A method to enhance the hit ratio by a combination of structure-based drug screening and ligand-based screening

Katsumi Omagari[1,5]
Daisuke Mitomo[1]
Satoru Kubota[1]
Haruki Nakamura[2,3]
Yoshifumi Fukunishi[2,4]

[1]Japan Biological Informatics Consortium (JBiC), Koto-ku, Tokyo, Japan; [2]Biomedicinal Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST), Koto-ku, Tokyo, Japan; [3]Institute for Protein Research, Osaka University, Suita, Osaka, Japan; [4]Pharmaceutical Innovation Value Chain, BioGrid Center Kansai, Toyonaka, Osaka, Japan; [5]Department of Virology, Medical School, Nagoya City University, Mizuho-ku, Nagoya, Japan

Correspondence: Yoshifumi Fukunishi
Biomedicinal Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6, Aomi, Koto-ku, Tokyo 135-0064, Japan
Tel +81 3 3599 8290
Fax +81 3 3599 8099
Email y-fukunishi@aist.go.jp

Abstract: We examined the procedures to combine two different *in silico* drug-screening results to achieve a high hit ratio. When the 3D structure of the target protein and some active compounds are known, both structure-based and ligand-based *in silico* screening methods can be applied. In the present study, the machine-learning score modification multiple target screening (MSM-MTS) method was adopted as a structure-based screening method, and the machine-learning docking score index (ML-DSI) method was adopted as a ligand-based screening method. To combine the predicted compound's sets by these two screening methods, we examined the product of the sets (consensus set) and the sum of the sets. As a result, the consensus set achieved a higher hit ratio than the sum of the sets and than either individual predicted set. In addition, the current combination was shown to be robust enough for the structural diversities both in different crystal structure and in snapshot structures during molecular dynamics simulations.

Keywords: *in silico* screening, consensus score, protein-based screening, protein-ligand docking, conformation of active site

Introduction

In silico drug screening results by structure-based screening methods strongly depend on the target protein's 3D structure (Kontoyianni et al 2005; Warren et al 2006). Generally speaking, *in silico* screening succeeds in providing good database enrichment in approximately half of the cases, and it fails in hit compound prediction in another half of the cases. In some cases, the prediction results are much worse than the results obtained by random screening. In addition, even a slight structural change around the binding site will sometimes have a large effect on the docking scores (DeWeese-Scott and Moult 2004; Oshiro et al 2004). To overcome this problem, when several docking programs and scoring functions are available, consensus scoring methods are used (Charifson et al 1999; Wang and Wang 2001; Clark et al 2002; Wang and Wand 2003; Yang et al 2005; Oda et al 2006; Teramoto and Fukunishi 2007). The basic idea of the consensus score is to take the product of sets of two or more compound's predicted by different scoring methods. Namely, the candidate hit compounds are the top scoring compounds common to two or more scoring functions (Charifson et al 1999). Several variations of the consensus scores have been proposed. For example, the ranking number of a compound is given by an average of the ranking numbers of two or more scoring functions (Wang and Wang 2001; Clark et al 2002; Wang and Wand 2003; Yang et al 2005; Oda et al 2006; Teramoto and Fukunishi 2007). Instead of modifying the docking score, a combination of the quantitative structure-activity relationship (QSAR) method and the scoring function was also examined (Hetényi et al 2006).

In structure-based screening, the 3D structure of target protein must be known. There have been many reports about the relationship between target modeling methods and

hit ratios (Bissantz et al 2003; Diller and Li 2003; McGovern and Shoichet 2003; DeWeese-Scott and Moult 2004; Oshiro et al 2004; Nicholls 2008; Sheridan 2008). Some reports have suggested that the hit ratios depend on the structure changes in the ligand-binding regions (DeWeese-Scott and Moult 2004; Oshiro et al 2004). McGovern and Shoichet (2003) have reported that the *holo* crystal structures would provide better enrichment than the *apo* crystal structures, and that the *apo* structures would provide better enrichment than the homology-modeled structures. However, the number of examples was small, and a more comprehensive study is needed to better our understanding of how to prepare the most suitable 3D-structure model for the target protein (Lerner et al 2007).

We have developed a protein-compound docking program and some *in silico* screening methods based on the protein-compound affinity matrix (Fukunishi et al 2005a, 2005b, 2006a, 2006b, 2006c, 2006d, 2007). One screening method is a structure-based screening method called the multiple target screening (MTS) method (Fukunishi et al 2005b, 2006a), and the other is a ligand-based screening method called the docking score index (DSI) method (Fukunishi et al 2006b, 2006c, 2006d, 2007). When the 3D structure of the target protein is known and some active compounds are known, we can apply both the structure-based and the ligand-based screening methods (Charifson et al 1999). A simple method to combine the results from these two methods is to take the sum of the two sets of predicted compounds. Another method is to take the consensus set (product of the two sets) of the predicted compounds. In the present study, we attempted to determine which method gives a better hit-ratio. In addition, we observed the robustness of these methods against the structural diversities, using several different crystal structures and snap-shot structures during the molecular dynamics trajectories.

Our docking program, Sievgene, showed a standard accuracy in a self-docking test (Fukunishi et al 2005a), and the cross-docking results could be improved by the assistance of the known protein-compound complex structures by using the maximum volume overlap method (Fukunishi and Nakamura 2008). Here, the dependence of screening performance on the target structure was examined.

Methods
Conventional screening (CS) method and protein-compound docking procedure
In the conventional screening (CS) method, compounds that show a strong affinity with the target protein are selected by protein-compound docking software based on the 3D structure of the target protein.

In the present study, all protein-compound dockings were performed by our program, called Sievgene (Fukunishi et al 2005a), which is a protein-ligand flexible docking program for *in silico* drug screening. This program generates many conformers (default is up to 100 conformers) for each compound, and keeps the target protein structure rigid but with soft interaction forces altering its structure to some extent (Fukunishi et al 2005a). Our docking program, Sievgene, is a part of the myPresto (prestoX) system, which is available, free for academic use, from the following website, http://presto. protein.osaka-u.ac.jp/myPresto4/index_e.html.

Multiple target screening (MTS) method
In the present study, the MTS method was used to select the sum of the compound sets predicted by the multiple active site correction (MASC) scoring method and the original MTS method, as described previously (Fukunishi et al 2005b). The original version of the MTS method is based on a protein-compound affinity matrix, and the compounds that show strong affinity with the target protein among many proteins are selected as candidate hit compounds.

First, let us briefly explain the MTS method. We prepared a set of proteins $P = \{p_1, p_2, p_3, ..., p_M\}$, where p_a represents the a-th protein. The total number of proteins is M. We also prepared a set of compounds $X = \{x^1, x^2, ..., x^N\}$, where x^i represents the i-th compound. The total number of compounds is N. For each protein p_a, all compounds of set X are docked to protein p_a with score s_a^i between the a-th protein and the i-th compound. Here, s_a^i corresponds to the binding free energy; a lower s_a^i means a higher affinity between the a-th protein and the i-th compound.

For the i-th compound, $\{s_a^i; a = 1, ..., M\}$ were sorted in descending order, and the order n_a^i assigned to each a-th protein depended on its value s_a^i. For example, when $n_a^i = 1$, the a-th protein binds the i-th compound with the strongest affinity. When $n_a^i = M$, the a-th protein binds with the weakest affinity. This procedure was repeated until the order $\{n_a^i; a = 1, ..., M/i = 1, ..., N\}$ was determined for all compounds. Next, we focus on the target a-th protein. The compounds having the order $n_a^i = 1$ are assigned as members in the compound group-1, compounds having $n_a^i = 2$ are assigned as members in compound group-2, and so on. Among the group-1 members, the compound with the lowest s_a^i should be the most probable hit compound. If there is no compound in group-1, the compound with the lowest n_a^i in group-2 should be the most probable hit

compound. This procedure is repeated until the most probable hit compound is found. This procedure is the so-called MTS method.

The MASC score s'^i_a for the a-th protein and the i-th compound has been reported by Vigers and Rizzi (2004) as follows:

$$s'^i_a = (s^i_a - \mu_i)/\sigma_i, \qquad 1$$

where s^i_a is the raw docking score for the a-th protein and the i-th compound, and μ_i and σ_i are the average and standard deviation of the raw docking scores across all proteins for the i-th compound, respectively. In this method, s'^i_a is used for screening instead of s^i_a.

Both the MTS and the MASC scoring methods are applied in this study, and the combination of the results by these two methods (sum of sets) is taken as the set of candidate hit compounds. Namely, to get the top-ranked N compounds, the same numbers of compounds are tak6 MTS and the MASC scoring methods, and the sum of the two sets gives the total N compounds.

Direct score modification (DSM) MTS method

In the DSM-MTS method, the score of a compound score on the a-th protein is a weighted average of the raw scores of that compound on proteins that are similar to the a-th protein. The DSM-MTS method modifies the raw docking score as follows (Fukunishi et al 2006a):

$$s^{DSM\,i}_a = \frac{\sum_b s^i_b R^b_a}{\sum_b R^b_a} c \qquad 2$$

where is the so-called "DSM score" of the i-th compound $s^{DSM\,i}_a$ for the a-th protein, s^i_b is the raw docking score of the i-th compound for the b-th protein, and R^b_a is the correlation coefficient between the a-th and b-th proteins

$$R^b_a = \frac{\sum_i \left(s^i_b - \frac{\sum_i s^i_b}{N_c} \right)\left(s^i_a - \frac{\sum_i s^i_a}{N_c} \right) + \varepsilon}{\sqrt{\sum_i \left(s^i_b - \frac{\sum_i s^i_b}{N_c} \right)^2 \cdot \sum_i \left(s^i_a - \frac{\sum_i s^i_a}{N_c} \right)^2} + \varepsilon}. \qquad 3$$

In the DSM-MTS method, the potential active compounds were selected by the MTS method through the DSM score instead of the original docking score.

Machine-learning score modification (MSM) MTS method

In the MSM-MTS method, the score of a compound score on a protein is a weighted average of the raw scores of that compound on proteins and the weight is determined to maximize the hit ratio of the known active compounds. Equation 2 is rewritten as

$$s^{new\,i}_a = \sum_b s^i_b M^b_a \qquad 4$$

where $s^{new\,i}_a$, s^i_b, and M^b_a are the new docking scores of the i-th compound with the a-th protein, the raw docking score of the i-th compound with the b-th protein, and the parameters, respectively.

If known active compounds are available, we can determine M^b_a in eq. 4 to maximize the database enrichment (Fukunishi et al 2006a). Let x and $f(x)$ be the numbers of compounds (%) selected from the total compound library and from the database enrichment curve, respectively. The surface area under the database enrichment curve (q) is a measure of the database enrichment.

$$q = \int_0^{100} f(x)dx \qquad 5$$

Higher q values correspond to better database enrichment, and $0 < q < 100$. The optimal M^b_a is determined by a Monte Carlo method to maximize the q value. The a-b element of the new matrix $\boldsymbol{M}$ ($M^{new\,b}_a$) is given by $M^{new\,b}_a = M^b_a + \eta^b_a$; here, η^b_a is a random number. Using the newly generated matrix, the new docking score is calculated by eq. 4. Then an *in silico* screening method based on the new matrix $\boldsymbol{M}$ gives the q value by eq. 5. The best matrix $\boldsymbol{M}$, which gives the highest q value, is selected as the seed matrix for the next optimization step. This process is repeated until the q value shows convergence.

Machine-learning docking score index (ML-DSI) method

The docking score index (DSI) method is a ligand-based screening method that utilizes molecular descriptors. The descriptors of compounds are docking scores with many proteins, and compounds that are similar to active compounds are selected as candidate hit compounds. In the framework of the DSI method, a measure to represent the distance between two compounds is determined based on the protein-compound affinity matrix. From the covariance matrix of the compounds, principal component analysis (PCA) is performed to find similar clusters of compounds. This DSI method has

been described in detail in previous papers (Fukunishi et al 2006b, 2007) and is briefly introduced below.

We prepare a set of proteins $P = \{p_1, p_2, p_3, ..., p_{Nr}\}$, where p_i represents the i-th protein and Nr the total number of proteins, and a set of compounds $X = \{x^1, x^2, ..., x^{Nc}\}$, where x^k represents the k-th compound and Nc the total number of compounds. For each protein p_i, all compounds of the set X are docked to the protein p_i with a score of s_i^k between the i-th protein and the k-th compound. Here, s_i^k corresponds to the binding free energy.

The covariance matrix M^P of the proteins is defined as

$$M^P_{ij} = \frac{1}{N_c} \sum_{k=1}^{N_c} (s_i^k - \overline{s_i})(s_j^k - \overline{s_i}),$$ 6

and

$$\overline{s_i} = \frac{1}{N_c} \sum_{k}^{N_c} s_i^k$$ 7

where the upper bar represents an average. Let ϕ_j be the j-th eigenvector of M^P with an eigenvalue ε_j, and let the order of ε_j be descendant. The vector of docking scores for the k-th compound $X_k = (s_1^k, s_2^k, ..., s_{Nr}^k)$ is represented by the linear combination of ϕ_j

$$X_k = \sum_{j=1}^{Nr} c_j^k \phi_j.$$ 8

The coefficient $\{c_j^k\}$ represents the j-th coordinate of the PCA space of the k-th compound. In this study, we call this coefficient $\{c_j^k\}$ the "docking score index (DSI)".

Candidate hit compounds are selected using the following method. In the PCA space, compounds that are close to the known active compounds are selected as the candidate hit compounds. In the original version of the DSI method, the distance from the k-th compound to the average position of the active compounds (D_k) is defined as

$$D_k = \sqrt{\sum_{j=\{\alpha1, \alpha2, ..., \alpha Nselect\}} (c_j^k - \overline{c_j})^2}$$ 9

and

$$\overline{c_j} = \sum c_j^{active} / N_a$$ 10

where c_j^{active} and N_a are the DSI values of the active compounds and the total number of active compounds. The suffix j runs over the selected axes $\{\alpha_1, \alpha_2, ..., \alpha_{Nselect}\}$.

The principal component axes are selected in the following manner. The contribution of each principal component is estimated using a database enrichment curve. The surface area under the database enrichment curve q_α is evaluated for the α-th principal component axis; namely, the suffix j in eq. 9 is set as α and N_{select} is set as 1, and the database enrichment curve f_α is calculated for the α-th axis, the same as in eq. 5. The q_α values are calculated by

$$q_\alpha = \int_0^{100} f_\alpha(x) dx,$$ 11

where x and $f_\alpha(x)$ are the percentages of compounds that are selected from the total compound library and the database enrichment curve, respectively.

The axes are sorted in descending order with respect to the q_α value. The q value given by eq. 5 is a measure of the database enrichment in addition to q_α in eq 11. The q value is calculated by changing the number of axes (N_{select}) used in eq. 9 to find the optimal N_{select} value, which gives the maximum q value.

To apply the DSI method, the known active compounds must be available; hence, the docking score can be modified to increase the database enrichment. If the new docking score is given by the linear combination of the docking scores with many proteins, as given by eq. 4, we can optimize the coefficients M_a^b to maximize the q value as in the MSM-MTS method. The MSM-MTS method is a QSAR method with machine learning. Many reports about the machine learning in ligand-based drug screening have been published and showed that the machine learning is very effective in drug screening (Anzali et al 1996; Livingstone 1996; Zupan and Gasteiger 1999; Burkard 2005; Plewczynski et al 2006; Ivanciuc 2007).

Materials
Protein set
Our screening methods are based on a protein-compound affinity matrix. We therefore must prepare a set of proteins including target proteins. The set of proteins consisted of the 20 target proteins and a basic protein set of 150 proteins. The target protein structures were 20 protein structures of the human immunodeficiency virus protease-1 (HIVP), cyclooxygenase-2 (COX2), thermolysin (THR), and gluta-thione S-transferase (GST) (Appendix A). Each protein has one *apo* structure. The complex structures of these target proteins include different ligands, and they show different binding modes.

For each target structure, an individual protein set was prepared. Each protein set consisted of a basic protein set and

the target structure itself. All structures of the basic protein set were crystal structures. The target structures were prepared in two ways, as a crystal structure and as a final snapshot structure generated by a molecular dynamics simulation. The preparation procedure is described in the next section.

The basic protein set was composed of 150 proteins, whose crystal structures of protein-ligand complexes are known (Appendix B). They were the same ones used to evaluate the docking programs, GOLD and FlexX (Nissink et al 2002), and they were also used in our previous works to evaluate our *in silico* screening program (Fukunishi et al 2006a). The data set contains a rich variety of proteins and compounds whose structures were all determined by high-quality experiments with a resolution of less than 2.5 Å. Almost all of the atom coordinates are supplied, except for those of the hydrogen atoms, and the all-atomic structures around the ligand proteins are quite reliable. In none of the complex structures are there any covalent interactions between the proteins and ligands.

Preparation of protein structures

Two types of protein models were prepared. The first model (model A) is the original crystal structure, which includes the *holo* and *apo* crystal structure. The bound ligand, co-activator, and water molecules were removed from the crystal structure. Lost hydrogen atoms were added to the protein structures without water and cofactors by using the program Tplgene in the myPresto suite (Fukunishi et al 2003, 2005a). The charges of protein atoms were originated from AMBER parm99 (Wang et al 2000; Case et al 2004).

The other model (model B) generated from the *holo* and *apo* crystal structure given by MD in explicit water was prepared as follows. The whole structure of each protein was embedded in a sphere of TIP3P (Jorgensen et al 1983; Jorgensen and Madura 1985) water (CAP water) including ion particles of 0.1% Na^+ and Cl^- in order to neutralize the total charge of the systems. The center of the sphere was set at the mass center of the protein, and the radius was 42, 52, 42, and 37 Å for HIVP, COX2, THR, and GST, respectively. Before performing MD calculations for the entire system, an MD calculation for only the solvent parts (solvent water and counter ions) was performed with the protein, ligand, and metal ions coordinates fixed, so as to bring the solvent parts sufficiently closer to an equilibrium state. MD simulations of the entire system were performed using 0.5-fsec time steps with the temperature set at 310 K and the fast multipole method (Greengard and Rokhlin 1987) being used for calculation of the Coulombic interaction. The cut-off

distance of the van der Waals interaction was 10.0 Å. The final snapshot structures of *holo* and *apo* crystal structures at 500 psec simulations were used as model B. The MD simulations were performed by using the cosgene/myPresto (Fukunishi et al 2003).

Preparation of chemical compound libraries

The used compound library consisted of 110 known ligands for 20 target proteins, and 11,050 compounds of the Coelacanth chemical compound library (Coelacanth Corporation, East Windsor, NJ, USA), which is a random library, as a decoy set.

The 3D coordinates of 11,050 chemical compounds of the Coelacanth chemical compound library were generated by the Concord program (Tripos, St. Louis, MO, USA) from the 2D Sybyl SD files provided by the Coelacanth Chemical Corporation. The 3D coordinates of the known ligands were generated by the Chem3D program (Cambridge Software, Cambridge, MA, USA). The atomic charges of each compound in the compound library were determined by the Gasteiger method (Gasteiger and Marsili 1978; Gasteiger and Marsili 1980).

Results

The protein-compound affinity matrixes were calculated for each target, and we then applied CS, MTS, DSM-MTS, MSM-MTS, and ML-DSI methods. We took the consensus of the predicted compound's sets by the MSM-MTS and ML-DSI methods and designated it as the "consensus set." In contrast, we collected the sum of the predicted compound's sets by the MSM-MTS and ML-DSI methods and called it the "sum set." Protein models A and B were used as the crystal structures and the MD snapshot structures, respectively. Figure 1 shows the schematic representation of the screening procedure.

Table 1 shows the q values and hit ratios at the first 1% of the entries in the database obtained by various screening methods for the 20 target proteins. Among these screening methods, the consensus set gave the best q value and the highest hit ratio for models A and B. The q value and the hit ratio for model A are close to the values for model B, respectively. In addition, the q value and hit ratio did not significantly depend on differences in the target protein.

The sum set gave the second best q value among these screening methods. The hit ratio by the sum set was almost the average value of that by the MSM-MTS method and that by the ML-DSI method. Similar to with the consensus set, the q value and the hit ratio for model A was close to the values for model B. The q value and hit ratio did not significantly
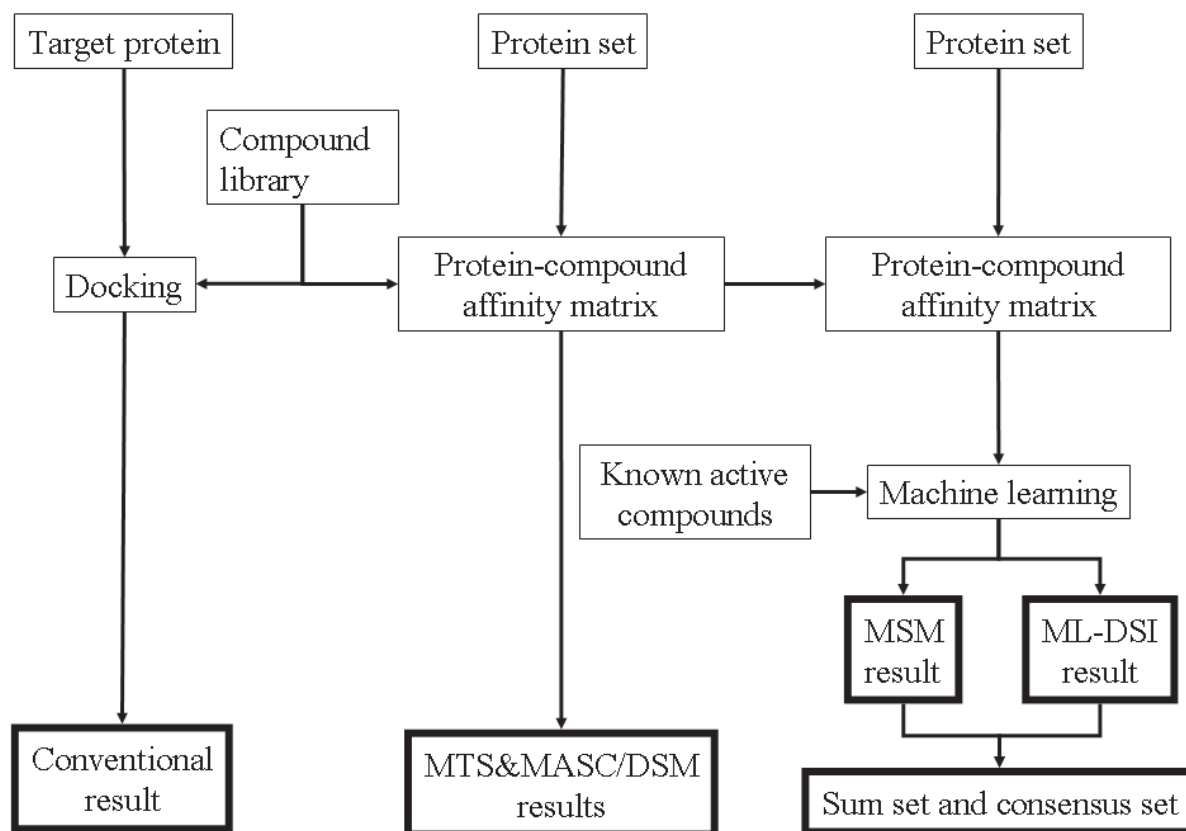
Figure 1 Schematic representation of the screening methods in the current study. The same procedure was applied to models A and B. The protein set consists of the proteins listed in Appendixes A and B.

depend on differences in the target protein. The q values obtained by the consensus set, the sum set, and the MSM-MTS and ML-DSI methods were 89.6–98.6, 90.4–97.2, 85.8–97.4, and 75.6–99.5, respectively.

Without known active compounds, we can apply the CS, MTS, and DSM-MTS methods. The MTS method gave the best q value, the DSM-MTS method gave the second best q value, and the CS method gave the worst q value among these three methods. In contrast, the CS method gave the best hit ratio, the DSM-MTS method gave the second best hit ratio, and the MTS method gave the worst hit ratio among these three methods. In general, the results obtained by the CS, MTS, and DSM-MTS methods strongly depended on the target structure model. These three methods gave good q values for HIVP. The MTS and DSM-MTS methods gave good q values for COX2, while the CS method gave poor q values for COX2. These three methods gave poor q values for THR and GST, with the DSM-MTS method giving the worst q values for GST.

Figure 2 shows the database enrichment curves obtained by the MSM-MTS method, the ML-DSI method, the consensus set, and the sum set. The database enrichment curves

by the consensus and sum sets were close to the database enrichment curve by the MSM-MTS method rather than that by the ML-DSI method. The database enrichment curve by the consensus set was the highest among these four curves in the range of the number of selected compounds <20%. The database enrichment curve by the sum set was quite close to that by the consensus set. When the number of selected compounds was more than 20%, the sum set gave the highest database enrichment curve among these four curves.

The RMSD values between models A and B were approximately 2 Å, and all values were less than 3 Å. The RMSD values of model B were approximately 1 Å bigger than those of model A. The q values of model B by the CS, the MTS and the DSM-MTS methods were much worse than the q values of model A. The result is consistent with the previous report (McGovern and Shoichet 2003). When active compounds are known for the target protein, the MSM-MTS and the ML-DSI methods are available. The q values by the MSM-MTS and the ML-DSI methods did not depend on the RMSD value or the difference in modeling methods. This result suggests that the MSM-MTS and ML-DSI methods are robust against

Table 1 q values and hit ratios for target protein models A and B

Protein[a]	PDB[b]	RMSD[c] A	RMSD[c] B	Conventional A	Conventional B	MTS&MASC[d] A	MTS&MASC[d] B	DSM[e] A	DSM[e] B	MSM[f] A	MSM[f] B	ML-DSI[g] A	ML-DSI[g] B	Sum[h] A	Sum[h] B	Consensus[i] A	Consensus[i] B
HIVP	3hvp*	0.0	2.0	89.4	85.4	89.0	83.5	89.0	89.2	91.2	90.9	83.1	82.2	90.6	90.9	89.6	90.3
	1aid	1.4	1.8	85.3	87.6	86.9	87.5	88.8	89.6	91.7	91.7	80.0	83.4	92.4	91.5	90.5	90.2
	1hpx	2.3	2.6	95.4	80.2	94.7	81.5	94.3	89.7	92.2	92.0	84.4	80.6	91.8	91.4	90.7	91.6
	1htfl	2.4	2.3	88.9	87.7	82.0	83.2	92.5	90.6	91.0	90.5	83.3	85.9	90.9	90.4	92.8	91.5
	1htf2	2.4	2.0	83.2	87.2	71.9	80.0	93.0	90.7	92.2	92.7	77.4	75.6	91.9	91.5	91.8	91.3
	1ivp	2.2	2.5	82.7	64.4	74.8	49.5	90.9	49.5	91.0	91.4	80.4	77.6	91.0	91.2	91.0	90.0
	4phv	2.4	2.6	92.3	83.9	79.8	73.9	90.4	91.5	92.2	91.3	79.3	84.8	91.2	91.0	89.9	90.3
average				88.2	82.3	82.7	77.0	91.3	84.4	91.7	91.5	81.1	81.5	91.4	91.1	90.9	90.7
σ_q				4.4	7.7	7.5	11.9	1.9	14.3	0.5	0.7	2.3	3.5	0.6	0.4	1.0	0.6
COX2	5cox*	0.0	1.6	46.7	33.1	89.2	70.5	55.5	49.1	94.3	95.0	89.4	80.1	96.8	96.2	98.2	97.8
	1cx2	0.7	1.6	55.8	44.3	92.0	83.0	98.7	97.4	95.2	94.8	91.0	82.2	95.9	96.8	98.0	98.5
	1pxx	0.8	1.4	46.3	33.2	85.8	78.4	99.1	78.4	95.4	95.0	88.5	84.0	96.8	96.7	98.1	98.3
	4cox	0.6	1.3	45.0	16.4	82.2	52.9	95.6	31.7	97.4	95.7	99.5	89.0	96.5	97.2	98.5	98.6
	6cox	0.6	1.5	55.2	12.3	91.3	31.9	95.1	93.7	94.9	95.4	81.7	87.6	96.1	96.0	98.4	97.9
average				49.8	27.9	88.1	63.3	88.8	70.1	95.4	95.2	90.0	84.5	96.4	96.6	98.2	98.2
σ_q				4.7	11.8	3.7	18.8	16.7	25.6	1.0	0.3	5.7	3.3	0.3	0.4	0.2	0.5
THR	1l3f*	0.0	1.3	44.5	9.0	63.1	24.9	69.9	13.6	93.5	95.5	81.5	79.3	95.3	93.7	95.9	95.7
	1tlp	1.0	1.4	42.9	42.3	57.7	59.5	16.7	16.8	95.0	95.2	87.1	83.7	95.2	95.2	95.8	95.7
	1tmn	0.9	1.2	46.1	42.1	56.1	60.5	15.8	87.8	95.4	95.0	85.4	84.2	95.0	95.2	95.9	96.1
	2tmn	0.9	1.5	52.3	47.4	75.7	73.0	53.5	38.1	95.5	95.6	84.8	85.8	95.4	95.4	96.0	96.0
average				46.4	35.2	63.2	54.5	39.0	39.1	94.9	95.3	84.7	83.2	95.2	94.8	95.9	95.9
σ_q				3.6	15.3	7.7	17.9	23.5	29.7	0.8	0.2	2.0	2.4	0.1	0.7	0.1	0.2
GST	16gs*	0.0	2.6	69.3	72.7	50.8	60.9	28.0	24.2	86.3	85.8	82.3	82.4	91.6	92.8	92.7	94.3
	18gs	0.4	2.0	71.5	70.4	62.9	48.6	29.4	26.5	85.9	87.4	87.9	83.3	91.4	93.4	94.9	93.0
	2gss	0.3	1.3	67.4	70.4	54.3	59.7	26.5	28.8	87.8	86.1	86.0	85.9	92.4	93.7	93.3	95.5
	3pgt	0.5	1.7	68.2	63.8	46.4	34.7	25.0	23.3	87.5	88.1	82.5	77.7	92.1	92.3	94.7	94.9
average				69.1	69.3	53.6	51.0	27.2	25.7	86.8	86.9	84.6	82.3	91.9	93.1	93.9	94.4
σ_q				1.5	3.3	6.1	10.6	1.6	2.1	0.8	0.9	2.4	2.9	0.4	0.6	0.9	0.9
q value average				66.4	56.7	74.3	63.9	67.4	60.0	91.7	91.4	81.4	83.7	92.9	92.5	93.6	93.4
hit ratio at 1%				23.8	16.2	15.1	7.3	21.1	14.7	35.8	29.3	16.7	16.1	28.8	25.4	47.2	46.0

Notes: The q value is the area under the database enrichment curve. The q value is 50 for random screening and the maximum value is 100. The hit ratio is the hit ratio at the first 1% of the entries in the database. The σ_q value is the standard deviation of the q values. Model A includes the original crystal structures of the target proteins. Model B includes the model structures of the target proteins obtained by the MD simulations in explicit water. a: name of target protein. b: PDB code. "*" represents apo structure. c: RMSD value from the apo structure of protein model A. d: combination of MTS and MASC methods (sum set) with original docking score. e: MTS method with the DSM method. f: MSM method with the MSM-MTS method). g: ML-DSI method. h: sum set of predicted compounds by the MSM-MTS and the ML-DSI methods. i: consensus set of predicted compounds by the MSM-MTS and the ML-DSI methods.
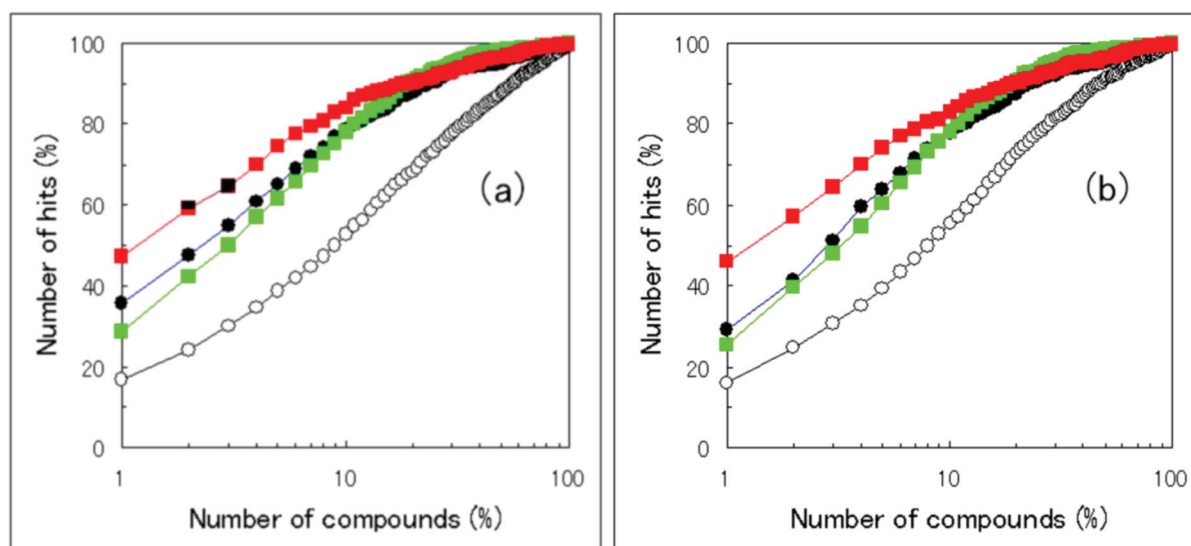
Figure 2 Database enrichment curves for models A and B. Filled circles, open circles, green squares, and red squares represent the results by the MSM-MTS method, the ML-DSI method, the sum sets of predicted compounds by the MSM-MTS and the ML-DSI methods, and the consensus sets of predicted compounds by the MSM-MTS and the ML-DSI methods, respectively. a: database enrichment curves for model A, in which the target protein structures are the original crystal structures. b: database enrichment curves for model B, in which the target protein structures are the model structures obtained by the MD simulations in explicit water.

the structure change on induced fitting of the target protein as well as the consensus set and the sum set.

In all cases, the q values for *apo* structures were not as small as those for *holo* structures. In some cases, the q values for *apo* structures were better than the averaged q values. These results show that *in silico* screening could be applied to both *apo* and *holo* structures.

Discussion

Let us select the top-ranked N compounds by one method, which we will call "method 1" (in this study, the MSM-MTS method) and the other top-ranked N compounds by another method, called "method 2" (in this study, the ML-DSI method). Suppose that N_a^1 active compounds are included in the N compounds by method 1, N_a^2 active compounds are included in the N compounds by method 2, and N_a active compounds are selected by both methods 1 and 2. The number of inactive compounds by method 1 is $N - N_a^1$ and that by method 2 is $N - N_a^2$. Suppose that the number of inactive compounds selected by both methods 1 and 2 is N_i. The hit ratio by the consensus set of compounds selected by methods 1 and 2 is $N_a/(N_a + N_i)$, and the hit ratio by the sum set of compounds selected by methods 1 and 2 is $(N_a^1 + N_a^2 - N_a)/(2N - N_i - N_a)$. The hit ratio by the consensus set increases when N_a increases or N_i decreases. The hit ratio by the sum set increases when N_i increases or N_a decreases. For example, if N = 100, $N_a^1 = 36$, $N_a^2 = 17$, $N_a = 13$, and $N_i = 15$, the hit ratio for model A in Table 1

could be explained by these numbers. The results in Table 1 suggest that the N_a value is large and the N_i value is small. Thus the same active compounds should frequently be selected by both the MSM-MTS and the ML-DSI methods in the top-ranked compounds.

Klon and colleagues (2004a, 2004b, 2004c) have proposed a combination of structure-based and ligand-based screening. In their method, the conventional structure-based screening was performed by a protein-compound docking program such as Glide, FlexX (Rarey et al 1996), or GOLD (Jones et al 1997). These three programs succeeded in finding the active compounds. The top-ranked compounds found by the structure-based screening were designated as candidate active compounds, while all other compounds were designated as candidate inactive compounds. The fingerprints were then calculated for all compounds in the database. A Bayesian model was trained using the fingerprints from the candidate active and inactive compounds. Finally, all of the compounds were re-ranked according to the Bayesian model. This method worked well and succeeded in increasing the hit ratio. This result suggests that the top-ranked active compounds by a structure-based screening have some common 2D descriptors (Klon et al 2004a, 2004b, 2004c). Thus the current consensus set of predicted compounds by structure-based and ligand-based screening should in general give a high hit ratio.

In the current study, the ML-DSI method was called the ligand-based screening method, while the target protein

structure was used in the ML-DSI method. The DSI method is a kind of QSAR method, and the screening results by the DSI method do not strongly depend on the existence of the target structure in the used protein set (Fukunishi et al 2006a). It is also true for the ML-DSI method (Fukunishi et al 2006b). Thus, the ML-DSI method can be called as the ligand-based screening method.

In some cases, model B, which is a snap shot structure obtained by the MD, gave better hit ratio than model A, which is an energy-optimized crystal structure. The MD simulation gave the ligand-binding protein of slightly larger size than the energy-optimized crystal structure some times. The screening results obtained by the energy-minimized structures of *apo* proteins in vacuum were almost random screening results. The energy minimization in vacuum gave the ligand-binding protein of smaller size than the energy-optimized crystal structure of *holo* protein. The slightly large size protein can accept the active compounds, which are different from the ligand of the *holo* protein structure, while the small size protein cannot accept any active compound. That should be why model B could give the better results than model A in some cases.

Conclusion

We examined a method for combining the predicted compound sets obtained by structure-based and ligand-based screenings. We found that the product of the two sets (consensus set) could give a better hit ratio than the individual hit ratios obtained by the structure-based and ligand-based screenings. The database enrichment by the sum of the two sets (sum set) was close to that obtained by the product of the two sets, but the hit ratio by the consensus set was higher than that by the sum set when less than 20% of the compounds of the database were selected.

For structure-based screening, an X-ray crystallography structure can give better results than a modeled structure given by a molecular dynamics simulation. When several active compounds are identified, both the X-ray crystallography structure and the MD structure can give a high hit ratio, and the hit ratios are robust against structural changes in the target protein such as induced fitting. However, without any known active compounds, the results by the CS, MTS, and DSM-MTS methods strongly depend on structural changes of the target proteins.

We could not say which structure is better, the *apo* structure or *holo* structure, when Sievgene was used as docking software. Generally speaking, *holo* structures are more suitable for *in silico* screening and docking study than *apo* structures.

In some cases, screening based on *apo* structures could provide better a hit ratio than that on *holo* structures.

Acknowledgments

Appendix A

The protein databank (PDB) identifier list of the target protein set is: the PDB codes of COX2's crystal structures are 1cx2, 1pxx, 4cox, 6cox, and 5cox; those of THR are 1tlp, 1tmn, 2tmn, and 1l3f; those of HIVP are 1aid, 1hpx, 1ivp, 1htf1, 1htf2, 4phv, and 3hvp; those of GST are 18gs, 2gss, 3pgt, and 16gs. The *apo* forms are 5cox, 1l3f, 3hvp, and 16gs.

Appendix B

The protein databank (PDB) identifier list of the basic protein set is: 1a28, 1a42, 1a4g, 1a4q, 1abe1, 1abe2, 1abf1, 1abf2, 1aco, 1ai5, 1aoe, 1apt, 1apu, 1aqw, 1atl, 1b58, 1b9v, 1bma, 1byb, 1byg, 1c1e, 1c5c, 1c83, 1cbs, 1cbx, 1cdg, 1ckp, 1com, 1coy, 1cps, 1cvu, 1d0l, 1d3h, 1dd7, 1dg5, 1dhf, 1dog, 1dr1, 1ebg, 1eed, 1ejn, 1epb, 1epo, 1ets, 1f0r, 1f0s, 1f3d, 1fen, 1fkg, 1fki, 1fl3, 1glp, 1hdc, 1hfc, 1hos, 1hpv, 1hsb, 1hsl, 1htf1, 1htf2, 1hyt, 1ida, 1ivb, 1jap, 1lah, 1lcp, 1lic, 1lna, 1lst, 1mdr, 1mld, 1mmq, 1mrg, 1mts, 1mup, 1nco, 1ngp, 1nis, 1okl, 1pbd, 1phd, 1phg, 1poc, 1ppc, 1pph, 1pso, 1qbr, 1qbu, 1qpq, 1rds, 1rne, 1rnt, 1rob, 1snc, 1srj, 1tlp, 1tmn, 1tng, 1tnh, 1tni, 1tnl, 1tyl, 1xid, 1xie, 1yee, 2aad, 2ack, 2ada, 2cht, 2cmd, 2cpp, 2ctc, 2fox, 2gbp, 2ifb, 2pk4, 2qwk, 2tmn, 3cla, 3cpa, 3erd, 3ert, 3tpi, 4lbd, 4phv, 5abp1, 5abp2, 5cpp, 5er1, 6rnt, and 7tim. For 1abe, 1abf, 5abp, and 1htf, two protein proteins were prepared, since these proteins each bind two kinds of ligands.

References

Anzali S, Barnickel G, Krug M, et al.; Devillers J, (Ed). 1996. Neural networks in QSAR and drug design. London: Academic Press, pp. 209–22.

Burkard U. 2005. In: Gasteiger J, Engel T, (Ed). Chemoinfomatics. Weinheim: Wiley-VCH, pp. 435–81.

Bissantz C, Bernard P, Hibert M, et al. 2003. Protein-based virtual screening of chemical databases. II. Are homology models of G-Protein Coupled Receptors suitable targets? *Proteins*, 50:5–25.

Case DA, Darden TA, Cheatham III TE, et al. 2004. AMBER8. University of California, San Francisco.

Charifson PS, Corkery JJ, Murcko MA, et al. 1999. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem*, 42:5100–9.

Clark RD, Strizhev A, Leonard JM, et al. 2002. Consensus scoring for ligand/protein interactions. *J Mol Graph Model*, 20:281–95.

DeWeese-Scott C, Moult J. 2004. Molecular modeling of protein function regions. *Proteins*, 55:942–61.

Diller DJ, Li R. 2003. Kinases, homology models, and high throughput docking. *J Med Chem*, 46:4638–47.

Fukunishi Y, Mikami Y, Nakamura H. 2003. The filling potential method: A method for estimating the free energy surface for protein-ligand docking. *J Phys Chem B*, 107:13201–10.

Fukunishi Y, Mikami Y, Nakamura H. 2005a. Similarities among receptor proteins and among compounds: analysis and application to in silico ligand screening. *J Mol Graph Model*, 24:34–45.

Fukunishi Y, Mikami Y, Kubota S, et al. 2005b. Multiple target screening method for robust and accurate in silico ligand screening. *J Mol Graph Model*, 25:61–70.

Fukunishi Y, Mikami Y, Takedomi K, et al. 2006a. Classification of chemical compounds by protein-compound docking for use in designing a focused library. *J Med Chem*, 49:523–33.

Fukunishi Y, Hojo S, Nakamura H. 2006b. An efficient in silico screening method based on the protein-compound affinity matrix and its application to the design of a focused library for cytochrome P450 (CYP) ligands. *J Chem Inf Model*, 46:2610–22.

Fukunishi Y, Kubota S, Nakamura H. 2006c. Noise reduction method for molecular interaction energy: application to in silico drug screening and in silico target protein screening. *J Chem Inf Model*, 46:2071–84.

Fukunishi Y, Kubota S, Kanai C, et al. 2006d. A virtual active compound produced from the negative image of a ligand-binding protein, and its application to in-silico drug screening. *J Comput Aided Mol Des*, 20:237–48.

Fukunishi Y, Kubota S, Nakamura H. 2007. Finding ligands for G protein-coupled receptors based on the protein-compound affinity matrix. *J Mol Graph Model*, 25:633–43.

Fukunishi Y, Nakamura H. 2008. Prediction of protein-ligand complex structure by docking software guided by other complex structures. *J Mol Graph Model*, 26:1030–3.

Gasteiger J, Marsili M. 1978. A New Model for Calculating Atomic Charges in Molecules. *Tetrahedron Lett*, 19:3181–4.

Gasteiger J, Marsili M. 1980. Iterative partial equalization of orbital electronegativity – A rapid access to atomic charges. *Tetrahedron*, 36:3219–28.

Greengard L, Rokhlin V. 1987. A fast algorithm for particle simulations. *J Comp Phys*, 73:325–48.

Hetényi C, Paragi G, Maran U, et al. 2006. Combination of a modified scoring function with two-dimensional descriptors for calculation of binding affinities of bulky, flexible ligands to proteins. *J Am Chem Soc*, 128:1233–9.

Ivanciuc O; Lipkowitz KB, Cundari TR, (Ed). 2007. Reviews in Computational Chemistry. Weinheim: Wiley-VCH, 23:291–400.

Jorgensen WL, Chandrasekhar J, Madura J, et al. 1983. Comparison of simple potential. Functions for simulating liquid water. *J Chem Phys*, 79:926–35.

Jorgensen WL, Madura JD. 1985. Temperature and size dependence for Monte Carlo simulations of TIP4P water. *Mol Phys*, 56:1381–92.

Jones G, Willett P, Glen RC, et al. 1997. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol*, 267:727–48.

Klon AE, Glick M, Davies JW. 2004a. Combination of a naive Bayes classifier with consensus scoring improves enrichment of high-throughput docking results. *J Med Chem*, 47:4356–9.

Klon AE, Glick M, Davies JW. 2004b. Application of machine learning to improve the results of high-throughput docking against the HIV-1 protease. *J Chem Inf Comput Sci*, 44:2216–24.

Klon AE, Glick M, Thoma M, et al. 2004c. Finding more needles in the haystack: A simple and efficient method for improving high-throughput docking results. *J Med Chem*, 47:2743–9.

Kontoyianni M, Sokol GS, McClellan LM. 2005. Evaluation of library ranking efficacy in virtual screening. *J Comput Chem*, 26:11–22.

Lerner MG, Bowman AL, Carlson HA. 2007. Incorporating dynamics in E. coli dihydrofolate reductase enhances structure-based drug discovery. *J Chem Inf Model*, 47:2358–65.

Livingstone D; Devillers J, (Ed). 1996. Neural networks in QSAR and drug design. London: Academic Press, pp. 157–76.

McGovern SL, Shoichet BK. 2003. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J Med Chem*, 46:2895–907.

Nicholls A. 2008. What do we know and when do we know it? *J Comput Aided Mol Des*, 22:239–55.

Nissink JW, Murray C, Hartshorn M, et al. 2002. A new test set for validating predictions of protein-ligand interaction. *Proteins*, 49:457–71.

Oda A, Tsuchida K, Takakura T, et al. 2006. Comparison of consensus scoring strategies for evaluating computational models of protein-ligand complexes. *J Chem Inf Model*, 46:380–91.

Oshiro C, Bradley EK, Eksterowicz J, et al. 2004. Performance of 3D-database molecular docking studies into homology models. *J Med Chem*, 47:764–7.

Plewczynski D, Spieser SA, Koch U. 2006. Assessing different classification methods for virtual screening. *J Chem Inf Model*, 46:1098–106.

Rarey M, Kramer B, Lengauer T, et al. 1996. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol*, 261:470–89.

Sheridan RP. 2008. Alternative global goodness metrics and sensitivity analysis: heuristics to check the robustness of conclusions from studies comparing virtual screening methods. *J Chem Inf Model*, 48:426–33.

Teramoto R, Fukunishi H. 2007. Supervised consensus scoring for docking and virtual screening. *J Chem Inf Model*, 47:526–34.

Vigers GP, Rizzi JP. 2004. Multiple active site corrections for docking and virtual screening. *J Med Chem*, 47:80–9.

Wang J, Cieplak P, Kollman PA. 2000. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J Comput Chem*, 21:1049–74.

Wang R, Wang S. 2001. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J Chem Inf Comput Sci*, 41:1422–6.

Wang R, Lu Y, Wang S. 2003. Comparative evaluation of 11 scoring functions for molecular docking. *J Med Chem*, 46:2287–303.

Warren GL, Andrews CW, Capelli AM, et al. 2006. A critical assessment of docking programs and scoring functions. *J Med Chem*, 49:5912–31.

Yang JM, Chen YF, Shen TW, et al. 2005. Consensus scoring criteria for improving enrichment in virtual screening. *J Chem Inf Model*, 45:1134–46.

Zupan J, Gasteiger J. 1999. Neural networks in chemistry and drug design. Weinheim: Wiley-VCH.