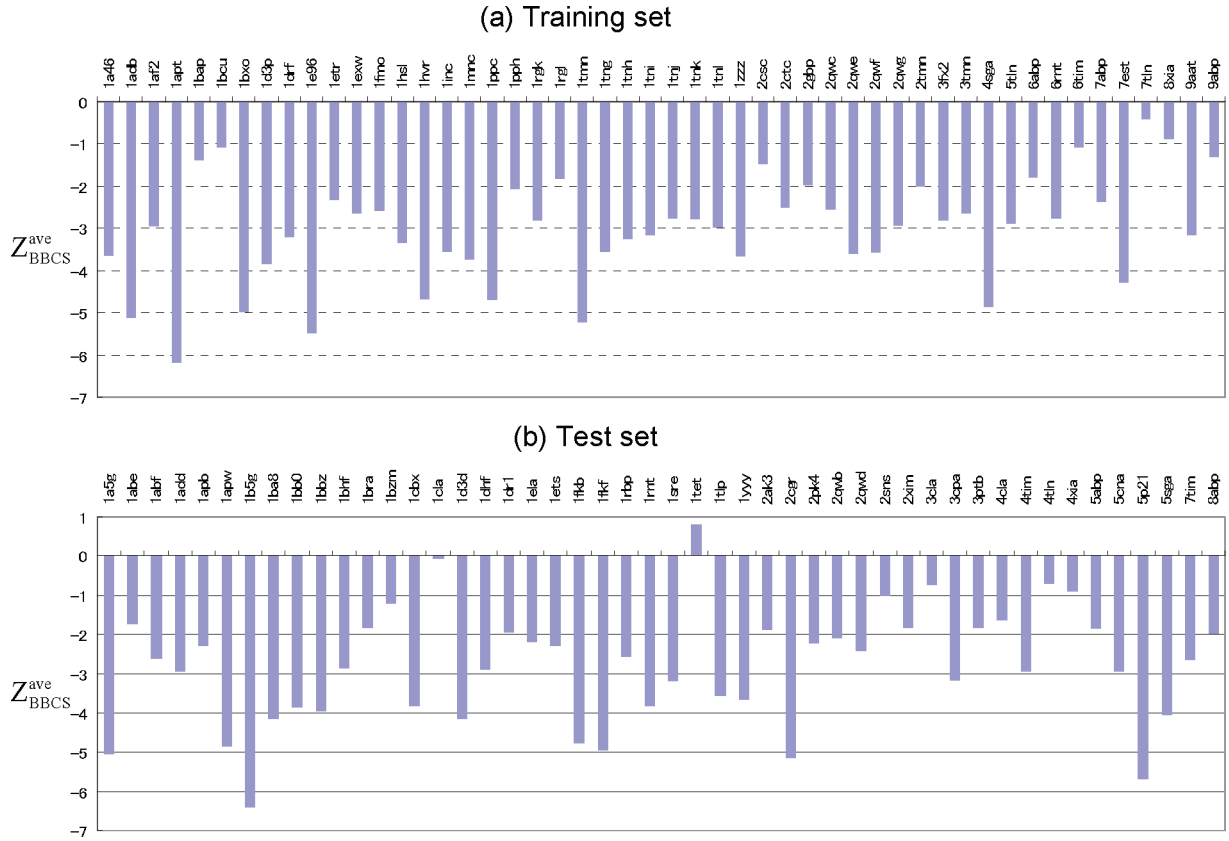

Figure 4. Average of *Z*-score, Z^{ave}_{BBCS}, for each complex in the training and test sets. The average is taken over all parameter sets contained in multiple parameter sets p_{Mul}.

Table 2. Distribution of $Z^{ave}_{BBCS} - Z_{FlexX}$

	%			
$Z^{ave}_{BBCS} - Z_{FlexX}$	Training set		Test set	
-1.5∼-1.0	2.00		10.42	
-1.0∼-0.5	6.00	76.00	12.50	68.75
-0.5∼-0.0	68.00		45.83	
0.0∼0.5	22.00		20.83	
0.5∼1.0	0.00	24.00	10.42	31.25
1.0∼1.5	2.00		0	

Table 3. Success Rate of Recognizing the Crystal Structure (Percentage of Complexes Whose Crystal Structure or Low rmsd Structures Are Top-Ranked)

	%					
	BBCS		CS			
	RbyW[a]	RbyN[b]	RbyW[a]	RbyN[b]	single scoring[c]	original FlexX
crystal	52.1	39.6	18.8	20.8	33.3	33.3
rmsd < 1 Å	72.9	60.4	45.8	41.7	52.1	54.1
rmsd < 2 Å	83.3	75.0	72.9	70.8	72.9	75.0

[a] Rank-by-wcs. [b] Rank-by-number. [c] Parameter set optimized by Z-score approach.

Figure 4 shows the average *Z*-scores calculated by

$$Z^{ave}_{BBCS} = \frac{1}{N_{pred}} \sum_{j=1}^{N_{pred}} Z(p^{(j)}) \qquad (15)$$

for each complex in the training and test sets. For the training set (Figure 4a), all complexes give negative values, which indicate the crystal structure was discriminated from the decoys. The larger the negative value is, the clearer the discrimination becomes. This is also the case with the test set (Figure 4b) except for 1tet. Thus we expect that p_{Mul} determined in this manner will be effective for unknown data.

Table 4. Comparison of BBCS and Various Scoring with Respect to the Rank *R* of crystal structure[a]

	%		
	>0	=0	<0
	D_{S-B}[b]		
rank-by-wcs	58.3	41.7	0
rank-by-number	39.6	60.4	0
	D_{F-B}[c]		
rank-by-wcs	54.2	45.8	0
rank-by-number	35.4	62.5	2.1
	D_{C-B}[d]		
rank-by-wcs	62.5	20.8	16.7
rank-by-number	56.3	20.8	22.9

[a] Percentage of complexes giving positive, zero, and negative D_{X-B} values are shown for all 48 complexes in the test set. D_{X-B} ($= R_X - R_{BBCS}$) is the rank difference. [b] Rank difference between single scoring with *Z*-score approach and BBCS. [c] Rank difference between original FlexX and BBCS. [d] Rank difference between CS and BBCS.

It is interesting to compare Z^{ave}_{BBCS} with Z_{FlexX} calculated with the original parameter set[45] of the FlexX scoring function. Table 2 lists the distribution of $Z^{ave}_{BBCS} - Z_{FlexX}$. Negative values of $Z^{ave}_{BBCS} - Z_{FlexX}$ mean that p_{Mul} provides a larger score gap than the original parameter set does, i.e., higher performance in discriminating the crystal structure from decoys. High percentages for the training set (76%) and test set (68.75%) were obtained.

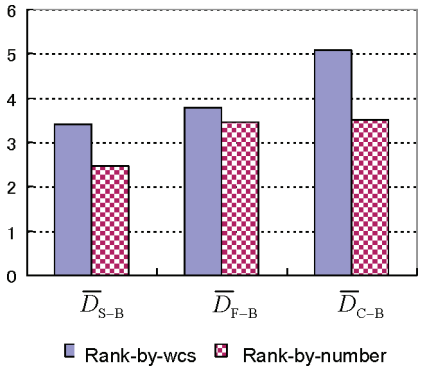

Figure 5. Average of D_{X-B} values, $\overline{D}_{X-B}$, for values over 48 complexes. Note that the averages of R_{single}, R_{FlexX} and R_{CS} are 10.4, 10.8, and 11.5, respectively.

Table 5. Rank Difference Averaged over Complexes Whose R_X Falls in the Indicated Range

	average rank difference		
	rank-by-wcs		rank-by-number
R_{single}		$\bar{D}_{S-B}$[a]	
1−5	0.53		0.27
6−10	3.00		1.33
11−100	10.83		8.50
R_{FlexX}		$\bar{D}_{F-B}$[b]	
1−5	0.61		0.258
6−10	2.80		1.20
11−100	12.42		10.17
R_{CS}		$\bar{D}_{C-B}$[c]	
1−5	−0.25(0.33)[d]		−0.57(0.15)[d]
6−10	6.00		1.83
11−100	15.36		12.43

[a] Average of rank differences between single scoring with *Z*-score approach and BBCS. [b] Average of rank differences between original FlexX and BBCS. [c] Average of rank differences between CS and BBCS. [d] $\bar{D}_{C-B}$ value when 4xia is omitted.

Validation of BBCS. To determine whether multiple parameter sets used in BBCS are more effective than a single parameter set optimized using all complexes in the training set without bootstrapping, we compared the results of BBCS with those of one reparameterization of the FlexX scoring function, referred to as "single scoring". To make a fair comparison, the parameter set for single scoring was also optimized with the *Z*-score approach.

We compared their performances against 48 complexes in the test set in two ways. The first comparison was with respect to the success rate of recognizing the crystal structure, which was calculated as the percentage of complexes whose crystal structure or low rmsd structures are top-ranked relative to decoys. The results are summarized in Table 3, where results mentioned in the next subsections are also included. BBCS attains the highest success rate of 52%, compared with 18–33% of the others in recognizing crystal structure. It is clear that both BBCSs, especially BBCS by rank-by-wcs, perform better than single scoring. This conclusion remains valid, even if we relax the success criteria by including low rmsd structures with rmsd < 1.0 or 2.0 Å as hits.

The second comparison was of the rank R of the crystal structure relative to 100 decoys. Table 4 lists the percentage of complexes giving positive, zero, and negative D_{S-B}, where D_{S-B} is an indicator defined by

$$D_{S-B} = R_{single} - R_{BBCS} \quad (16)$$

with R_{single} and R_{BBCS} the ranks of crystal structure by single scoring and by BBCS, respectively. A positive D_{S-B} means that BBCS is superior to single scoring. Surprisingly, the percentage of negative D_{S-B} is zero, indicating that BBCS is better than single scoring without exception. From a probabilistic viewpoint, this is not amazing even if BBCS were worse than single scoring for a small number of complexes, because BBCS's advantage is solely due to the reduced statistical error as can be seen from eq 14. Figure 5 shows the average rank difference $\bar{D}_{S-B}$ over complexes. The improvement in rank attained by BBCS is about 2 to 3 per 100 decoys on average against single scoring's average rank of 10.4. As listed in Table 5, we examined $\bar{D}_{S-B}$ by varying the R_{single} range. For moderately ranked structures in the range $R_{single} = 6-10$, there is an appreciable improvement $\bar{D}_{S-B} = 3$ for rank-by-rank. Note that rank-by-wcs always gives better results than rank-by-number.

Comparison of BBCS and the Original FlexX Scoring Function. Now that we have established the advantage of BBCS over single scoring on parameter sets determined by the same *Z*-score approach, let us proceed to a comparison of BBCS with the original FlexX scoring function.[45] Table 3 shows that BBCS has a higher success rate for the recognizing crystal structure or low rmsd structures than the original FlexX, especially for rank-by-wcs. Figure 6 shows the D_{F-B} value of each complex in the test set, where

$$D_{F-B} = R_{FlexX} - R_{BBCS} \quad (17)$$

with R_{FlexX} is the rank of the crystal structure given by the original FlexX. Table 4 indicates that BBCS provides a better ranking than the original FlexX for high percentage of complexes. In particular, rank-by-wcs provides an improvement over original FlexX for 54.2% of the complexes. Figure 5 shows the average rank difference $\bar{D}_{F-B}$ over complexes. The rank attained by BBCS is about 3 to 4 higher on average than the average rank of 10.8 attained by the original FlexX. As listed in Table 5, for moderately ranked structures in the range $R_{FlexX} = 6-10$, there is an appreciable improvement $\bar{D}_{F-B} = 2.8$ for rank-by-wcs. Furthermore, we compared BBCS and original FlexX for various combinations of TR and TE, as listed in Table 1 on the basis of the complex's main interaction. In either case, BBCS provides a better ranking that the original FlexX. (See Supporting Information S-2).

Comparing single scoring with the *Z*-score approach and original FlexX, one sees from Tables 3 and 4 that they are quite similar in performance in spite of the difference in optimization method.

In this paper, we have restricted ourselves to rescore conformations generated by AutoDock. When one wants to achieve best performance BBCS for the conformations generated by FlexX docking program, it may be advantageous to newly train scoring function only against a large number of conformations created by FlexX program. In practice, generated conformations are biased to some extent, because conformation search and scoring function are coupled. Thus it is not efficient to train against a set including conformations particularly disfavored by FlexX docking program. However, its determination is beyond the scope of this paper.

Comparison of BBCS and CS. The scoring functions used for CS were FlexX and D-score, G-score, PMF, and ChemScore, the latter four of which are included in CScore of the Sybyl module (Tripos Inc.). The number of predictors giving

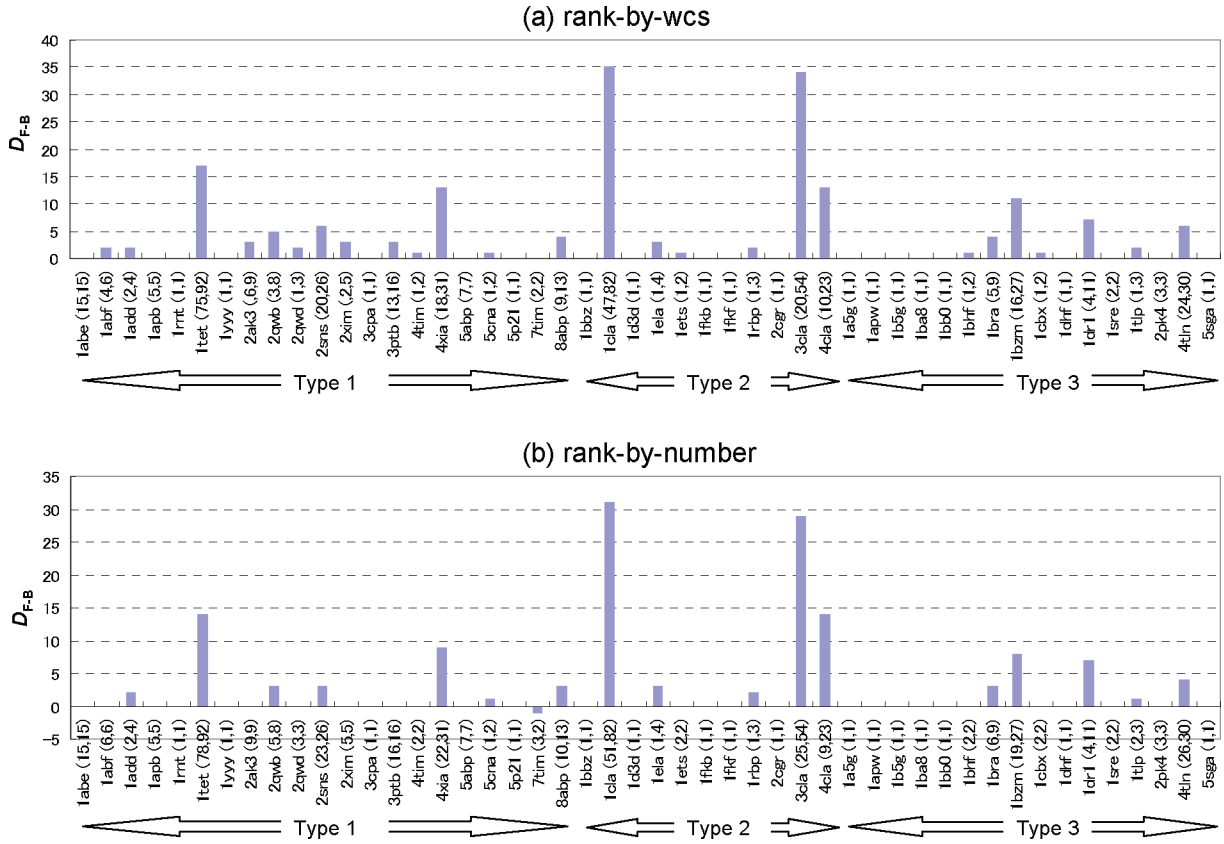

Figure 6. $D_{F-B} = R_{FlexX} - R_{BBCS}$ for each complex in the test set, where R_{FlexX} and R_{BBCS} are the rank of crystal structure by the original FlexX and by BBCS, respectively. The first and second values in parentheses are R_{BBCS} and R_{FlexX}, respectively. (a) Rank-by-wcs. (b) Rank-by-number.

the best or better performance for each method should be used when different methods are compared. Thus, the number of predictors is 100 for BBCS and 5 for CS in this study.

Table 3 shows that compared with CS, BBCS gives a higher success rate for recognizing crystal structure, especially for rank-by-wcs. Figure 7 shows D_{C-B} for each complex in the test set, where

$$D_{C-B} = R_{CS} - R_{BBCS} \tag{18}$$

with R_{CS} the consensus rank of the crystal structure for CS. Table 4 indicates that BBCS provides a better ranking by percentage, especially 62.5% of the complexes for rank-by-wcs. Figure 5 shows average rank difference $\bar{D}_{C-B}$ over complexes. The improvement in rank attained by BBCS is about 3 to 5 on average against the CS's average rank of 11.5. As listed in Table 5, for moderately ranked structures with $R_{CS} = 6-10$, there is an appreciable rank improvement $\bar{D}_{C-B}$=6.0 for rank-by-wcs. Although $\bar{D}_{C-B}$ for $R_{CS} = 1-5$ is slightly negative, it is probably not a serious problem, because it is mostly due to a failure at 4xia among 28 complexes. When 4xia is eliminated, $\bar{D}_{C-B}$ is positive (in parentheses).

CS gave the worst performance among all scorings in Tables 3 and 4; it was even worse than the predictor FlexX contained in CS. The difference between BBCS and CS lies in whether predictors are constructed based on a statistically rational method (bootstrapping) or constructed by simply gathering scoring functions. For CS, there is no guarantee that predictors will supplement each other's weak points if ones that give quite different results or almost the same results are gathered.

CPU time required is another important standpoint. Note that as long as docked conformations are generated by single procedure and these conformations are rescored by scoring functions, CPU time is not proportional to the number of predictors, i.e. CPU time of BBCS is not simply 20 times longer than that of CS in this study. Actually, there was little difference of CPU time between them. This is because, for BBCS, once values of F_i in eq 1, are calculated as a function of coordinates, scores for any parameter sets are simply obtained as $\Sigma c_i F_i$, whereas for CS, all scoring functions must be separately calculated because their function forms are different.

Worst Case by BBCS. There remains some complexes for which BBCS still fails to highly rank crystal structure among decoys; namely, the worst case is 1tet for which R_{BBCS} was 75th in 101 protein–ligand docking structures for rank-by-wcs. Hence, it is interesting to examine why BBCS failed for 1tet. As shown in Figure 8a, the ligand is a citrate. Figure 8b shows the difference between ligand binding sites of the crystal structure and of the top-ranked (incorrect) structure. For the crystal structure, the citrate binds to the receptor on the surface. To our knowledge, many docking programs regard such structure as energetically and structurally unstable. On the other hand, for the top-ranked structure, the citrate binds to the receptor in the pocket, which is mostly regarded as a more plausible site. A correct binding site can not be predicted. The reason for this is explained as follows. The binding mode of the crystal structure in Figure 8c shows two important histidine residues. It is likely that these imidazole rings are positively charged and make strong

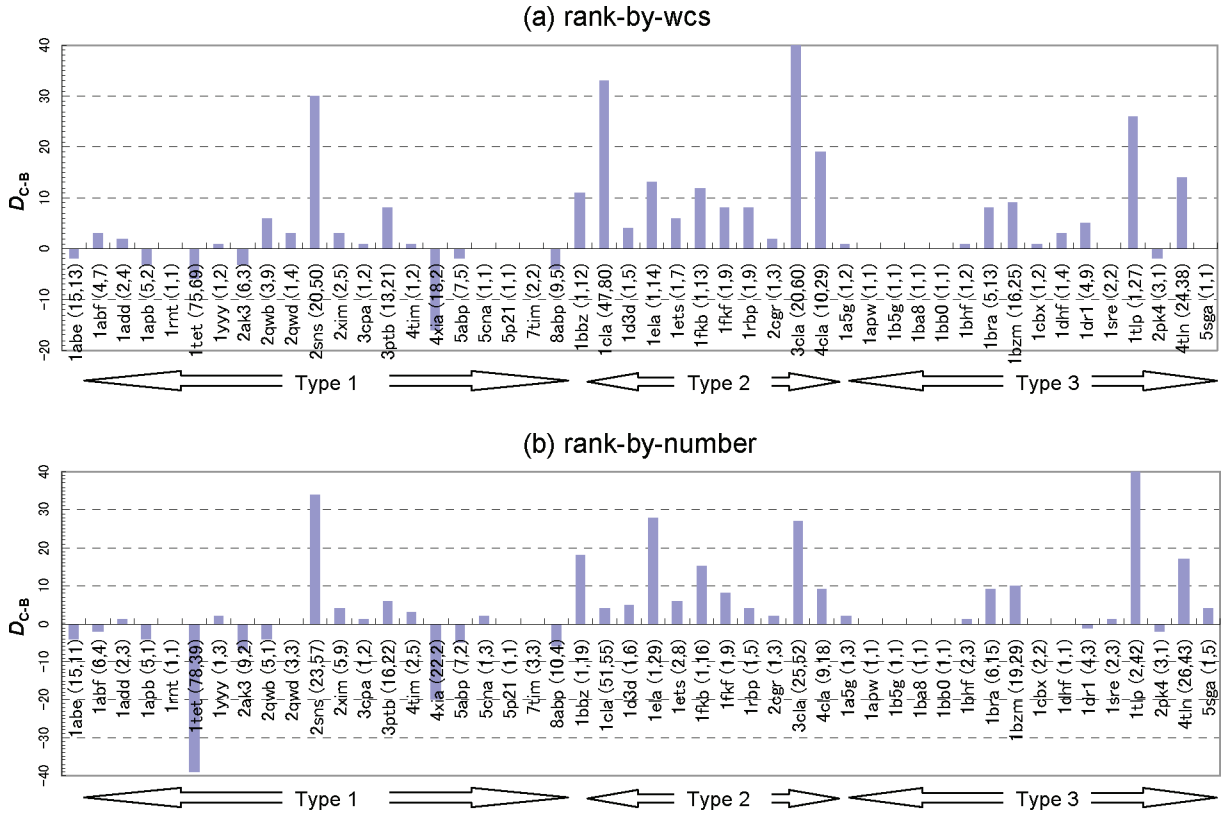

Figure 7. $D_{C-B} = R_{CS} - R_{BBCS}$ for each complex in test set, where R_{CS} and R_{BBCS} are the consensus rank of crystal structure by CS and by BBCS, respectively. The first and second values in parentheses are R_{BBCS} and R_{CS}, respectively. (a) Rank-by-wcs. (b) Rank-by-number.

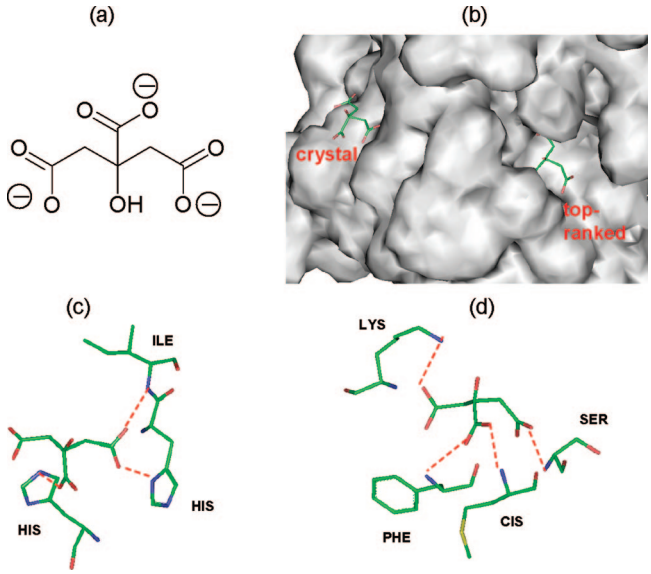

Figure 8. Structure of 1tet. (a) Ligand: citrate. (b) Binding site of crystal structure and the top-ranked (incorrect) structure obtained by BBCS. (c) Binding mode of crystal structure. (d) Binding mode of (incorrect) top-ranked structure.

ion–ion interactions with two carboxylate ions of citrate, as discussed by Shoham et al.[46] In our study, however, the histidine residues were treated as neutral. That is a main reason why the crystal structure was top-ranked. However, it must be remarked that one generally assumes pH = 7.0 for a blind trial, without considering delicate protonation state of histidines and that, in this case, histidines are neutral. On the condition that histidines are neutral, it is no wonder that the incorrect structure obtained top rank, because hydrogen bonds between ligand and four residues are plausibly formed for that structure. Thus, not only the performance of the scoring function but also adequate assignment of protonation state is very important.

CONCLUSION

BBCS was examined for pose prediction of protein–ligand docking. BBCS comes from two key ideas: the bagging to reduce the statistical error and the *Z*-score approach to distinguish the crystal structure from decoys. BBCS was shown to be an improvement over single scoring in a comparison of results obtained with the same *Z*-score approach. BBCS also performed better than original FlexX and CS in ranking the crystal structure. In particular, BBCS's success rate of recognizing the crystal structure at the top was higher. As for the strategies of combining multiple scoring functions, rank-by-wcs performed better than rank-by-number. Once the parameter sets optimized in BBCS are obtained, one can rescore placements generated by arbitrary docking algorithm in the following steps.

(1) Perform only scoring of their placements by FlexX module to obtain individual energy components.

(2) Calculate rescores of the placements by using the parameter sets optimized in BBCS.

(3) Perform consensus scoring based on the rescores obtained.

It would be interesting to apply BBCS to other scoring functions different in form from the FlexX function.

ACKNOWLEDGMENT

This work is supported in part by New Energy and Industrial Technology Development Organization, Japan

(NEDO) under the project of Development of Basic Technologies for Advanced Production Methods Using Microorganism Functions.

Supporting Information Available: Sections S1 and S2. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

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

(2) Abagyan, R. A.; Totrov, M. M.; Kuznetsov, D. A. ICM: a new method for structure modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488–506.

(3) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470–89.

(4) Welch, W.; Ruppert, J.; Jain, A. N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.* **1996**, *3*, 449–462.

(5) Jones, G.; Wilett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727–748.

(6) McMartin, C.; Bohacek, R. QXP: Powerful, rapid computer algorithms for structure-based drug design. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 333–344.

(7) Morris, G. M.; Goodsell, D. S.; Halliday, R.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639–1662.

(8) Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins* **1998**, *33*, 367–382.

(9) Hou, T.; J., W.; Chen, L.; Xu, X. Automated docking of peptides and proteins by using a genetic algorithm combined with a tabu search. *Protein Eng.* **1999**, *12*, 639–647.

(10) Liu, M.; Wang, S. MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 435–451.

(11) Perola, E.; Xu, K.; Kollmeyer, T. M.; Kaufmann, S. H.; Prendergast, F. G.; Pang, Y. P. Successful virtual screening of a chemical database for farnesyltransferase inhibitor leads. *J. Med. Chem.* **2000**, *43*, 401–408.

(12) Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided. Mol. Des* **2001**, *15*, 411–428.

(13) Zavodszky, M. I.; Sanschagrin, P. C.; Korde, R. S.; Kuhn, L. A. Distilling the essential features of a protein surface for improving protein. ligand docking, scoring, and virtual screening. *J. Comput. -Aided. Mol. Des.* **2002**, *16*, 883–902.

(14) Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499–511.

(15) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.

(16) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, *57*, 225–242.

(17) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **2004**, *56*, 235–249.

(18) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of docking performance: Comparative data on docking algorithms. *J. Med. Chem.* **2004**, *56*, 235–249.

(19) Kroemer, R. T.; Vulpetti, A.; McDonald, J. J.; Rohrer, D. C.; Trosset, J. Y.; Giordanetto, F.; Cotesta, S.; McMartin, C.; Kihlen, M.; Stouten, P. F. W. Assessment of docking poses: Interactions-based accuracy classification (IBAC) versus crystal structure deviations. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 871–881.

(20) Kontoyianni, M.; Sokol, G. S.; McClellan, L. M. Evaluation of library ranking efficacy in virtual screening. *J. Comput. Chem.* **2005**, *26*, 11–22.

(21) 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.

(22) 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.

(23) 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.

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

(25) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graph. Model.* **2002**, *20*, 281–295.

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

(27) 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.

(28) 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.

(29) 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.

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

(31) Breiman, L. Bagging predictors. *Machine Learning.* **1996**, *24*, 123–140.

(32) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, Pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **1995**, *21*, 167–195.

(33) Koretke, K. K.; Luthey-Schulten, Z. A.; Wolynes, P. G. Self-consistently optimized energy functions for protein structure prediction by molecular dynamics. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 2932–2937.

(34) Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular system: Application to protein structure prediction. *J. Chem. Phys.* **2002**, *116*, 9058–9067.

(35) Fujitsuka, Y.; Takada, S.; Luthey-Schulten, Z. A.; Wolynes, P. G. Optimizing physical energy functions for protein folding. *Proteins* **2004**, *54*, 88–103.

(36) Chikenji, G.; Fujitsuka, Y.; Takada, S. Protein folding mechanisms and energy landscape of src SH3 domain studied by a structure prediction toolbox. *Chem. Phys.* **2004**, *307*, 157–162.

(37) Chikenji, G.; Fujitsuka, Y.; Takada, S. Shaping up the protein folding funnel by local interaction: Lesson from a structure prediction study. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 3141–3146.

(38) Fujitsuka, Y.; Chikenji, G.; Takada, S. SimFold energy function for de novo protein structure prediction: Consensus with Rosetta. *Proteins* **2006**, *62*, 381–398.

(39) Jin, W.; Kambara, O.; Sasakawa, H.; Tamura, A.; Takada, S. De novo design of foldable proteins with smooth folding funnel: automated negative design and experimental verifications. *Structure* **2003**, *11*, 581–590.

(40) Liu, Z.; Dominy, B. N.; Shakhnovich, E. I. Structural Mining: Self-Consistent Design on Flexible Protein-Peptide Docking and Transferable Binding Affinity Potential. *J. Am. Chem. Soc.* **2004**, *126*, 8515–8528.

(41) Camacho, C. J.; Vajda, S. Protein docking along smooth association pathways. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 10636–10641.

(42) Antes, I.; Merkwirth, C.; Lengauer, T. J. POEM: Parameter Optimization Using Ensemble Methods: Application to Target Specific Scoring Functions. *J. Chem. Inf. Model.* **2005**, *45*, 1291–1302.

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

(44) Böhm, H. J. Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 309–323.

(45) *FlexX*, version 1.12.2 L; BioSolveIT GmbH: Sankt Augustin, Germany.

(46) Shoham, M.; Scherf, T.; Anglister, J.; Levitt, M.; Merritt, E. A. HOL, WGJ. Structure diversity in a conserved cholera toxin epitope involved in ganglioside binding. *Protein Sci.* **1995**, *4*, 841–848.

CI700204V