# Improvement of Protein-Compound Docking Scores by Using Amino-Acid Sequence Similarities of Proteins

Yoshifumi Fukunishi*,†,‡ and Haruki Nakamura†,§

Biological Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6, Aomi, Koto-ku, Tokyo 135-0064, Japan, Pharmaceutical Innovation Value Chain, BioGrid Center Kansai, 1-4-2 Shinsenri-Higashimachi, Toyonaka, Osaka 560-0082, Japan, and Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

Received August 16, 2007

The low accuracy of predicted docking scores is critical at in silico drug screening. In order to improve the accuracy of docking scores, we approximated the protein-compound binding free energy as a linear combination of the raw docking scores of a target compound with many different protein pockets. The coefficients of the linear combination were estimated by the similarities among proteins, simply by using the amino-acid sequence similarities or identities of the proteins. This method was applied to in silico screening of the active compounds of five target proteins, and it increased the hit ratio by approximately four to five times compared to that given only by the raw docking scores in every case. The hit ratio also became robust against differences of target proteins.

## 1. INTRODUCTION

One of the most important goals of the genome project is the discovery of new drug targets and the development of new drugs. Comprehensive protein-compound interaction data have been reported by initiatives such as the PubChem project (http://pubchem.ncbi.nlm.nih.gov/); however, the number of known drug target proteins is still only about 300.[1] Many potential drug target proteins remain to be examined. To reduce the cost of drug development, in silico screening has been widely applied to many targets.[2−12]

The problems of in silico screening are the low accuracy (low hit ratio) and the strong target dependence of the hit ratio.[13,14] Many docking programs have been developed,[13−19] and still the accuracy of the binding free energy estimation remains low,[18,19] resulting in low database enrichment from in silico screening. To improve database enrichment, one approach is improvement of the docking score itself.[20,21] But the benefits of this improvement are limited, since this method does not take into account structural changes to the protein, compound, and solvent molecules. No efficient method to reduce the target dependence of the hit ratio is currently known. Some docking programs show high docking accuracies and hit ratios for some targets and low docking accuracies and hit ratios for other targets.[13,14] There is no way of knowing a priori which program is best for a given target.

Another approach to improving the hit ratio is the application of statistical analysis to reduce the computational error. One such method is the application of the protein-compound affinity matrix. The multiple-active-site correction (MASC) scoring method uses the deviation of the docking score instead of the raw docking score.[22] The multiple-target screening (MTS) method and its variations compare the docking scores of many proteins for one compound instead of comparing the docking scores of many compounds for one target protein.[23,24] Details of the MASC scoring and the MTS methods are given in the methods section in this manuscript. If active compounds for the target protein are known, compounds similar to the known active compounds can be found by analyzing the protein-compound affinity matrix.[25−30]

We previously proposed a method in which the docking score of a target protein for a particular compound was modified by a linear combination of the docking scores of the target protein plus several those of various other proteins for the compound[24,30]

$$s_a^{\mathrm{mod}^i} = \sum_b s_b^i M_a^b \tag{1}$$

where $s_a^{\mathrm{mod}^i}$, $s_b^i$, and $M_a^b$ are the modified docking score of the *a*th protein and the *i*th compound, the raw docking score of the *b*th protein and the *i*th compound, and the constant coefficient, respectively. The problem was how to determine the coefficient $M_a^b$ without any experimental observation of the binding free energy. Previously, we calculated $M_a^b$ only from the correlation coefficient between the raw docking score of the target protein and that of the other protein, assuming that the similarity between the local protein pockets was described by the docking scores themselves.

This method achieved a database enrichment factor of ∼23 times when 1% of the compounds of the library were selected,[24] and it improved the hit ratio by about three times on average compared to the raw docking score. However, it could not reduce the target dependence of the hit ratio. Namely, the hit ratio was sometimes worse than that of a random screening, because the hit ratio was too sensitive to

* Corresponding author e-mail: y-fukunishi@aist.go.jp.
† National Institute of Advanced Industrial Science and Technology (AIST).
‡ BioGrid Center Kansai.
§ Osaka University.

10.1021/ci700306s CCC: $40.75 
Published on Web 01/01/2008

the 3D structures of the binding pockets of the target proteins. In the current study, in order to overcome those problems and to develop a method applicable to actual in silico drug screening, we introduced amino-acid sequence similarities to compute the coefficients for the linear combination of raw scores instead of the previous coefficients. Consequently, significantly high hit ratios were attained, and the target dependence of the hit ratios was drastically reduced, providing a much more robust method.

## 2. METHODS

**2.1. Sequence-Based Direct Score Modification Method.** Previously, we proposed the so-called Direct Score Modification (DSM) method,[24] which modified the docking score as in the form of eq 1. Here, the DSM method is briefly described. We prepared a set of proteins and a set of compounds, and the affinity matrix was calculated. Let $s_a^i$ be the raw docking score between the $a$th protein and the $i$th compound. The modified docking score $s_a^{\mathrm{DSM}i}$ is defined as

$$s_a^{\mathrm{DSM}i} = \frac{\sum_b s_b^i R_a^b}{\sum_b R_a^b} \tag{2}$$

where $R_a^b$ is the correlation coefficient between the $a$th and the $b$th proteins

$$R_a^b = \frac{\sum_i \left(s_b^i - \frac{\sum_i s_b^i}{\mathrm{Nc}}\right)\left(s_a^i - \frac{\sum_i s_b^i}{\mathrm{Nc}}\right) + \epsilon}{\sqrt{\sum_i \left(s_b^i - \frac{\sum_i s_b^i}{\mathrm{Nc}}\right)^2 \cdot \sum_i \left(s_a^i - \frac{\sum_i s_b^i}{\mathrm{Nc}}\right)^2} + \epsilon} \tag{3}$$

Here, $\epsilon$ is a small number to avoid the trouble of division by zero when the correlation coefficient is zero, and Nc is the number of compounds.

$R_a^b$ is a similarity measure between proteins $a$ and $b$. In the present study, instead of eq 3, we adopt the amino-acid sequence similarity or identity between proteins $a$ and $b$ as $R_a^b$ as follows

$$R_a^b = S_a^b \frac{1}{1 + e^{-c(x-0.5)}} \tag{4}$$

where $S_a^b$ is the amino-acid sequence similarity or identity between proteins $a$ and $b$, $c$ is a coefficient, and $x$ is the ratio of overlapped sequence, respectively. The sequence alignment program, FASTA, shows the overlapped sequence between proteins $a$ and $b$.[31] The $x$ value is (the length of the overlapped sequence)/(the sequence length of protein $a$). In some cases, FASTA showed high similarity with very short overlapped sequences (less than 10 amino acids). In the present study, since the global similarity with similar fold should be more important than a short sequence motif, we take both the overlapped length and the similarity into account.

The method using sequence similarity is called the sequence-similarity DSM (ssDSM) method, and the method using the sequence identity as $S_a^b$ is called the sequence-identity DSM (siDSM) method. Sequence similarity and identity are measured by the FASTA34 program with the BLOSUM50 similarity matrix.[31,32]

**2.2. Machine-Learning Score Modification (MSM) Method.** If known active compounds are available, we can determine $M_a^b$ in eq 1 to maximize the database enrichment.[24] 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.

$$q = \int_0^{100} f(x)\mathrm{d}x \tag{5}$$

Higher $q$ values correspond to better database enrichment, and $0 < q < 100$. The optimal $M_a^b$ is determined by a Monte Carlo method to maximize the $q$ value. The $a$-$b$ element of the new matrix $\boldsymbol{M}$ ($M_a^{\mathrm{new}b}$) is given by $M^{\mathrm{new}}{}_a{}^b = M_a^b + \eta_a^b$; here, $\eta_a^b$ is a random number and $-1 < \eta_a^b < 1$. Using the newly generated matrix, the new docking score is calculated by eq 1. Then an in silico screening method based on the new matrix $\boldsymbol{M}$ gives the $q$ value of 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.

**2.3. In Silico Screening method with the Combined MTS and MASC Scoring Method.** We combined the MTS method[23] and MASC scoring method[22] as an in silico screening method. The MTS and the MASC scoring methods can select different compounds; thus, the combination of the results of these two methods is taken as the set of candidate hit compounds.[23]

First, let us briefly explain the MTS method. We prepared a set of protein pockets $P = \{p_1,p_2,p_3,...,p_M\}$, where $p_a$ represents the $a$th pocket. The total number of pockets 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 pocket $p_a$, all compounds of set $X$ are docked to pocket $p_a$ with score $s_a^i$ between the $a$th pocket 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 pocket 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 pocket depended on its value $s_a^i$. For example, when $n_a^i = 1$, the $a$th pocket binds the $i$th compound with the strongest affinity. When $n_a^i = \mathrm{M}$, the $a$th pocket binds with the weakest affinity. This procedure was repeated until the order $\{n_a^i;\ a = 1,...,M \mid i = 1,...,N\}$ was determined for all compounds.

Next, we focused on the target $a$th pocket. The compounds having the order $n_a^i = 1$ were assigned as members in the compound group-1, compounds having $n_a^i = 2$ were 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, then the compound with the lowest $s_a^i$ in group-2 should be the most probable hit compound. This procedure is repeated until the most probable hit compound is found.

Second, let us explain the MASC score. The MASC score $s'^i_a$ for the *a*th pocket and the *i*th compound has been reported by Vigers and Rizzi as follows[22]

$$s'^i_a = (s_a^i - \mu_i)/\sigma_i \quad (6)$$

where $s_a^i$ is the raw docking score for the *a*th pocket and the *i*th compound, and $\mu_i$ and $\sigma_i$ are the average and standard deviation of the raw docking scores across all pockets for the *i*th compound, respectively. In this method, $s'^i_a$ is used for screening instead of $s_a^i$.

Both the MTS and the MASC scoring methods are applied in this study, and the combination of the results of these two methods (sum of sets) is taken as the set of candidate hit compounds. Namely, to get the top ranked *N* compounds, about *N*/2 different compounds are taken from the top ranked compounds obtained by the MTS and the MASC scoring methods, respectively, and the sum of the two sets gives the total *N* compounds.

Protein-compound docking simulation was performed by our program named Sievgene,[19] 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.[19] This docking program was developed with a performance yielding about 50% of the reconstructed complexes at a distance of less than 2 Å rmsd for the 132 complexed receptors with the compounds in PDB.[19] The results predicted by our program were almost the same as the results of other docking programs; indeed, we expected that the results obtained by other docking program would show the same trends as the results from our docking program.[33] Our docking program, Sievgene, is a part of the myPresto (prestoX) system, which is available, free for academic use, from the Web site http://www.jbic.or.jp/activity/st_pr_pj/mypresto/index_mypr.html.

## 3. PREPARATION OF MATERIALS

To evaluate our method, we performed a protein-compound docking simulation based on the soluble protein structures registered in the Protein Data Bank (PDB). Since the used protein sets and compound set are exactly the same as those used in our previous study, we can compare the current results to the previous results. Here we describe the data set again.[24]

The protein−ligand complex structures were suitable for the docking study, since the ligand pockets were clearly determined. A total of 180 proteins were selected from the PDB; 142 complexes were selected from the database used in the evaluation of the GOLD[17] and FlexX,[16,33] and the other 38 complexes were selected from the PDB. The former 142-protein data set contains a rich variety of proteins and compounds whose structures have all been determined by high-quality experiments with a resolution of less than 2.5 Å. Almost all the atom coordinates are supplied except the hydrogen atoms, and the atomic structures around the ligand pockets are reliable. Thus, this data set was used in the clustering analysis of proteins and in silico screening. From the original data set, the complexes containing a covalent bond between the protein and ligand were removed, since our docking program cannot perform protein−ligand docking when a covalent bond exists between the protein and the ligand. The other 38 structures included the human immunodeficiency virus protease-1 (HIV protease-1), cyclooxygenase-2 (COX-2), and glutathione S-transferase (GST). The PDB identifiers are summarized in Appendix A. All water molecules and cofactors were removed from the proteins, and all missing hydrogen atoms were added to form all-atom models of the proteins.

Four subsets of proteins were selected from the entire 180 proteins of a clustering method.[23] The clustering method was applied to the 166 proteins other than the 14 target proteins to select candidate proteins. The 180-protein set is called protein set A. The four subsets were named protein sets B, C, D, and E; and these sets consisted of 123, 93, 63, and 24 proteins, respectively. The lists of the PDB codes of the four subsets are summarized in Appendix A.

Our target proteins were the macrophage migration inhibitory factor (MIF, PDB code: 1gcz), COX-2 (1cx2, 1pxx, 3pgh, 4cox, 5cox, and 6cox), HIV (1aid, 1hpx, and 1ivp), thermolysin (2tmn), and GST (18gs, 2gss, and 3pgt).

The compound set for validation tests consisted of 14 inhibitors of MIF, 28 inhibitors of thermolysin, 15 inhibitors of COX-2, 20 inhibitors of HIV, and 12 inhibitors of GST as the active compounds, along with 11,050 potential-negative compounds from the random compound library of the Coelacanth Chemical Corporation (East Windsor, NJ). Typically, only one hit compound can be found out of $10^4$ randomly selected compounds; we therefore expected that there would be no more than a few, if any, hit compounds among these 11 212 compounds.

The list of the active compounds is summarized in Appendix B and the chemical structures of these compounds are depicted in the Supporting Information (labeled Supplementary figure).

The size distribution of compounds is as follows: 0−19 atoms, 0.1%; 20−29 atoms, 1.2%; 30−39 atoms, 1.6%; 40−49 atoms, 9.3%; 50−59 atoms, 22.5%; 60−69 atoms, 37.9%; 70−79 atoms, 20.5%; and more than 80 atoms, 7.0%. The average compound size was 64.3 atoms.

The 3D coordinates of the above 11 050 random compounds were generated by the Concord program (Tripos, St. Louis, MO) from 2D Sybyl SD files provided by the Coelacanth Chemical Corporation. The 3D coordinates of the inhibitors were generated by the Chem3D program (Cambridge Software, Cambridge, MA) from the reference literatures of our previous study.[24] The active compound's set and the decoy compound's set were prepared by using the different programs, but all the compound's structures were energy-minimized with AMBER force field in the docking process of Sievgene.[19] Conformations of the ligands, which were extracted from the protein−ligand complexes, were randomized before the docking study. The atomic charge of each ligand was determined by the Gasteiger method.[34,35] The atomic charges of proteins were the same as the atomic charges of AMBER parm99.[36]

**Table 1.** Average $q$ Values of 14 Target Proteins Using the ssDSM and siDSM Methods, and Their Dependence on the Parameter c

| | ssDSM $c = 0$ | ssDSM $c = 1$ | ssDSM $c = 20$ | siDSM $c = 0$ | siDSM $c = 1$ | siDSM $c = 20$ |
|---|---|---|---|---|---|---|
| protein set E | 68.0 | 70.5 | 69.2 | 70.8 | 72.4 | 71.7 |
| protein set D | 77.8 | 75.9 | 70.0 | 76.0 | 76.0 | 72.5 |
| protein set C | 76.5 | 78.5 | 80.5 | 80.4 | 80.2 | 80.5 |
| protein set B | 80.9 | 83.0 | 83.5 | 84.4 | 84.8 | 84.6 |
| protein set A | 81.0 | 83.4 | 81.2 | 83.5 | 83.3 | 82.3 |

**Table 2.** Average Hit Ratios at 1% Compounds Selected of 14 Target Proteins Using the ssDSM and siDSM Methods, and Their Dependence on the Parameter $c$

| | ssDSM $c = 0$ | ssDSM $c = 1$ | ssDSM $c = 20$ | siDSM $c = 0$ | siDSM $c = 1$ | siDSM $c = 20$ |
|---|---|---|---|---|---|---|
| protein set E | 10.5 | 11.1 | 11.4 | 11.4 | 11.0 | 11.2 |
| protein set D | 16.7 | 20.6 | 14.7 | 17.4 | 21.3 | 21.3 |
| protein set C | 19.9 | 33.4 | 28.4 | 26.5 | 30.3 | 29.9 |
| protein set B | 27.5 | 37.3 | 24.1 | 37.4 | 34.6 | 31.1 |
| protein set A | 35.2 | 45.8 | 36.1 | 38.2 | 41.9 | 39.3 |

## 4. RESULTS

**In Silico Drug Screening Results of the ssDSM and siDSM Methods.** The ssDSM and siDSM methods were applied to the drug screening of the five target proteins—MIF, COX-2, thermolysin, HIV protease-1, and GST—and these results were compared with the results of the original raw docking score and the DSM method. In addition, the scores of the machine-learning score modification (MSM) method, which optimized the coefficients in eq 1 using information of the known compounds,[24] were also compared. Since the MSM method maximizes the $q$ value, which is the surface area under the database enrichment curve,[24] the $q$ value given by the MSM method should be the theoretical upper limit realized by the approximation of eq 1. First, we examined the parameter dependence of the hit ratio of the siDSM/ssDSM methods; next, we observed the details of the results of the siDSM/ssDSM method with an optimal parameter.

Table 1 shows the average $q$ values, which are averages of the 14 $q$ values of the MIF, COX-2, thermolysin, HIV protease-1, and GST, and Table 2 shows the average hit ratios with 1% of compounds selected of these target proteins. The coefficient $c$ of eq 4 was changed from 0 to 20. When $c = 0$, $R_a^b = S_a^b$. The $q$ values did not depend on the parameter $c$ so much; on the contrary, the hit ratio strongly depended on the parameter $c$. In many cases, the hit ratio became the highest value when $c = 1$. Thus the parameter $c$ was set to 1 in the following study.

Figures 1 and 2 show the average database enrichment curves, which are averages of the 14 database enrichment curves of the target proteins. Also, the $q$ values of the 14 target proteins for set A are summarized in Table 3. The results for sets B, C, D, and E are summarized in the Supporting Information.

When 180 proteins (protein set A) were used, the database enrichment was drastically improved by the siDSM and ssDSM methods compared to the results of the raw docking score. The important part of the database enrichment curve is the slope around the origin of the axis, since the purpose of the in silico screening is to select a small number of compounds from the large number of compounds of the library. The slopes of the siDSM and ssDSM methods were much better than that of the raw score and the DSM method. 41.9%, 45.8%, 22.3%, and 8.9% of the active compounds were found within the first 1% of the database of the siDSM, ssDSM, DSM, and the raw score methods, respectively. The enrichment factors of the siDSM and ssDSM methods were 4.7 times and 5.4 times higher than that of the raw score method. The average $q$ values of the siDSM and ssDSM methods were 83.3 and 83.4, respectively, which were much better than the random screening. The hit ratios and the $q$ values of the siDSM and ssDSM methods were almost equivalent to the values of the MSM method reported previously. These results show that the hit ratios of the siDSM and ssDSM methods almost reached the theoretical upper limit of the approximation of eq 2.

The standard deviations of $q$ values ($\sigma_q$) of the siDSM and ssDSM methods were smaller than that of the DSM method and were almost equal to that of the MSM method. The sequence-based DSM, especially the siDSM method, is

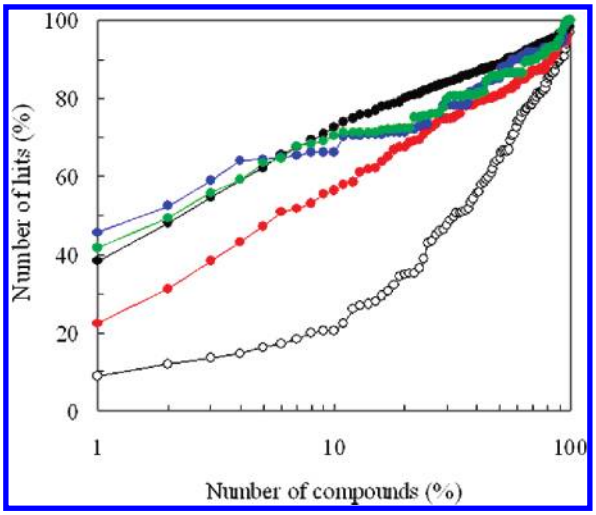


**Figure 1.** Averaged database enrichment curves of 14 proteins using an affinity matrix of 180 proteins (protein set A). Open circles, filled circles, red circles, blue circles, and green circles represent the averaged database enrichments by the raw docking score, the MSM method, the docking score method modified by the DSM method, the ssDSM method, and the siDSM method, respectively.

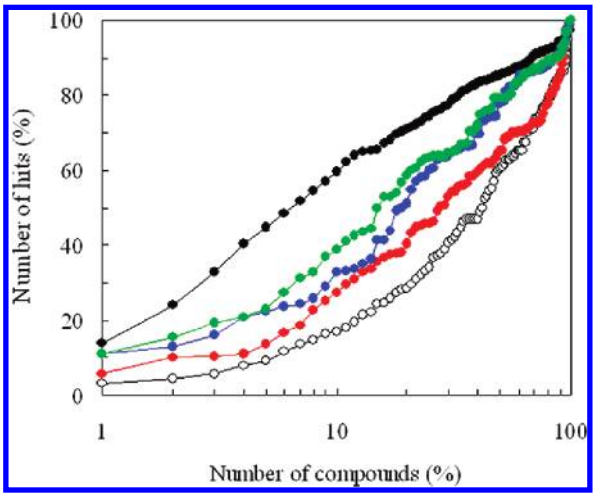


**Figure 2.** Averaged database enrichment curves of 14 proteins using an affinity matrix of 24 proteins (protein set E). Open circles, filled circles, red circles, blue circles, and green circles represent the averaged database enrichments by the raw docking score, the MSM method, the docking score modified by the DSM method, the ssDSM method, and the siDSM method, respectively.

**Table 3.** Database Enrichments of 14 Target Proteins for Protein Set A and Their Average Using the Raw Docking Score, Docking Score Modified by the DSM Method, and the MSM Method, Respectively, and Their Dependence on the Number of Proteins

| | *q*-value | | | | |
|---|---|---|---|---|---|
| PDB | raw[a] | DSM[b] | ssDSM[c] | siDSM[d] | MSM[e] |
| 18gs | 54.8 | 67.7 | 67.2 | 67.3 | 71.4 |
| 1aid | 63.2 | 90.9 | 96.3 | 97.1 | 93.3 |
| 1cx2 | 59.7 | 88.2 | 89.6 | 88.9 | 93.3 |
| 1hpx | 58.6 | 96.1 | 95.7 | 96.8 | 95.7 |
| 1ivp | 65.7 | 95.1 | 96.6 | 96.5 | 94.7 |
| 1gc2z | 53.2 | 93.9 | 91.4 | 90.2 | 85.4 |
| 1pxx | 73.6 | 87.9 | 89.8 | 88.3 | 90.2 |
| 2gss | 68.1 | 57.0 | 67.2 | 66.8 | 61.4 |
| 2tmn | 51.2 | 87.2 | 49.6 | 51.2 | 89.2 |
| 3pgh | 60.5 | 84.1 | 89.6 | 88.9 | 88.6 |
| 3pgt | 72.3 | 63.4 | 66.1 | 66.8 | 68.7 |
| 4cox | 46.4 | 88.9 | 89.6 | 88.9 | 89.9 |
| 5cox | 61.6 | 28.8 | 89.6 | 88.9 | 81.7 |
| 6cox | 68.2 | 69.5 | 89.6 | 88.9 | 87.7 |
| average | 61.2 | 78.5 | 83.4 | 83.3 | 86.3 |
| $\sigma_q$ | 7.7 | 18.4 | 14.0 | 13.7 | 10.3 |
| hit ratio at 1% | 8.9 | 22.3 | 45.8 | 41.9 | 38.3 |

[a] *q* value by the raw docking score. [b] *q* value by the DSM method. [c] *q* value by the sequence-similarity DSM method. [d] *q* value by the sequence-identity DSM method. [e] *q* value by the MSM method.

robust against structural change of the target proteins. The robustness of the siDSM method was almost the same as that of the MSM method. These results show that the siDSM and ssDSM methods reduced the target dependence of the hit ratio by almost half compared to the original DSM method.

When 123 proteins (protein set B) were used, the database enrichments of the siDSM and ssDSM methods were drastically improved compared to the results of the original docking score and the DSM method: i.e., the same result as when the 180 proteins were used. The slopes around the origin of the axis of the siDSM and ssDSM methods were much better than that of the raw docking score and were almost the same value as that of the MSM method. Of the active compounds, 34.6%, 37.3%, 23.6%, and 8.5% were found within the first 1% of the database of the siDSM, ssDSM, DSM, and the raw score methods, respectively. The $\sigma_q$ values of the siDSM and ssDSM were smaller than that of the DSM method and were almost equivalent to that of the MSM method.

When 93 proteins (protein set C) and 63 proteins (protein set D) were used, the database enrichments of the siDSM and ssDSM methods were improved compared to the results of the raw docking score and the DSM method. The slopes around the origin of the axis of the siDSM and ssDSM methods were better than that of the raw docking score and the DSM method, as when the 180 and 123 proteins were used. These results were slightly lower than the result of the MSM method. The $\sigma_q$ values of the siDSM and ssDSM methods were smaller than the $\sigma_q$ value of the DSM method and were the almost the same as that of the MSM method.

When 24 proteins (protein set E) were used, the database enrichments of the siDSM and ssDSM methods were better than the results of the raw docking score and the DSM method. Six *q* values out of 14 of the raw docking score were less than 50, and four *q* values of the DSM method were less than 50. These results suggest that these methods were worse than a random screening in these cases. On the contrary, no *q* value of the siDSM and ssDSM methods was less than 50. The $\sigma_q$ values of the siDSM and ssDSM methods were almost the same as that of the MSM method. The siDSM and ssDSM methods showed advantages in both the hit ratio and structural dependence of the hit ratio compared to the raw score and DSM methods.

In every case (i.e., protein sets A−E), the ssDSM and siDSM method increased the hit ratio by 3−5 times compared to the raw score method, and these hit ratios were higher than the hit ratio of the DSM method. In addition, the ssDSM and siDSM methods decreased the $\sigma_q$ by approximately half compared to the DSM method.

In terms of the hit ratio at 1% compounds, the siDSM and ssDSM expanded the applicable targets compared to the raw score and the DSM method. If the hit ratio at 1% compounds is <1%, such a screening method is worse than a random screening. When protein set A was used, 9 targets out of 14 (=64%), 2 targets out of 14 (=14%), 1 target out of 14 (=7%), and 1 target out of 14 (=7%) had hit ratios of <1% of the raw score, DSM, ssDSM, and siDSM methods, respectively. When protein set E was used, 10 targets out of 14 (=71%), 7 targets out of 14 (=50%), 4 targets out of 14 (=29%), and 4 targets out of 14 (=29%) had hit ratios of <1% of the raw score, DSM, ssDSM, and siDSM methods, respectively. If more than 50% targets using a particular screening method have hit ratios of <1%, then we cannot adopt the method. Thus, only the ssDSM and siDSM methods can be applied when protein set E is used.

Compared to the raw score method, the DSM method decreased the *q* values for three targets out of 14 targets (=21%) by 10−30%, and one *q* value (=7%) was <50 when protein set A was used. The DSM method decreased the *q* values for six targets (=43%) by a few percentage points to 20%, and four *q* values (=29%) were <50 when protein set E was used. If the *q* value is <50, then the screening method should not be used, since random screening gives a better hit ratio. The hit ratio of the DSM method shows strong target dependence, and the target dependence becomes serious when the number of proteins becomes small. On the contrary, no *q* value obtained by the siDSM method was <50 when protein sets A and E were used. Compared to the raw score method, the siDSM method decreased the *q* values for two targets (=14%) by only a few percentage points when protein set A was used, and it decreased the *q* values for two targets (=14%) by a few percentage points to 10% when protein set E was used. Obviously, the siDSM method reduced the target dependence of the hit ratio compared to the DSM method, and the target dependence was small even if the number of proteins was small. The ssDSM method showed the same trend as the siDSM method.

## 5. DISCUSSION

Equation 2 shows that the major contribution to the new docking score comes from the docking scores of similar proteins. If the number of proteins is large enough to find a protein similar to the target protein, eq 2 can work effectively to improve the docking score. On the contrary, if the number of proteins is small and there is no protein similar to the target protein, then eq 2 cannot work. Thus, the selection of proteins is important, and the results of the siDSM and

ssDSM methods depend on the number of proteins. The $q$ values of the siDSM and ssDSM methods are almost proportional to the number of proteins. Namely, the $q$ value of the siDSM $= 0.08 \times$ (number of proteins) $+ 71.89$, and the correlation coefficient is 0.89. The $q$ values of the ssDSM $= 0.09 \times$ (number of proteins) $+ 70.02$, and correlation coefficient is 0.94. These results suggest that the $q$ value could be increased until the number of proteins reaches 350.

The $q$ values of 18gs and 2gss of the siDSM and ssDSM methods are the same values. The amino-acid sequence of 18gs is exactly the same as that of 2gss. Also, the $q$ values of 3pgh, 4cox, 5cox, and 6cox of these methods are the same, and the amino-acid sequences of these proteins are exactly the same to each other. If the amino-acid sequence of protein $a$ is equal to that of protein $b$, the coefficients of protein $a$ in eq 2 are exactly the same as those of protein $b$ by definition. Thus, the database enrichment of protein $a$ is equal to that of protein $b$. The same discussion can be applied to the MSM method. But the coefficients of eq 1 were not ideally optimized in the actual calculation. The optimization process of the MSM method is a Monte Carlo method, so that the same protein did not provide the same $q$ values.

A docking score is improved by the docking scores of similar proteins of eq 2. If there is no similar protein to a target protein in the protein set, eq 2 cannot improve the docking score of the target protein. Thus, the user should add some proteins homologous to the target protein or add target proteins with different ligands into the protein set to achieve a higher hit ratio. The target protein structures generated by a molecular dynamics simulation could be added into the protein set, if the experimental structures of homologous proteins are not available. In practical use, this problem was not so serious in our research. We applied our method to several targets (human hematopoietic prostaglandin D synthase (H-PGDS), orotidine 5-monophosphate decarboxylase (OMPDC), TNF-alpha converting enzyme, and etc.) and succeeded to find active compounds for each target.[37,38] In these cases, several 3D structures of each target were available. The number of potential targets was reported as several thousands, and the number of practical potential targets was approximately 1000.[39] Currently, several thousands protein−ligand complex structures were registered in the PDB, the sequence-based DSM could be applied to many targets.

The coefficient $R_a^b$ in eq 4 is not symmetrical ($R_a^b$ is not equal to $R_b^a$), while the $R_a^b$ in eq 3 is symmetrical. The MSM method is a sort of a single layer perceptron, thus the $M_a^b$ in eq 1 should not be symmetric as parameters of a learning machine. Considering the MSM method, symmetry of $R_a^b$ is not required in the sequence-based DSM method. We tried several other symmetrical definitions of $R_a^b$, the screening results did not strongly depend on the difference of definitions. Among these trials, the coefficient $R_a^b$ in eq 4 gave the best hit ratio.

The sequence-based DSM method outperformed the original DSM method, probably because the proteins are more correctly classified by the current DSM method than the original one. In our previous paper, we showed that the protein classification could be possible by using the protein-compound affinity matrix. But the classification results based on the docking affinity contained some error (miss-classification). In our previous work, the protein set included 6 HIV protease-1s, and the whole proteins were classified into 7 clusters.[19] Four HIV protease-1s out of 6 were included in one cluster, which consisted of 21 proteins, and the other two HIV protease-1s were included in the other cluster. On the contrary, such classification error is quite rare by using amino-acid sequence similarity. To compare the protein pockets, the amino-acid sequence similarity should be better than the docking-affinity based method. For this reason, the sequence-based DSM would outperform the original DSM method.

## 6. CONCLUSION

We developed the siDSM and ssDSM methods, which improve docking scores based on the protein-compound affinity matrix and the amino-acid sequences of proteins. The new docking score of a protein is modified by a linear combination of the docking scores of other proteins, and the coefficients of the linear combination are given by the amino-acid sequence similarities among proteins. The siDSM and ssDSM methods increased the hit ratio by approximately four times compared to the raw docking score and 1.5−2 times comparing to the DSM method when 1% of compounds of the used library were selected. The target dependence of the hit ratio of various in silico screening methods has been suggested by previous reports to be a serious problem.[13,14] The siDSM and ssDSM methods reduced the target dependence of the hit ratio. Namely, these methods reduced the deviation of the hit ratios by half compared to the original DSM method. in silico screenings are worse than a random screening for particular target proteins. The siDSM and ssDSM reduced the probability of such worst cases and expanded the applicable targets compared to the raw score and the DSM method.

The result of the ssDSM method is slightly better than that of the siDSM method in many cases. But, in some cases, the siDSM method gave better results than the ssDSM method. The results of the siDSM and ssDSM methods depend on the number of proteins that are used in the protein-compound affinity matrix; the larger the number of proteins, the better the hit ratio. When the number of proteins was more than 123, the database enrichments of the siDSM and ssDSM methods were almost equivalent to the result of the MSM method, which gives the theoretical upper limit of the $q$ value of the approximation of eq 1. The MSM method requires information about known active compounds of the target protein. On the contrary, the siDSM and ssDSM methods do not require such information. Thus, when there is no known active compound for the target, the siDSM/ssDSM methods are recommended.

## APPENDIX A

The selected 180 proteins of protein set A were as follows: 1gcz, 1cx2, 1pxx, 3pgt, 4cox, 5cox, 6cox, 1aid, 1hpx, 1ivp, 2tmn, 18gs, 2gss, 3pgh, 12as, 16gs, 1a28, 1a42, 1a4g, 1a4q, 1abe, 1abf, 1aco, 1ady, 1aer, 1ai5, 1aoe, 1apt, 1apu, 1aqw, 1asz, 1atl, 1aux, 1b58, 1b76, 1b9v, 1bdg, 1bma, 1byb, 1byg, 1c1e, 1c5c, 1c83, 1cbs, 1cbx, 1cdg, 1ckp, 1com, 1coy, 1cps, 1cqe, 1csn, 1cbx, 1cdg, 1ckp, 1com, 1coy, 1cps, 1cqe, 1csn, 1cvu, 1d0l, 1d3h, 1dd7, 1dg5, 1dhf, 1dog, 1dr1, 1ebg, 1eed, 1efv, 1ejn, 1epb, 1epo, 1eqg, 1eqh, 1ets, 1f0r,

1f0s, 1f3d, 1fen, 1fkg, 1fki, 1fl3, 1glg, 1glp, 1gol, 1gtr, 1hck, 1hdc, 1hfc, 1hos, 1hpv, 1hsb, 1hsl, 1htf, 1hyt, 1hfc, 1hos, 1hpv, 1hsb, 1hsl, 1htf, 1hyt, 1ida, 1ivb, 1jap, 1l3f, 1lah, 1lcp, 1ldm, 1lic, 1lna, 1lst, 1mbi, 1mdr, 1gcz, 1mld, 1mmq, 1mmu, 1mrg, 1mts, 1mup, 1nco, 1ngp, 1nis, 1nks, 1okl, 1pbd, 1pdz, 1phd, 1phg, 1poc, 1ppc, 1pph, 1pso, 1pyg, 1qbr, 1qbu, 1qh7, 1qpq, 1rds, 1rne, 1pxx, 1pyg, 1qbr, 1qbu, 1qh7, 1qpq, 1rds, 1rne, 1rnt, 1rob, 1s2a, 1s2c1, 1s2c2, 1ses, 1snc, 1so0, 1srj, 1tlp, 1tmn, 1tng, 1tnh, 1tni, 1tnl, 1tyl, 1xid, 1xie, 1yee, 2aac, 2aad, 2ack, 2ada, 2cht, 2cmd, 2cpp, 2ctc, 2fox, 2gbp, 2gbp, 2ifb, 2pk4, 2qwk, 2tmd, 3cla, 3cpa, 3erd, 3ert, 3hvp, 3r1r, 3tpi, 4est, 4lbd, 4phv, 5abp, 5cpp, 5er1, 6rnt, and 7tim. For 1abe, 1abf, 5abp, and 1htf, two receptor pockets were prepared since these proteins bind two ligands each.

The selected 123 proteins of protein set B were as follows: 1gcz, 1cx2, 1pxx, 3pgt, 4cox, 5cox, 6cox, 1aid, 1hpx, 1ivp, 2tmn, 18gs, 2gss, 3pgh, 1a28, 1a42, 1a4g, 1a4q, 1abf, 1aco, 1ai5, 1aoe, 1aqw, 1atl, 1b58, 1bkc, 1bma, 1bqq, 1byb, 1byg, 1c1e, 1c5c, 1c83, 1cbs, 1cbx, 1cdg, 1ckp, 1cle, 1com, 1coy, 1cps, 1cvu, 1d0l, 1d3h, 1dd7, 1dg5, 1dhf, 1dog, 1ebg, 1ejn, 1epb, 1ets, 1f0r, 1f0s, 1f3d, 1fen, 1fki, 1fl3, 1glp, 1hdc, 1hfc, 1hos, 1hpv, 1hsb, 1hsl, 1htf, 1hyt, 1ivb, 1jap, 1lah, 1lcp, 1ldm, 1lic, 1lna, 1lst, 1mbi, 1mdr, 1mmq, 1mrg, 1mts, 1mup, 1nco, 1ngp, 1nis, 1okl, 1pbd, 1pdz, 1poc, 1ppc, 1pph, 1qbr, 1qpq, 1r55, 1rne, 1rob, 1snc, 1srj, 1tlp, 1tng, 1tnh, 1tni, 1tnl, 1xid, 1xie, 1yee, 2ack, 2ada, 2cht, 2ctc, 2fox, 2gbp, 2ifb, 2pk4, 2qwk, 3cla, 3cpa, 3erd, 3ert, 3tpi, 4aah, 4est, 4lbd, and 4phv.

The selected 93 proteins of protein set C were as follows: 1gcz, 1cx2, 1pxx, 3pgt, 4cox, 5cox, 6cox, 1aid, 1hpx, 1ivp, 2tmn, 18gs, 2gss, 3pgh, 1a28, 1a42, 1aco, 1ai5, 1aoe, 1atl, 1b58, 1bkc, 1bqq, 1byb, 1c5c, 1c83, 1cbs, 1cbx, 1cle, 1com, 1coy, 1cps, 1cvu, 1d3h, 1dd7, 1dg5, 1dhf, 1dog, 1ebg, 1ejn, 1epb, 1ets, 1f3d, 1fen, 1fl3, 1glp, 1hfc, 1hos, 1hsb, 1hsl, 1hyt, 1ivb, 1jap, 1lah, 1lcp, 1ldm, 1lst, 1mbi, 1mdr, 1mmq, 1mrg, 1mup, 1nco, 1ngp, 1nis, 1okl, 1pbd, 1pdz, 1poc, 1ppc, 1qpq, 1r55, 1rne, 1rob, 1snc, 1srj, 1tni, 1tnl, 1xid, 1xie, 1yee, 2ack, 2ada, 2cht, 2ctc, 2fox, 2gbp, 3cpa, 3erd, 3ert, 3tpi, 4aah, and 4lbd.

The selected 63 proteins of protein set D were as follows: 1gcz, 1cx2, 1pxx, 3pgt, 4cox, 5cox, 6cox, 1aid, 1hpx, 1ivp, 2tmn, 18gs, 2gss, 3pgh, 1a28, 1ai5, 1b58, 1bqq, 1c83, 1cbx, 1cdg, 1com, 1coy, 1cvu, 1d3h, 1dog, 1epb, 1fen, 1fki, 1fl3, 1hfc, 1hos, 1jap, 1lcp, 1ldm, 1mbi, 1mdr, 1mld, 1mmq, 1mrg, 1mup, 1ngp, 1okl, 1pbd, 1pdz, 1pso, 1qbu, 1qpq, 1tng, 1xie, 1yee, 2ack, 2ada, 2cmd, 2ctc, 2fox, 2ifb, 2pk4, 3cpa, 3ert, 3tpi, 4aah and 4lbd.

The selected 24 proteins of protein set E were as follows: 1gcz, 1cx2, 1pxx, 3pgt, 4cox, 5cox, 6cox, 1aid, 1hpx, 1ivp, 2tmn, 18gs, 2gss, 3pgh, 1d3h, 1fl3, 1hfc, 1mup, 1ngp, 1pbd, 2ada, 2cmd, 2ctc, and 4aah.

## APPENDIX B

MIF inhibitors used in this study were as follows: **1**: c1(O)cc2c(cc1)c(c(c1cc(c(cc1)O)O)co2)═O. **2**: c1(CSc2ccncc2)cc(c(O)cc1)O. **3**: c1(O)cc2c(cc1)cc(c1ccccc1)c(o2)═O. **4**: c1(O)cc2c(cc1)cc(C(S)═[NH2])c(o2)═O. **5**: c12[C@@H]3[C@H]([C@@H](C(═O)O)Nc1ccc(c2)Br)CC═C3. **6**: c1(O)cc2c(cc1)cc(C(═O)C)c(o2)═O. **7**: COC(═O)C[C@@H]1O[NH]═C(C1)c1ccc(O)cc1. **8**: CCOC(═O)c1cc2ccc(cc2oc1═O)O. **9**: c12c([C@@H](C(═O)O)C[C@H]1c1cccs1)cc(CCc1c3c(ccc1)cccc3)c(c2)NC(C(C)C)═O. **10**: C(c1ccc(Cl)cc1)(c1ccc(Cl)cc1)(c1ccc(Cl)cc1)CC(═O)O. **11**: c1(CCCC2CCCCC2)oc(C(═O)N2CCN(CC2)Cc2ccccc2)c(C)n1. **12**: O\C(═C\c1ccc(cc1)O)C(O)═O. **13**: D-dopachrome. **14:** 5,6-dihydroxyindole-2-carboxylic acid (DHICA).

COX-2 inhibitors used in this study: **15**: Sc-558 (1-phenylsulfonamide-3-trifluoromethyl-5-parabromophenylpyrazole). **16**: diclofenac. **17**: indomethacin. **18**: arachidonic acid. **19**: diflunisal. **20**: etodolac. **21**: ketoprofen. **22**: naproxen. **23**: nimesulide. **24**: prostaglandin H2. **25**: piroxicam. **26**: rofecoxib. **27**: sulindac. **28**: suprofen.

HIV protease-1 inhibitors used in this study were as follows ([The] PDB code in parentheses is the complex structure from which the compound originated.): compound **29**: C1(c2ccc(F)cc2)(SCCS1)CCCN3CCC(c4ccc(Cl)cc4)(O)CC3 (1aid), compound **30**: c1(OCC2N(S(N(C(C(C2O)O)COc3ccccc3)Cc4ccccc4)(═O)═O)Cc5ccccc5)ccccc1 (1ajv), compound **31**: [4r-(4alpha,5alpha,6beta,7beta)]-3,3′-[[tetrahydro-5 6-dihydroxy-2-oxo-4,7-bis(phenylmethyl)-1h-1,3 diazepine-1,3(2h)-diyl] bis(methylene)]bis[N-2 thiazolylbenzamide (1bv7), compound **32**: C(N(Cc1ncccc1)C)(═O)NC(C(═O)NC(C(C(C(NC(═O)C(C(C)C)NC(N(Cc2ncccc2)C)═O)Cc3ccccc3)(O)O)(F)F)Cc4ccccc4)C(C)C (1dif), compound **33**: C(N1C(C(═O)NC(C)(C)C)CSC1)(═O)C(C(NC(═O)C(NC(═O)COc2[c]3[c](cncc3)ccc2)CSC)Cc4ccccc4)O (1hpx), compound **34**: C(═O)(C(NC(═O)C(CC(C)C)N)CCC(═O)N)NC(C(═O)NC(C(═O)O)CO)CCC(═O)O (1hte), compound **35**: C(═O)(C1C(SC(C(C(═O)NCc2ccccc2)NC(═O)Cc3ccccc3)N1)(C)C)NC(Cc4ccccc4)CO (1htf), compound **36**: c12c(cccc1)NC(═N2)CNC(═O)CC(C(NC(═O)C3C(SC(C(C(═O)NCc4ccccc4)NC(═O)Cc5ccccc5)N3)(C)C)Cc6ccccc6)O (1htg), compound **37**: 2-phosphoglycolic acid (1hvi), compound **38**: C1(N(C(C(C(C(N1Cc2c[c]3[c](cc2)cccc3)Cc4ccccc4)O)O)Cc5ccccc5)Cc6c[c]7[c](cc6)cccc7)═O (1hvr), compound **39**: 2-carbonylquinoline - phenylalaninol group -decahydro-1-methylisoquinoline-2-carbonyl - tertiary-butylamino group (1hxb), compound **40**: ritonavir (1hxw), compound **41**: naphthyloxyacetyl - cyclohexyl alapsi(Choh−Choh)-Val -2-aminomethyl-pyridine (1ivp), compound **42**: 2-carbonylquinoline - phenylalanylmethane -3-(carboxyamide (2-carboxyamide-2-tertbutylethyl)) penta (1jld), compound **43**: C1(N(C(C(C(C(N1Cc2ccc(cc2)CO)Cc3ccccc3)O)O)Cc4ccccc4)Cc5ccc(cc5)CO)═O (1mes), compound **44**: tertiary-butoxyformic acid - phenylalaninol group - dimethylamine -phenylalaninol group - tertiary-butoxyformic acid (1odw), compound **45**: (5r,6r)-2,4-bis-(4-hydroxy-3-methoxybenzyl)-1,5dibenzyl-3-oxo-6-hydroxy-1,2,4-triazacycloheptane (1pro), compound **46**: C1(C(═C(C═C(O1)C(Cc2ccccc2)CC)O)C(c3 cc(ccc3)NC(═O)CCNC(═O)OC(C)(C)C)C4CC4)═O (2upj), compound **47**: N,N-bis- (2(R)-hydroxy-1 (S)-indanyl-2,6- (R,R) -diphenylmethyl-4-hydroxy-1,7-heptandiamide (4hpv).

GST inhibitors used in this study were (The PDB code in parentheses is the complex structure from which the compound originated.): compound **48**: benzylcysteine - phenylglycine (10gs), compound **49**: glutathione - [2,3-dichloro-4- (2-methylene-1-oxobutyl) phenoxyacetic acid (11gs), compound **50**: S-nonyl-cysteine (12gs), compound **51**: 1-(S-glutathionyl)-2,4-dinitrobenzene (18gs), compound **52**: glutamyl group - S-(4-bromobenzyl)cystine (1aqv), compound **53**: glutamyl group - S-(2,3,6-trinitrophenyl)cysteine (1aqx), compound **54**: S-hexylglutathione (1pgt), compound **55**:

cibacron blue (20gs), compound **56**: chlorambucil (21gs), compound **57**: ethacrynic acid (2gss), compound **58**: (9r,-10r)-9-(S-glutathionyl)-10-hydroxy-9,10 dihydrophenanthrene (2pgt), compound **59**: 2-amino-4- [1- (carboxymethyl-carbamoyl)-2- (9-hydroxy-7 8-dioxo- 7,8,9,10- tetrahydrobenzo [def] chrysen-10 ylsulfanyl) -ethylcarbamoyl]-butyric acid (3pgt).

Thermolysin inhibitors used in this study were as follows (The PDB code in parentheses is the complex structure from which the compound originated.): compounds **60**: aspartic acid, compound **61**: aspartame, compound **62**: phenyl alanine, compound **63**: l-benzylsuccinate (1hyt), compound **64**: phenylalanine phosphinic acid - deamino-methyl-phenylalanine (1os0), compound **65**: (6-methyl-3,4-dihydro-2H-chromen-2-yl) methylphosphonate (1pe5), compound **66**: 2-(4-methylphenoxy) ethylphosphonate - 3-methylbutan-1-amine (1pe7), compound **67**: 2-ethoxyethylphosphonate - 3-methylbutan-1-amine (1pe8), compound **68**: (2-sulfanyl-3-phenylpropanoyl)-Phe-Tyr (1qf0), compound **69**: [2(R,S)-2-sulfanylheptanoyl]-Phe-Ala (1qf1), compound **70**: [(2S)-2-sulfanyl-3-phenylpropanoyl]-Gly-(5-phenylproline) (1qf2), compound **71**: n-(1-(2(R, S)-carboxy-4-phenylbutyl) cyclopentylcarbonyl)- (S)-tryptophan (1thl), compound **72**: (R)-retrothiorphan (1z9g), compound **73**: (S)-thiorphan (1zdp), compound **74**: hydroxamic acid (4tln), compound **75**: phenylalanine phosphinic acid (4tmn), compound **76**: Honh-benzylmalonyl-L-alanylglycine-P-nitroanilide (5tln), compound **77**: Cbz-Gly$^P$-Leu−Leu (Zg$^P$Ll) (5tmn), compound **78**: Cbz-Gly$^P$-(O)-Leu−Leu (Zg$^P$(O)Ll) (6tmn), compound **79**: $CH_2CO(N-OH)$Leu-$OCH_3$ (7tln), compound **80**: benzyloxycarbonyl-D-Ala (1kto), compound **81**: benzyloxycarbonyl-L-Ala (1kl6), compound **82**: benzyloxycarbonyl-D-Thr (1kro), compound **83**: benzyloxycarbonyl-L-Thr (1kj0), compound **84**: benzyloxycarbonyl-D-Asp (1ks7), compound **85**: benzyloxycarbonyl-L-Asp (1kkk), compound **86**: benzyloxycarbonyl-D-Glu (1kr6) and compound **87**: benzyloxycarbonyl-L-Glu (1kjp).

## ACKNOWLEDGMENT

This work was supported by grants from the New Energy and Industrial Technology Development Organization of Japan (NEDO) and the Ministry of Economy, Trade, and Industry (METI) of Japan.

**Supporting Information Available:** MIF, HIV protease-1, GST, and thermolysin inhibitors and MIF, COX-2, HIV protease-1, GST, and thermolysin compounds (labeled Supplementary figure) and database enrichments of 14 target proteins for protein set A. This material is available free of charge via the Internet at http://pubs.acs.org.

## REFERENCES AND NOTES

(1) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How many drug targets are there? *Nat. Rev. Drug Discovery* **2006**, *5*, 993−996.

(2) Orita, M.; Yamamoto, S.; Katayama, N.; Aoki, M.; Takayama, K.; Yamagiwa, Y.; Seki, N.; Suzuki, H.; Kurihara, H.; Sakashita, H.; Takeuchi, M.; Fujita, S.; Yamada, T.; Tanaka, A. Coumarin and chomen-4-one analogues as tautomerase inhibitors of macrophage migration inhibitory factor: discovery and X-ray crystallography. *J. Med. Chem.* **2001**, *44*, 540−547.

(3) Cotesta, S.; Giordanetto, F.; Trosset, J.-Y.; Crivori, P.; Kroemer, R. T.; Stouten, P. F. W.; Vulpetti, A. Virtual screening to enrich a compound collection with CDK2 inhibitors using docking, scoring, and composite scoring models. *Proteins* **2005**, *60*, 629−643.

(4) Schellhammer, I.; Rarey, M. FlexX-Scan: Fast, structure-based virtual screening. *Proteins* **2004**, *57*, 504−517.

(5) Evers, A.; Hessler, G.; Matter, H.; Klabunde, T. Virtual Screening of Biogenic Amine-Binding G-Protein Coupled Receptors: Comparative Evaluation of Protein- and Ligand-Based Virtual Screening Protocols. *J. Med. Chem.* **2005**, *48*, 5448−5465.

(6) Howard, M. H.; Cenizal, T.; Gutteridge, S.; Hanna, W. S.; Tao, Y.; Totrov, M.; Wittenbach, V. A.; Zheng, Y.-J. A Novel Class of Inhibitors of Peptide Deformylase Discovered through High-Throughput Screening and Virtual Ligand Screening. *J. Med. Chem.* **2004**, *47*, 6669−6672.

(7) Godden, J. W.; Stahura, F. L.; Bajorath, J. POT-DMC: A Virtual Screening Method for the Identification of Potent Hits. *J. Med. Chem.* **2004**, *47*, 5608−5611.

(8) Zhao, L.; Brinton, R. D. Structure-Based Virtual Screening for Plant-Based ER-Selective Ligands as Potential Preventative Therapy against Age-Related Neurodegenerative Diseases. *J. Med. Chem.* **2005**, *48*, 3463−3466.

(9) Mestres, J.; Veeneman, G. H. Identification of "Latent Hits" in Compound Screening Collections. *J. Med. Chem.* **2003**, *46*, 3441−3444.

(10) Shacham, S.; Marantz, Y.; Bar-Haim, S.; Kalid, O.; Warshaviak, D.; Avisar, N.; Inbal, B.; Heifetz, A.; Fichman, M.; Topf, M.; Naor, Z.; Noiman, S.; Becker, O. M. PREDICT modeling and in-silico screening for G-protein coupled receptors. *Proteins* **2004**, *57*, 51−86.

(11) Cavasotto, C. N.; Orry, A. J. W.; Abagyan, R. A. Structure-based identification of binding sites, native ligands and potential inhibitors for G-protein coupled receptors. *Proteins* **2003**, *51*, 423−433.

(12) Katada, S.; Hirokawa, T.; Oka, Y.; Suwa, M.; Touhara, K. Structure basis for a broad but selective ligand spectrum of a mouse olfactory receptor: mapping the odorant-binding site. *J. Neurosci.* **2005**, *25*, 1806−1815.

(13) 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.

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

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

(16) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(17) Jones, G.; Willet, 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.

(18) 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.

(19) Fukunishi, Y.; Mikami, Y.; Nakamura, H. Similarities among receptor pockets and among compounds: Analysis and application to in silico ligand screening. *J. Mol. Graphics Modell.* **2005**, *24*, 34−45.

(20) Zhang, C.; Liu, S.; Zhu, Q.; Zhou, Y. A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *J. Med. Chem.* **2005**, *48*, 2325−2335.

(21) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.

(22) Vigers, G. P. A.; Rizzi, J. P. Multiple active site corrections for docking and virtual screening. *J. Med. Chem.* **2004**, *47*, 80−89.

(23) Fukunishi, Y.; Mikami, Y.; Kubota, S.; Nakamura, H. Multiple target screening method for robust and accurate in silico ligand screening. *J. Mol. Graphics Modell.* **2005**, *25*, 61−70.

(24) Fukunishi, Y.; Kubota, S.; Nakamura, H. Noise reduction method for molecular interaction energy: application to in silico drug screening and in silico target protein screening. *J. Chem. Inf. Model.* **2006**, *46*, 2071−2084.

(25) Briem, H.; Kuntz, I. D. Molecular similarity based on DOCK-generated fingerprints. *J. Med. Chem.* **1996**, *39*, 3401−3408.

(26) Lessel, U. F.; Briem, H. Flexsim-X: A method for the detection of molecules with similar biological activity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 246−253.

(27) Briem, H.; Lessel, U. F. In vitro and in silico affinity fingerprints: finding similarities beyond structural classes. *Perspect. Drug Discovery Des.* **2000**, *20*, 231−244.

(28) Fukunishi, Y.; Mikami, Y.; Takedomi, K.; Yamanouchi, M.; Shima, H.; Nakamura, H. Classification of chemical compounds by protein-compound docking for use in designing a focused library. *J. Med. Chem.* **2006**, *49*, 523−533.

(29) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, A.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* **1995**, *2*, 107−118.
(30) Fukunishi, Y;, Hojo, S.; Nakamura, H. An efficient in silico screening method based on the protein-compound affinity matrix and its application to the design of a focused library for cytochrome P450 (CYP) ligands. *J. Chem. Inf. Comput. Sci.* **2006**, *46*, 2610−22.
(31) Pearson, W. R.; Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 2444−2448.
(32) Pearson, W. R. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **1990**, *183*, 63−98.
(33) Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein-ligand interaction. *Proteins* **2002**, *49*, 457−471.
(34) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(35) Gasteiger, J.; Marsili, M. A new model for calculating atomic charges in molecules. *Tetrahedron Lett.* **1978**, 3181−3184.
(36) Case, D. A.; Darden, T. A.; Cheatham, T. E., III.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W.S.; Kollman, P. A. *AMBER 8*; University of California: San Francisco, CA, 2004.
(37) Inoue, T.; Kado, Y.; Tokuoka, K.; Matsumura, H.; Kai, Y.; Mori, Y.; Adachi, H.; Takano, K.; Murakami, S.; Fukunishi, Y.; Nakamura, H.; Kinoshita, T.; Nakanishi, I.; Okuno, Y.; Minakata, S.; Sakata. T. Drug development value chain constructed by collaboration between the SOSHO project and the NPO BIOGRID. In *Portable synchrotron light sources and advanced applications*, Proceedings of the 2nd International Symposium on Portable Synchrotron Light Sources and Advanced Applications, Shiga, Japan, 2007; Yamada, H., Mochizuki-Oda, N., Sasaki, M., Eds.; AIP: NewYork, 2007; pp 85−88.
(38) Fukunishi, Y. Structure-based drug screening and ligand-based drug screening with machine learning, *Comb. Chem. High Throughput Screening* **2007**, accepted.
(39) Russ, A. P.; Lampel, S. The druggable genome: an update. *Drug Discovery Today* **2005**, *10*, 1607−1610.

CI700306S