Noise Reduction Method for Molecular Interaction Energy: Application to in Silico Drug Screening and in Silico Target Protein Screening

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

Biological Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST) and Japan Biological Information Research Center (JBIRC), Japan Biological Informatics Consortium (JBIC), 2-41-6, Aomi, Koto-ku, Tokyo 135-0064, Japan, and Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

Received April 27, 2006

We developed a new method to improve the accuracy of molecular interaction data using a molecular interaction matrix. This method was applied to enhance the database enrichment of in silico drug screening and in silico target protein screening using a protein−compound affinity matrix calculated by a protein−compound docking software. Our assumption was that the protein−compound binding free energy of a compound could be improved by a linear combination of its docking scores with many different proteins. We proposed two approaches to determine the coefficients of the linear combination. The first approach is based on similarity among the proteins, and the second is a machine-learning approach based on the known active compounds. These methods were applied to in silico screening of the active compounds of several target proteins and in silico target protein screening.

1. INTRODUCTION

Recent progress in genome-wide research has provided extensive biological molecular interaction data, namely, protein−protein and protein−DNA interaction data. One of the most important molecular interactions is the protein−compound interaction, which is essential in developing new drugs. In silico screening and high throughput screening experiments have provided extensive protein−compound interaction data,[1−11] but the accuracy of the data remains suboptimal. In particular, the low accuracy of in silico screening is a serious problem.

Many docking programs have been developed,[12−23] and still the accuracy of the binding free energy estimation remains about 2−3 kcal/mol.[16,23] The low accuracy of the binding free energy or docking score causes low database enrichment from in silico screening. To improve database enrichment, one approach is improvement of the docking score itself.[24,25] But the limitation of this improvement is obvious. The free energy is calculated from the partition function, which is based on a structural ensemble of numerous structures at a particular temperature; on the other hand, the docking score is calculated from a single protein−compound complex structure.

The other approach is 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.[26] The multiple-target screening (MTS) method compares the docking scores of many proteins for one compound instead of comparing the docking scores of many compounds for one target protein.[27] Details of the MASC scoring and the MTS methods are given in the methods section in this manuscript. If the active compounds for the target protein are known, similar compounds to the known active compounds can be found by principal component analysis for the protein−compound affinity matrix.[28]

Kauvar et al. [29] proposed to approximate the IC_{50} value of a compound for a target protein by a linear combination of the IC_{50} values of the compound for other proteins

$$\log(\mathrm{IC}_{50}(a,i)) = \sum_{b(b \neq a)} c_b^i \log(\mathrm{IC}_{50}(b,i)) \qquad (1)$$

where $\mathrm{IC}_{50}(a,i)$ is the IC_{50} value of the *a*th protein and the *i*th compound, and c_b^i is constant. The procedure to estimate the IC_{50} value of the target protein using eq 1 is as follows.

First, the target protein and the other N proteins are prepared. For M1 compounds, the IC_{50} values of the M1 compounds vs the target and the N proteins are observed experimentally. Second, the IC_{50} values of the M1 compound vs the target compound are approximated by linear combination of the IC_{50} values of the M1 compounds vs the N proteins using eq 1. The optimal coefficients (c_b^i) of the linear combination are determined. Third, the M2 compounds, which will be examined, are prepared, and the IC_{50} values of the M2 compounds vs the N proteins are observed experimentally. Finally, the IC_{50} values of the M2 compounds vs the target protein are evaluated by the IC_{50} values of the M2 compounds given by the third step and the coefficients of the linear combination given by the second step using eq 1.

This method works well; however, it requires a lot of experimental data. The lesson of this method is that the IC_{50} value or the binding free energy of the compound vs the target protein could be approximated by the IC_{50} values or the binding free energies of the compound vs other proteins.

In this study, we assumed that the binding free energy of the compound vs the target protein is improved by linear

* Corresponding author phone: +81-3-3599-8290; fax: +81-3-3599-8099; e-mail: y-fukunishi@jbirc.aist.go.jp.

† National Institute of Advanced Industrial Science and Technology.

‡ Japan Biological Informatics Consortium.

§ Osaka University.

10.1021/ci060152z CCC: $33.50
Published on Web 07/28/2006

combination of the binding free energies vs the target protein and other proteins

$$s_a^{new^i} = \sum_b s_b^i M_a^b \tag{2}$$

where $s_a^{new^i}$, s_b^i, and M_a^b are the modified docking score of the ath protein and the ith compound, the raw docking score of the bth protein and the ith compound, and the constant coefficient, respectively. The problem is how to determine the coefficient M_a^b without any experimental observation of the binding free energy.

The structural accuracy of the docking software is not so high. The success rate of reproduction of the protein−ligand complex structure within RMSD < 2 Å is almost 50%.[16,23] We assumed that the calculated docking score of the ath protein and the ith compound $s_a^{calc^i}$ is the sum of the true docking score $s_a^{true^i}$ and noise η_a^i

$$s_a^{calc^i} = s_a^{true^i} + \eta_a^i \tag{3}$$

where η_a^i is a random number that satisfies normal distribution and whose average value is zero and deviation is σ. We can expect that similar proteins give similar values of the docking score for a given compound. If $s_a^{calc^i}$ is calculated for similar N proteins, and these N scores are averaged, then the deviation of the averaged value is reduced to $\sigma/N^{1/2}$. Thus, the average value of the binding free energies of the protein by some trial dockings could be more accurate than that by a single trial of docking. Considering these aspects of the docking study, we approximated the coefficients of the linear combination by the correlation coefficients of the proteins. This method does not require any known active compound but requires only the protein−compound affinity matrix calculated by protein−compound docking software. This method is called “direct score modification method” in this study. The approximation was evaluated by in silico drug screening studies, and the database enrichment results of these trials showed that the approximation worked well.

Also, if known active compounds for the target protein were available, we proposed a method to determine the coefficients of eq 2 to maximize the database enrichment. The result by this method can be better than the former method. This method is called “machine-learning score modification method” in this study.

In addition to finding active compounds of target proteins, finding a target protein of a known biologically active compound is also important. Almost all human genes have been found by genome projects, and recent progress in the structural determination of proteins has rapidly increased the number of entries in the Protein Data Bank (PDB). Based on this progress, in silico target protein screening will be possible in the near future. In this study, we proposed in silico target protein screening and showed the results based on a small number of proteins.

2. METHODS

2.1. Direct Score Modification (DSM) Method. We prepared a set of proteins and a set of compounds, and the affinity matrix was calculated. Let s_a^i be a raw docking score between the ath protein and the ith compound. The new modified docking score $s_a^{new^i}$ is defined as

$$s_a^{new^i} = \frac{\sum_b s_b^i R_a^b}{\sum_b R_a^b} \tag{4}$$

where R_a^b is the correlation coefficient between the ath and the bth proteins

$$R_a^b = \frac{\sum_i \left(s_b^i - \frac{\sum_i s_b^i}{\mathrm{Nc}} \right)\left(s_a^i - \frac{\sum_i s_b^i}{\mathrm{Nc}} \right) + \epsilon}{\sqrt{\sum_i \left(s_b^i - \frac{\sum_i s_b^i}{\mathrm{Nc}} \right)^2 \cdot \sum_i \left(s_a^i - \frac{\sum_i s_b^i}{\mathrm{Nc}} \right)^2} + \epsilon} \tag{5}$$

Here, ϵ 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. In this study, ϵ is set as 0.001.

Also when the docking score consists of two terms, namely the hydrophilic term s(hp) and hydrophobic term s(hh)

$$s_a^i = s(\mathrm{hp})_a^i + s(\mathrm{hh})_a^i \tag{6}$$

then the new docking score is defined as

$$s_a^{new^i} = \frac{\sum_b s(\mathrm{hp})_b^i R(\mathrm{hp})_a^b}{\sum_b R(\mathrm{hp})_a^b} + \frac{\sum_b s(\mathrm{hh})_b^i R(\mathrm{hh})_a^b}{\sum_b R(\mathrm{hh})_a^b} \tag{7}$$

This method is called the direct score modification (DSM) method. We did not adopt a coefficient of determination (R^2) but a correlation coefficient (R). The reason is that when the docking scores of the bth protein show a negative correlation to the scores of the ath protein, the coefficient R_a^b must be negative.

2.2. Machine-Learning Score Modification (MSM) Method. When known active compounds are available, we can modify the docking score to increase the database enrichment. Suppose the new docking score is given by the linear combination of the docking scores with many proteins; we optimized the coefficients of the linear combination to maximize the database enrichment. Let $s_a^{new^i}$, s_b^i , and M_a^b be the new docking score of the ith compound with the ath protein, the raw docking score of the ith compound with the bth protein, and the constant coefficient, respectively.

$$s_a^{new^i} = \sum_b s_b^i M_a^b \tag{8}$$

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.

$$q = \int_0^{100} f(x)\mathrm{d}x \tag{9}$$

Higher q values correspond to better database enrichment, and $0 < q < 100$. For the random screening, $q = 50$.

The optimization procedure for M_a^b is as follows.

Step 1. The initial matrix M in eq 8 is set as a unit matrix ($M_a^b = \delta_a^b$). The all-new docking scores calculated by eq 8 are equal to the original docking scores. Then screening by the combined MTS and MASC scoring method, which is the in silico screening method described in the next section, gives the q value by eq 9.

Step 2. Many new matrixes M are generated from a seed matrix M using random numbers. In the first step, the seed matrix M is the initial matrix M, which is a unit matrix. The $a-b$ element of the new matrix M (M_a^{newb}) is given by $M_a^{newb} = M_a^b + \eta_a^b$; here, η_a^b is a random number and $-1 < \eta_a^b < 1$. In this study, the number of newly generated matrixes is set at 20.

Step 3. Using the newly generated matrix, the new docking score is calculated by eq 8. Then screening by the combined MTS and MASC scoring method gives the q value by eq 9. The best matrix M, which gives the highest q value, is selected as the seed matrix for step 2.

Steps 2 and 3 are repeated until the q value shows convergence; in this study, the number of cycles was set at 20. This method is called the machine-learning score modification (MSM) method.

2.3. In Silico Screening Method with the Combined MTS and MASC Scoring Method. We combined the multiple target screening (MTS) method[27] and the multiple active site correction (MASC) scoring method[26] as an in silico screening method. The MTS and the MASC scoring methods can select different compounds; thus, the combination of the results by these two methods is taken as the set of candidate hit compounds.[27]

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 ath 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 ith 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 ath pocket and the ith compound. Here, s_a^i corresponds to the binding free energy; a lower s_a^i means a higher affinity between the ath pocket and the ith compound.

For the ith compound, $\{s_a^i;\ a=1,...M\}$ were sorted in descending order, and the order n_a^i was assigned to each ath pocket depending on its value s_a^i. For example, when $n_a^i = 1$, the ath pocket binds the ith compound with the strongest affinity. When $n_a^i = M$, the ath 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 ath 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, 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 ath pocket and the ith compound has been reported by Vigers and Rizzi as follows[26]

$$s'^i_a = (s_a^i - \mu_i)/\sigma_i \tag{10}$$

where s_a^i is the raw docking score for the ath pocket and the ith compound, and μ_i and σ_i are the average and standard deviation of the raw docking scores across all pockets for the ith 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 by these two methods is taken as the set of candidate hit compounds.

Protein−compound docking simulation was performed by our in-house program named Sievgene,[23] 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 the soft interaction forces adapting its slight structural change to some extent.[23] 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.[23] The predicted results by our program was almost the same as the results by the other docking programs, we expected that the results obtained by the other docking program show the same trend as the results obtained by our docking program.[30] Our docking program, sievgene, is a part of the prestoX (myPresto) system, which is available from the Web site http://www.jbic.or.jp/activity/st_pr_pj/mypresto/index_mypr.html, and it is free for academic use.

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). Here, 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 and FlexX,[30] and the other 38 complexes were selected from the PDB. The former 142-proteins 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 include 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 the all-atom models of the proteins. Our target proteins are the macrophage migration inhibitory factor (MIF, PDB code:

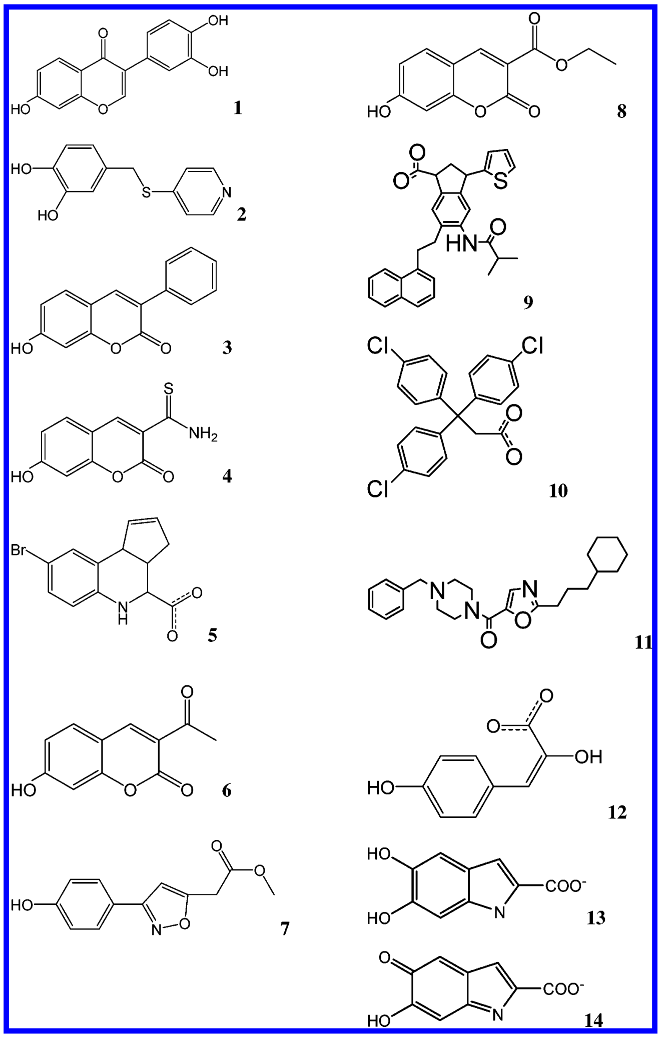

Figure 1. MIF active compounds.

1gcz), COX-2 (1cx2, 1pxx, 3pgh, 4cox, 5cox, and 6cox), HIV protease-1 (1aid, 1hpx, and 1ivp), thermolysin (2tmn), and GST (18gs, 2gss, and 3pgt).

Four subsets of proteins were selected from the entire 180 proteins by 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.

The compound set consisted of 14 inhibitors of MIF, 28 inhibitors of thermolysin, 14 inhibitors of COX-2, 19 inhibitors of HIV protease-1, 12 active inhibitors of GST, and 11 050 potential-negative compounds of the Coelacanth chemical compound library (Coelacanth Corporation, East Windsor, NJ), which is a random library. Usually only one hit compound is found out of 10^4 randomly selected compounds; thus, we expected that there was no or only a few hit compounds among these 11 138 compounds. The active compounds of MIF are depicted in Figure 1 and the other 73 active compounds are listed in Appendix B, and also the active compounds of COX-2, HIV protease-1, GST, and thermolysin are depicted in Figures 2−5, respectively. In Figure 1, compounds **7** and **12** were selected from the PDB, and compounds **1**, **3**, **4**, **6**, and **8** had been reported in a previous study.[1] The others (compounds **2**, **5**, **9**, **10**, and **11**) were prepared in our previous study.[28] Compound **13**

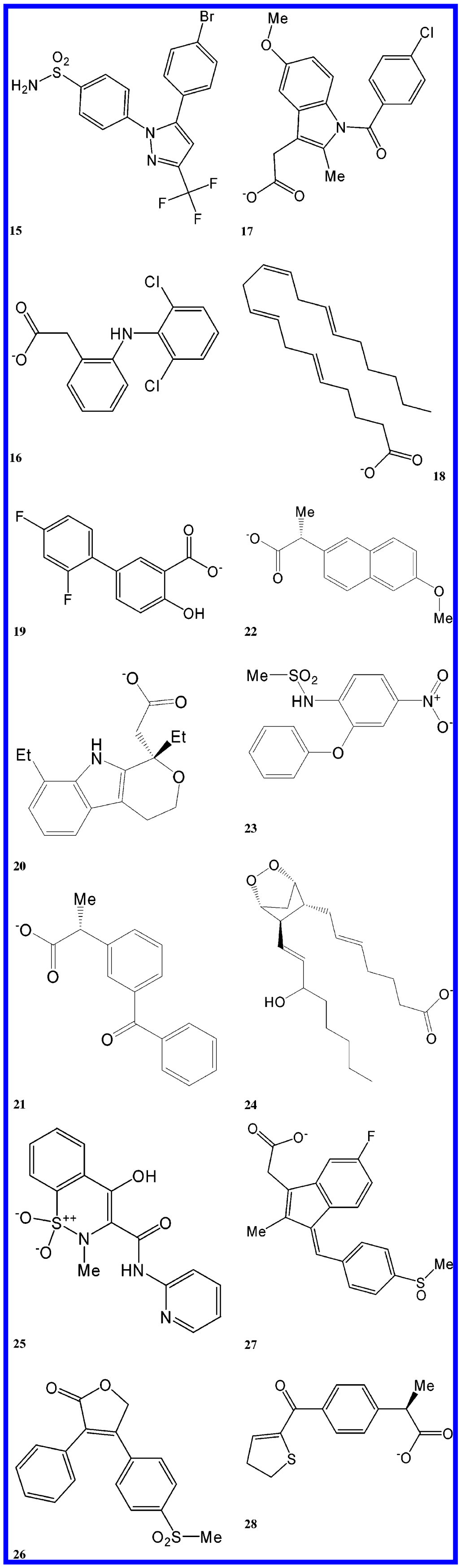

Figure 2. COX-2 active compounds. **15**: Sc-558 (1-phenylsulfonamide-3-trifluoromethyl-5-parabromophenylpyrazole). **16**: diclofenac. **17**: indomethacin. **18**: arachidonic acid. **19**: diflunisal. **20**: etodolac. **21**: ketoprofen. **22**: naproxen. **23**: nimesulide. **24**: prostaglandin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

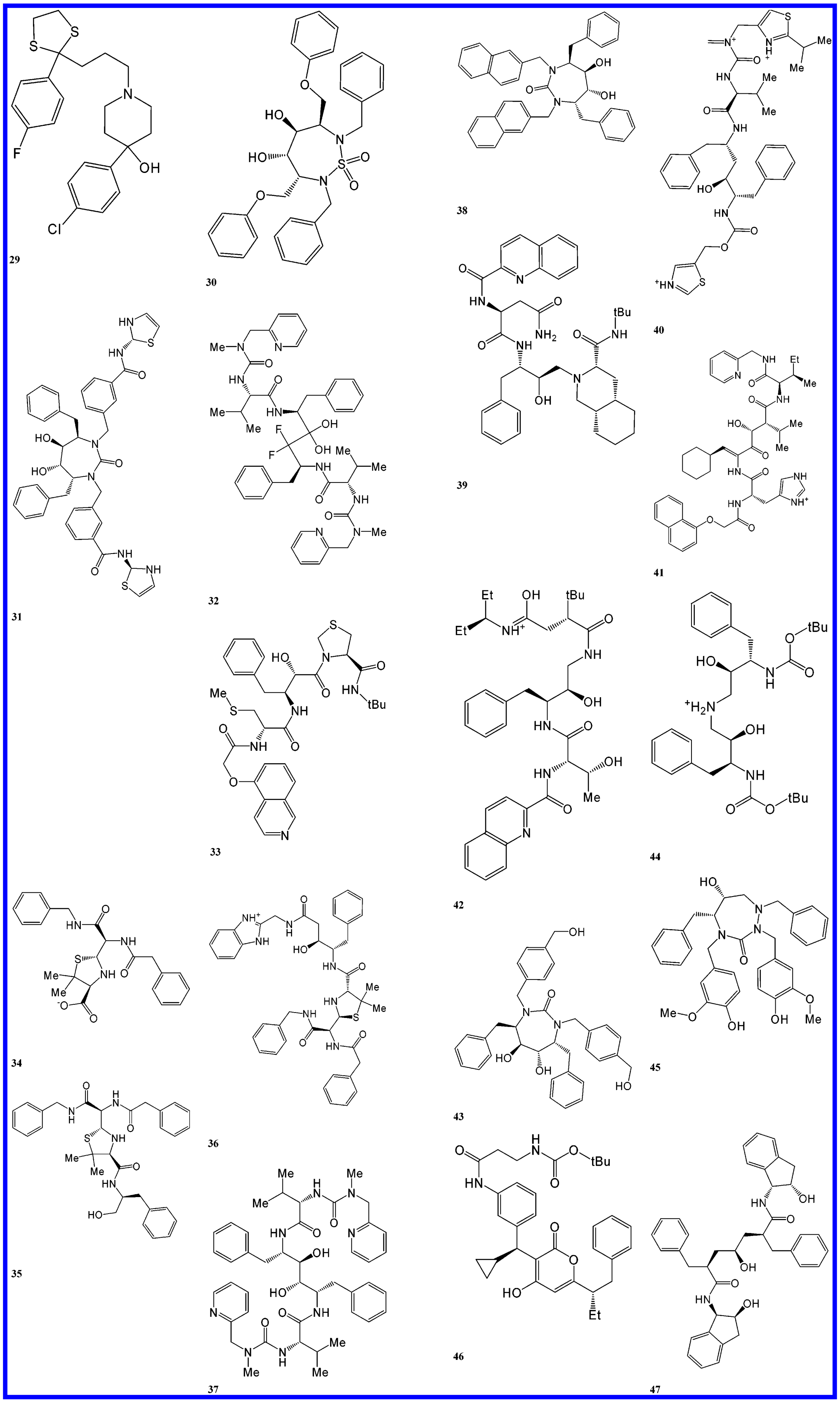

Figure 3. HIV protease-1 active compounds. **29**: ligand of 1aid. **30**: ligand of 1ajv. **31**: ligand of 1bv7. **32**: ligand of 1dif. **33**: ligand of 1hpx. **34**: ligand of 1hte. **35**: ligand of 1htf. **36**: ligand of 1htg. **37**: ligand of 1hvi. **38**: ligand of 1hvr. **39**: ligand of 1hxb. **40**: ligand of 1hxw. **41**: ligand of 1ivp. **42**: ligand of 1jld. **43**: ligand of 1mes. **44**: ligand of 1odw. **45**: ligand of 1pro. **46**: ligand of 2upj. **47**: ligand of 4phv.

and **14** are D-dopachrome and 5,6-dihydroxyindole-2-carboxylic acid (DHICA), which are native ligands of MIF, respectively.[1]

The size distribution of compounds is as follows: ratio of 0−19 atoms, 0.1%; ratio of 20−29 atoms, 1.2%; ratio of 30−39 atoms, 1.6%; ratio of 40−49 atoms, 9.3%; ratio of

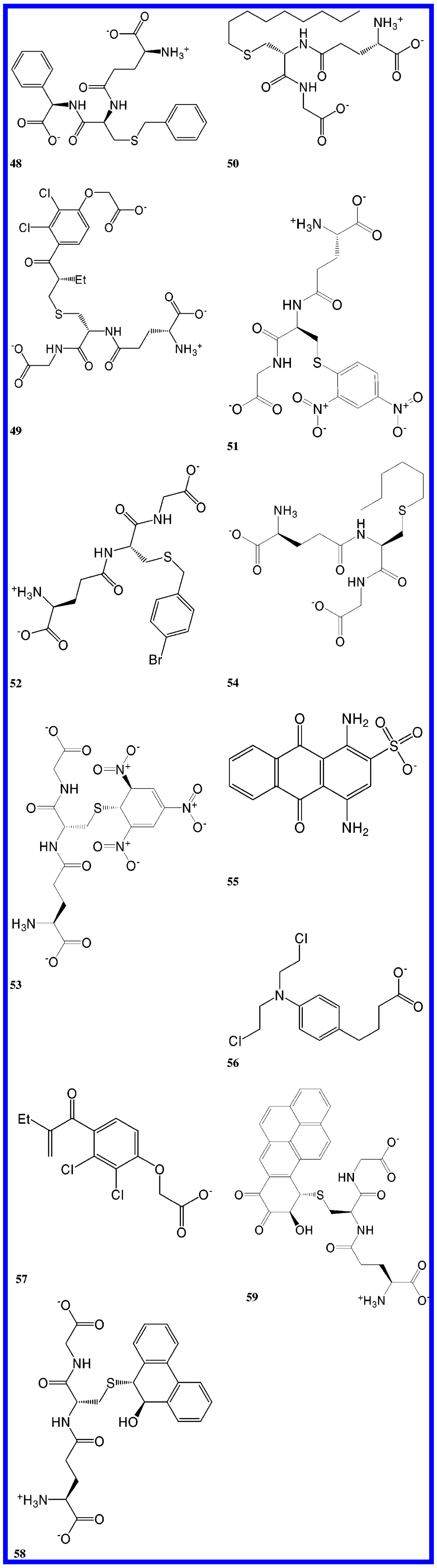

Figure 4. GST active compounds. **48**: ligand of 10gs. **49**: ligand of 11gs. **50**: ligand of 12gs. **51**: ligand of 18gs. **52**: ligand of 1aqv. **53**: ligand of 1aqx. **54**: ligand of 1pgt. **55**: ligand of 20gs. **56**: ligand of 21gs. **57**: ligand of 2gss. **58**: ligand of 2pgt. **59**: ligand of 3pgt.

50−59 atoms, 22.5%; ratio of 60−69 atoms, 37.9%; ratio of 70−79 atoms, 20.5%; and ratio of 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 the 2D Sybyl SD files provided by the Coelacanth Chemical Corporation. The 3D coordinates of the inhibitors were generated by ChemBats3D (Cambridge Software, Cambridge, MA). The 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.[31,32] The atomic charges of proteins were the same as the atomic charges of AMBER parm99.[33]

For in silico target protein screening, we prepared 132 protein−ligand complexes used in our previous study. These complexes are included in protein set A. From these protein−ligand complexes, 132 ligands and 132 target proteins were extracted. The PDB identifiers are summarized in the appendix as protein set F. All water molecules and cofactors were removed from the proteins, and all missing hydrogen atoms were added to form the all-atom models of the proteins. The conformations of all ligands were randomized before the docking simulation.

The size distribution of ligands was as follows: ratio of 1−9 atoms, 3.6%; ratio of 10−19 atoms, 15.2%; ratio of 20−29 atoms, 30.9%; ratio of 30−39 atoms, 15.4%; ratio of 40−49 atoms, 15.8%; ratio of 50−59 atoms, 10.6%; and the ratio of more than 60 atoms was 16.1%. The average ligand size was 37.1 atoms.

The atomic charge of each ligand was determined by the restricted electrostatic point charge (RESP) procedure using HF/6-31G*-level quantum chemical calculations.[34] We used GAMESS and Gaussian98 to perform the quantum chemical calculations.[35,36] The atomic charges of the proteins were the same as the atomic charges in AMBER parm99.[33]

4. RESULTS

4.1. In Silico Drug Screening Results by the DSM Method. The DSM method was applied to the drug screening of the five target proteins: MIF, COX-2, thermolysin, HIV protease-1, and GST. The docking scores were modified by eqs 4 and 5, and then drug screening was performed using the combined method of MTS and MASC scoring.

Figures 6−10 show the average database enrichment curves, which are averages of the 14 database enrichment curves of the MIF, COX-2, thermolysin, HIV protease-1, and GST. Also, the *q* values of the 14 target proteins are summarized in Table 1.

When all 180 proteins (protein set A) were used, the database enrichment was drastically improved compared to the results by 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 slope by the DSM method was much better than that by the raw docking score. 8.9% and 22.27% of the active compounds were found within the first 1% of the database by the raw score and by the DSM method, respectively. The average *q* value by the raw docking score was 61.2, which was better than the random screening; the *q* value by the DSM method reached 78.5. We tried the DSM method using eqs 6 and 7, and the results were almost equivalent to those of the DSM method using

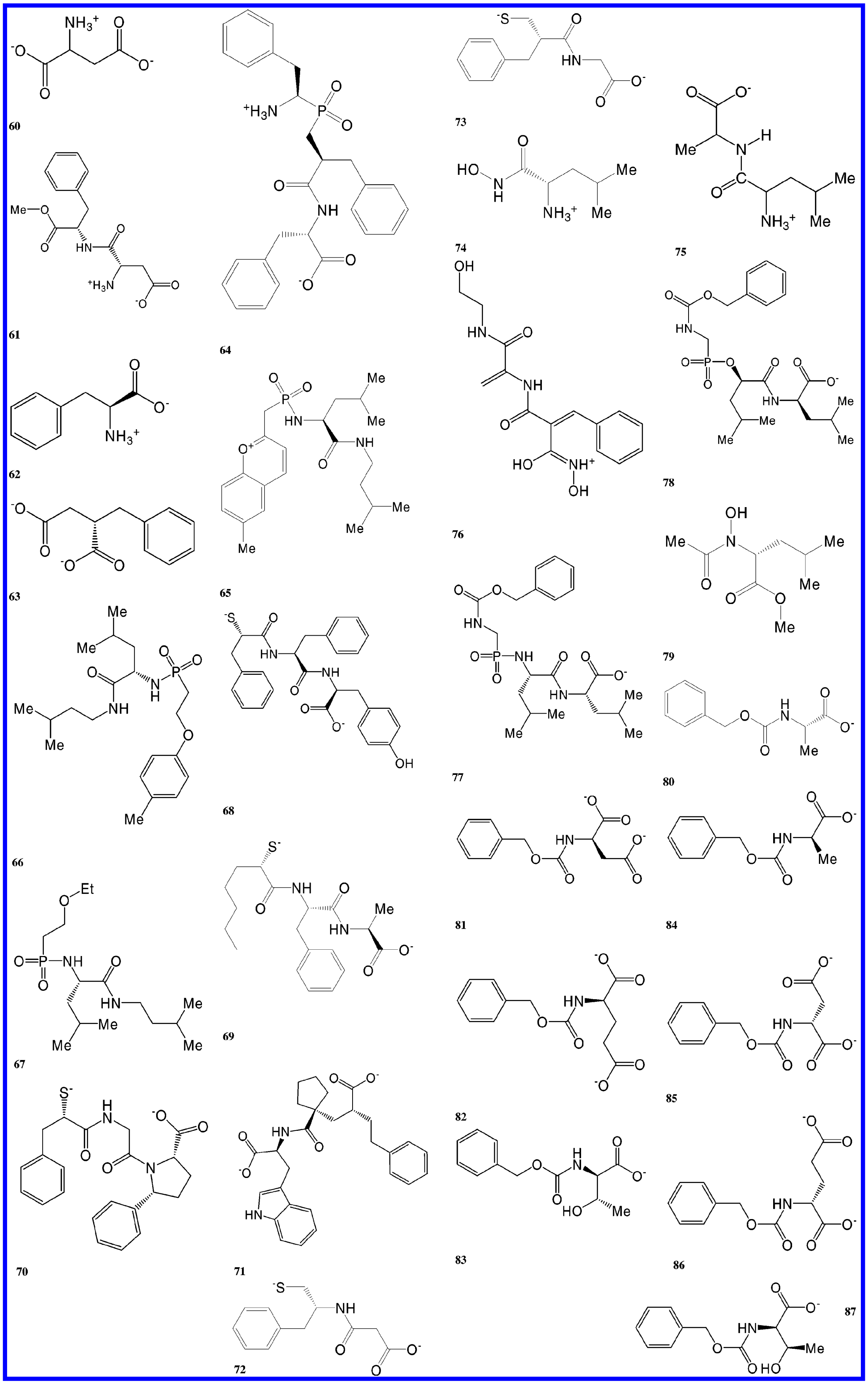

Figure 5. Thermolysin active compounds. **60**: aspartic acid. **61**: aspartame. **62**: phenyl alanine. **63**: ligand of 1hyt. **64**: ligand of 1os0. **65**: ligand of 1pe5. **66**: ligand of 1pe7. **67**: ligand of 1pe8. **68**: ligand of 1qf0. **69**: ligand of 1qf1. **70**: ligand of 1qf2. **71**: ligand of 1thl. **72**: ligand of 1z9g. **73**: ligand of 1zdp. **74**: ligand of 4tln. **75**: ligand of 4tmn. **76**: ligand of 5tln. **77**: ligand of 5tmn. **78**: ligand of 6tmn. **79**: ligand of 7tln. **80**: ligand of benzyloxycarbonyl-D-Ala. **81**: ligand of benzyloxycarbonyl-D-Asp. **82**: ligand of benzyloxycarbonyl-D-Glu. **83**: ligand of benzyloxycarbonyl-D-Thr. **84**: ligand of benzyloxycarbonyl-L-Ala. **85**: ligand of benzyloxycarbonyl-L-Asp. **86**: ligand of benzyloxycarbonyl-L-Glu. **87**: ligand of benzyloxycarbonyl-L-Thr.

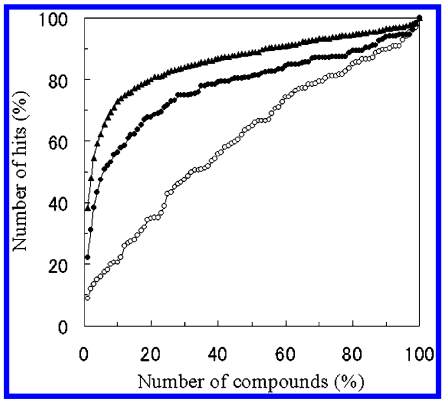

Figure 6. Averaged database enrichment curves of 14 proteins using an affinity matrix of 180 proteins (protein set A). Open circles, filled circles, and filled triangles represent the averaged database enrichments by the raw docking score, the docking score modified by the DSM method, and the docking score modified by the MSM method, respectively.

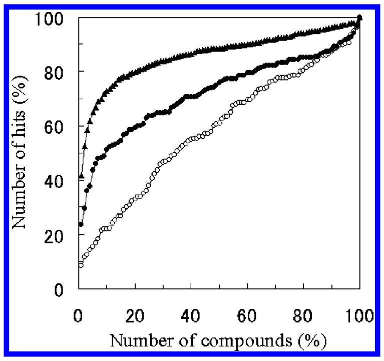

Figure 7. Averaged database enrichment curves of 14 proteins using an affinity matrix of 123 proteins (protein set B). Open circles, filled circles, and filled triangles represent the averaged database enrichments by the raw docking score, the docking score modified by the DSM method, and the docking score modified by the MSM method, respectively.

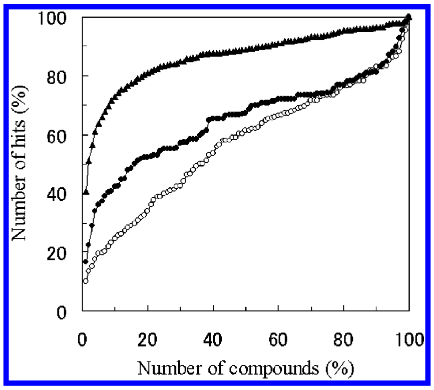

Figure 8. Averaged database enrichment curves of 14 proteins using an affinity matrix of 93 proteins (protein set C). Open circles, filled circles, and filled triangles represent the averaged database enrichments by the raw docking score, the docking score modified by the DSM method, and the docking score modified by the MSM method, respectively.

eqs 4 and 5. Thus, from that point on, only the DSM method using eqs 4 and 5 was applied.

When 123 proteins (protein set B) were used, the database enrichment was drastically improved compared to the results by the original docking score, i.e., the same result as when the 180 proteins were used. The slope around the origin of the axis by the DSM method was much better than that by the raw docking score. 8.48% and 23.63% of the active compounds were found within the first 1% of the database by the raw score and the DSM method, respectively. The average *q* value by the raw docking score was 58.8, which was better than that by the random screening; the *q* value by the DSM method reached 72.3. These values were slightly worse than those from protein set A.

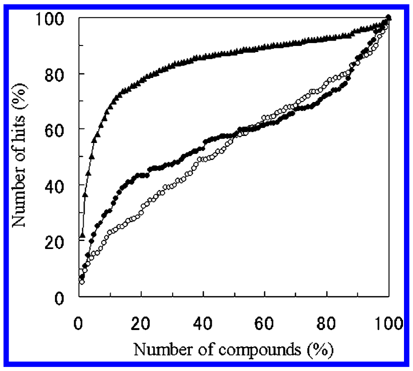

Figure 9. Averaged database enrichment curves of 14 proteins using an affinity matrix of 63 proteins (protein set D). Open circles, filled circles, and filled triangles represent the averaged database enrichments by the raw docking score, the docking score modified by the DSM method, and the docking score modified by the MSM method, respectively.

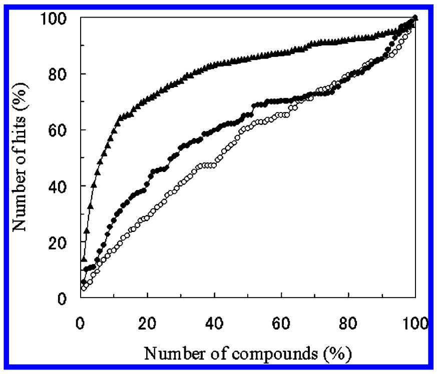

Figure 10. Averaged database enrichment curves of 14 proteins using an affinity matrix of 24 proteins (protein set E). Open circles, filled circles, and filled triangles represent the averaged database enrichments by the raw docking score, the docking score modified by the DSM method, and the docking score modified by the MSM method, respectively.

When 93 proteins (protein set C) were used, the database enrichment was slightly improved compared to the result by the raw docking score. The slope around the origin of the axis by the DSM method was slightly better than that by the raw docking score. 10.27% and 16.59% of the active compounds were found within the first 1% of the database by the raw score and the DSM method, respectively. The average *q* value by the original docking score was 57.0, which was better than that by the random screening; the *q* value by the DSM method reached 65.2.

When 63 proteins (protein set D) or 24 proteins (protein set E) were used, still the database enrichment was slightly improved compared to the result by the raw docking score, but the slope around the origin of the axis by the DSM method was almost the same as that by the raw docking score. 5.00% and 6.90% of the active compounds were found within the first 1% of the database for protein set D by the raw score and the DSM method, respectively, and 3.30% and 5.60% of the active compounds were found within the first 1% of the database for protein set E by the raw score and the DSM method, respectively. For protein set D, the average *q* value by the raw docking score was 54.8, and the

Table 1. Database Enrichments of 14 Target Proteins and Their Average Using the Raw Docking Score, Docking Scores Modified by the DSM and MSM Methods, Respectively, and Their Dependence on the Number of Proteins

				q-value							*q*-value		
type of target	PDB	Ns(0.4)[a]	Ns(0.5)[b]	Raw[c]	DSM[d]	MSM[e]	type of target	PDB	Ns(0.4)[a]	Ns(0.5)[b]	Raw[c]	DSM[d]	MSM[e]
Protein Set A: No. of Proteins 180													
MIF	1gcz	1	57	53.2	93.9	85.4	protease-1	1hpx	8	35	58.6	96.1	95.7
COX-2	1cx2	23	78	59.7	88.2	93.3		1ivp	21	70	65.7	95.1	94.7
	1pxx	1	62	73.6	87.9	90.2	thermolysin	2tmn	12	85	51.2	87.2	89.2
	3pgt	30	67	72.3	63.4	68.7	GST	18gs	35	80	54.8	67.7	71.4
	4cox	1	64	46.4	88.9	89.9		2gss	30	90	68.1	57.0	61.4
	5cox	17	70	61.6	28.8	81.7		3pgh	6	78	60.5	84.1	88.6
	6cox	1	56	68.2	69.5	87.7	average		15.1	69.3	61.2	78.5	86.3
HIV	1aid	26	78	63.2	90.9	93.3							
Protein Set B: No. of Proteins 123													
MIF	1gcz	1	40	46.9	86.9	88.1	protease-1	1hpx	6	35	57.7	95.3	95.2
COX-2	1cx2	14	48	55.4	67.3	90.0		1ivp	12	48	64.8	94.7	95.2
	1pxx	1	41	71.3	88.1	89.0	thermolysin	2tmn	11	55	47.9	83.5	87.2
	3pgt	15	39	72.1	62.3	66.9	GST	18gs	24	48	55.4	64.3	73.4
	4cox	1	41	43.1	78.7	91.1		2gss	21	60	67.5	56.6	61.1
	5cox	10	41	56.7	20.9	69.8		3pgh	5	48	56.3	70.6	89.4
	6cox	1	35	65.1	51.5	84.5	average		10.0	45.0	58.8	72.3	86.2
HIV	1aid	18	46	63.0	91.6	93.9							
Protein Set A: No. of Proteins 93													
MIF	1gcz	1	26	44.5	86.6	88.5	protease-1	1hpx	2	16	63.4	93.7	93.4
COX-2	1cx2	7	32	49.1	39.5	91.4		1ivp	7	32	68.7	93.3	93.4
	1pxx	1	26	69.7	84.5	89.6	thermolysin	2tmn	7	38	45.1	77.8	89.5
	3pgt	7	24	67.8	59.1	70.0	GST	18gs	13	31	65.1	60.4	73.5
	4cox	1	27	41.9	60.2	89.5		2gss	11	40	59.3	56.7	62.6
	5cox	4	25	46.8	19.6	81.5		3pgh	3	31	48.1	48.7	92.4
	6cox	1	21	61.2	39.3	85.6	average		5.3	28.0	57.0	65.2	86.8
HIV	1aid	9	29	67.2	93.9	93.7							
Protein Set A: No. of Proteins 63													
MIF	1gcz	1	16	45.4	95.4	88.5	protease-1	1hpx	2	6	62.5	76.3	92.3
COX-2	1cx2	3	20	45.8	43.2	94.1		1ivp	3	17	67.9	87.1	93.0
	1pxx	1	19	63.8	83.4	85.9	thermolysin	2tmn	4	24	43.5	16.1	81.0
	3pgt	4	18	68.9	57.3	67.8	GST	18gs	8	21	60.2	58.8	65.3
	4cox	1	16	38.0	35.1	89.2		2gss	5	23	64.3	58.5	61.0
	5cox	2	18	43.1	23.0	67.6		3pgh	2	20	41.6	31.1	86.8
	6cox	1	15	55.5	56.8	90.4	average		3.0	18.0	54.8	58.1	84.1
HIV	1aid	5	19	66.2	90.7	93.0							
Protein Set A: No. of Proteins 24													
MIF	1gcz	1	8	42.4	86.5	85.9	protease-1	1hpx	1	2	58.8	57.0	92.4
COX-2	1cx2	1	14	49.7	73.2	92.2		1ivp	1	7	65.2	79.7	88.5
	1pxx	1	12	67.4	83.4	86.5	thermolysin	2tmn	2	14	44.3	30.2	62.4
	3pgt	2	12	70.6	63.3	70.8	GST	18gs	5	14	60.2	61.2	65.1
	4cox	1	13	38.5	44.3	84.3		2gss	3	13	64.1	58.7	58.6
	5cox	1	12	47.7	25.8	77.2		3pgh	1	14	46.3	34.9	84.7
	6cox	1	10	59.0	73.0	89.3	average		1.7	11.0	55.5	61.3	80.6
HIV	1aid	3	13	63.2	87.1	90.2							

[a] Number of similar proteins, whose distance *D* is less than 0.4. [b] Number of similar proteins, whose distance *D* is less than 0.5. [c] *q* value by the raw docking score. [d] *q* value by the DSM method. [e] *q* value by the MSM method.

q value by the DSM method reached 58.1; for protein set E, the average *q* value by the raw docking score was 55.5, and the *q* value by the DSM method reached 61.3. The difference between the results for protein sets D and E was negligible.

Figure 11 shows the correlation between the average *q* value by the DSM method and the average number of proteins that are similar to the target protein (Ns). The similarity between two proteins was evaluated based on the protein–compound affinity matrix.[23] The distance between the *a*th target protein and the *b*th protein ($D(a,b)$) is defined as

$$D(a,b) = \sqrt{\frac{\sum_i (s_a^i - s_b^i)^2}{\mathrm{Nc}}} \qquad (11)$$

where s_a^i and Nc are the docking score between the *a*th protein and the *i*th compound and the number of compounds, respectively. Let Ns(x) be the number of proteins with $D(a,b) < x$. The number of proteins with $D < 0.4$ (Ns(0.4)) and that with $D < 0.5$ (Ns(0.5)) are summarized in Table 1. Figure 11 shows that the average *q* value was proportional to the average Ns; the correlation coefficient between the

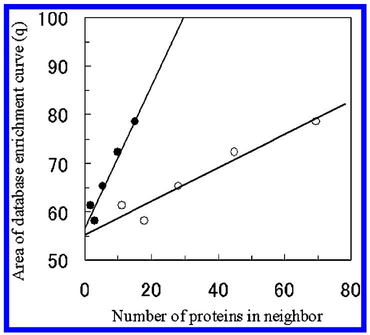

Figure 11. Correlation between the average *q* value by the DSM method and the average number of proteins that are similar to the target protein (Ns). The filled circles and open circles correspond to the results of Ns(0.4) and Ns(0.5), respectively.

average *q* value and Ns(0.4) was 0.97 and that between the average *q* value and Ns(0.5) was 0.96.

4.2. In Silico Drug Screening Results by MSM Method. The MSM method was applied to the drug screening of the five target proteins: MIF, COX-2, thermolysin, HIV protease-1, and GST. The docking scores were modified by eqs 8 and 9, and then the drug screening was performed using the combined MTS and MASC scoring method.

To evaluate the efficiency of this method, the Jack-knife test was applied: namely, the active compounds of each target protein were divided into two sets, the set of known active compounds for machine learning and the set of the hidden active compounds, which should be found by the software. The numbers of the compounds of these two sets were half and half of the whole known active compounds for each target protein. Ten pairs of these active compound sets were prepared for each target protein. Thus, a total of 140 (=14 targets × 10 trials) database enrichment curves were calculated for the 14 target proteins, and the results were averaged.

Figures 6−10 show the average database enrichment curves, which are averages of the 14 database enrichment curves of MIF, COX-2, thermolysin, HIV protease-1, and GST. Also, the *q* values of the 14 target proteins are summarized in Table 1.

When all 180 proteins (protein set A) were used, the database enrichment was drastically improved compared to the result by the raw docking score and the DSM method. The slope by the MSM method was much better than that by the raw docking score and the DSM method. 8.9% and 38.30% of the active compounds were found within the first 1% of the database by the raw score and the MSM method, respectively. The average *q* value by the MSM method reached 86.3, much better than the average *q* value by the raw docking score, 61.2.

When 123 proteins (protein set B), 93 proteins (protein set C), 63 proteins (protein set D), and 24 proteins (protein set E) were used, the trends of the database enrichment did not change so much: namely, the MSM method gave the best result among the three methods: 8.48% and 41.63% of the active compounds were found within the first 1% of the database for protein set B by the raw score and the MSM method, respectively; 10.27% and 40.79% of the active compounds were found within the first 1% of the database for protein set C by the raw score and the MSM method, respectively; 5.00% and 22.12% of the active compounds were found within the first 1% of the database for protein set D by the raw score and the MSM method, respectively; and 3.30% and 13.90% of the active compounds were found within the first 1% of the database for protein set E by the raw score and by MSM method, respectively. The *q* value by the MSM method did not depend on the protein set as much as the other two methods; the average *q* values were 86.3, 86.2, 86.8, 84.1, and 80.6 for protein sets A, B, C, D, and E, respectively.

4.3. In Silico Target Protein Screening Results by the Raw Score and the DSM Method. For in silico target protein screening, the ordinary screening method was applied: namely, for an active compound, the proteins were sorted according to the protein−compound docking score. In this study, we applied the DSM method using eqs 4 and 5 and the other DSM by replacing the suffix of eqs 4 and 5 as follows. Let s_a^i be the raw docking score between the *a*th protein and the *i*th compound. The new modified docking score $s_a^{new^i}$ is defined as

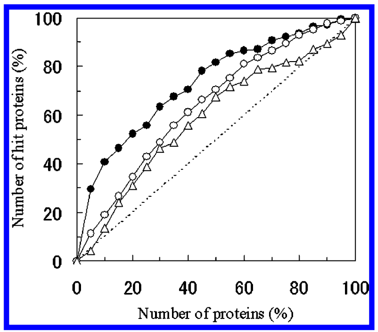

Figure 12. Database enrichment curves of in silico target protein screening. Filled circles, open circles, and open triangles represent database enrichment by the raw docking score, by the DSM method with eqs 12 and 13 and by the DSM method with eqs 4 and 5, respectively. The dashed line represents database enrichment by random screening.

$$s_a^{new^i} = \frac{\sum_j s_a^j R_j^i}{\sum_j R_j^i} \tag{12}$$

where R_j^i is the correlation coefficient between the *i*th and the *j*th compounds

$$R_j^i = \frac{\sum_a \left(s_a^i - \frac{\sum_a s_a^i}{\mathrm{Nr}}\right)\left(s_a^j - \frac{\sum_a s_a^j}{\mathrm{Nr}}\right) + \epsilon}{\sqrt{\sum_a \left(s_a^i - \frac{\sum_a s_a{}^i}{\mathrm{Nr}}\right)^2 \cdot \sum_a \left(s_a^j - \frac{\sum_a s_a{}^j}{\mathrm{Nr}}\right)^2} + \epsilon} \tag{13}$$

Here, ϵ is a small number to avoid the trouble of division by zero when the correlation coefficient is zero, and Nr is the number of proteins. In this study, ϵ is set as 0.001.

Figure 12 shows database enrichment by the raw score and the modified score by the DSM method with eqs 4 and 5 and that with eqs 12 and 13. The in silico target protein screening worked with the raw docking score. The docking score modification did not work at all. 29.6%, 11.3%, and 4.2% of the target proteins were found within the first 5% of the protein database by the raw score, by the DSM method with eqs 12 and 13, and by the DSM method with eqs 4 and 5, respectively.

5. DISCUSSION

The DSM method increased the *q* values for most targets but decreased the *q* values for a few. When 180 proteins were used, the *q* values of three proteins (2gss, 3pgt, and 5cox) out of 14 proteins were decreased. These three proteins showed the same trends even when the number of proteins was changed from 180 to 24. When 123, 93, 63, and 24 proteins were used, the *q* values of 4, 6, 8, and 6 proteins were decreased, respectively. The total number of target proteins was 14; thus, the DSM method was effective when more than 93 proteins were used. When the number of proteins was less than 93, the *q* values decreased in almost half the cases. Figure 11 showed that the average *q* value strongly depends on the number of proteins that are similar to the target protein (Ns). These two results are consistent and suggest that the protein set must include many proteins and that these proteins must be similar to each other to achieve high database enrichment for the DSM method.

Eq 5 shows that the major contribution to the new docking score comes from the docking score of a similar protein, which shows similar docking scores to those of the target protein. If the number of proteins is large enough to find a protein similar to the target protein, eqs 4 and 5 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 eqs 4 and 5 cannot work. Thus, the selection of proteins is important, and the result of the DSM method depends on the number of proteins.

The average *q* value by the DSM method is proportional to the average number of proteins that are similar to the target protein (Ns). However, each *q* value is not proportional to each Ns as shown in Table 1. For protein sets A, B, C, D, and E, the correlation coefficient between each *q* value by the DSM method and each Ns(0.4) was 0.36, 0.18, 0.003, 0.008, and 0.051, respectively. The correlation between the *q* value by the raw score and the DSM method was also weak. For protein sets A, B, C, D, and E, the correlation coefficients between the *q* values by the raw score and by the DSM method were 0.272, 0.108, 0.402, 0.665, and 0.508, respectively. Thus, database enrichment by the DSM method depends on neither the Ns value nor the *q* value by the raw score but rather on the target protein itself.

The average *q* value by the DSM method is proportional to the average *q* value by the raw score, with a correlation coefficient of 0.997 (average *q* value by the DSM = 3.188 × (average *q* value by the raw score) − 116.08), while the correlation between the average *q* value by the MSM method and that by the raw score is relatively weak, with a correlation coefficient of 0.626 (average *q* value by the MSM = 0.619 × (average *q* value by the raw score) + 49.23). Score modification by the DSM method did not reach the theoretical limit so that the *q* value by the DSM method could be improved by increasing the number of proteins. The *q* value by the DSM method was less than that by the MSM method by definition. On the contrary, the score modification by the MSM method reached the theoretical limit, and the *q* value by the MSM method was not improved by increasing the number of proteins. To improve the database enrichment, the docking software itself must be improved.

As shown in Table 1, the screening result depends on the choice of the protein set. The modified score by the DSM and MSM methods depend on the choice of the proteins, and also the screening result by the MASC scoring and the MTS methods depend on the choice of the protein even if the raw docking score was used. If the protein data set consisted of only the target protein structures with different ligands, the accuracy of the docking score could be improved by eqs 4 and 5, but the MTS method does not work well. Because, the candidate hit compound by the MTS method is the compound, which shows the highest docking score with the target protein. If the proteins of the data set were totally different to each other, the DSM method could not work but the MTS method could work. Because, R_a^b value of eq 5 becomes zero, when *a* is not equal to *b*. Thus the choice of the proteins is important; however, it is difficult to determine the best choice of the proteins before the in silico screening.

We used 6 target proteins for COX-2, 3 target proteins for HIV protease-1 and GST, and 1 target protein for MIF and thermolysin. Besides the target proteins, the protein data set includes the proteins, which are similar or exactly the same protein to the target protein, and there are total 7 proteins for COX-2 (1cx2, 1pxx, 3pgt, 4cox, 5cox, 6cox, and 1cvu), 7 proteins for HIV protease-1 (1aid, 1hpx, 1ivp, 1htf, 1hos, 1hpv, and 1qbu), and 2 proteins for thermolysin (2tmn and 1tmn) in the protein set. The number of the same proteins is different for these target proteins, but the trend of the results is similar to each other as shown in Table 1; the *q* values obtained by the MSM method is good, and the *q* values by the DSM and MSM methods is much higher than that by the raw score. Thus, the DSM and MSM methods are expected to work even if the protein data set consisted of the same protein to the target protein or not.

For protein set A, the averaged *q* values by the combined MTS and MASC scoring method were 61.2, 78.5, and 86.3 by the raw scoring, DSM, and MSM methods, respectively. The averaged *q* values by the MTS method were 60.8, 73.9, and 86.0 by the raw scoring, DSM, and MSM methods, respectively. The averaged *q* values by the MASC scoring method were 60.5, 79.5, and 86.1 by the raw scoring, DSM, and MSM methods, respectively. The DSM and MSM methods improve the results by the MASC scoring and MTS methods as well as that by the combined MTS and MASC scoring method. Thus, the DSM and MSM method could be applied to a general in silico screening method other than the combined MTS and MASC scoring method.

As shown in Table 1, the results obtained by the raw score and the results by the DSM method strongly depend on the target structure, and also the results by the MSM method depend on the target structure. Namely, the *q* values of 4cox and 1pxx obtained by the raw score are 46.4 and 73.6, respectively, and the *q* values of 5cox and 4cox obtained by the MSM method are 28.8 and 88.9, respectively. The structure dependence by the MSM is rather smaller than the raw score and the DSM method; the MSM method gives the best *q* value of 93.3 for 1cx2 and the worst *q* value of 68.7 for 3pgt among the five *q* values of COX-2. These proteins are all COX-2, but the screening results are quite different to each other. Structure change due to the different ligands makes the difference. The DSM and MSM methods could improve the screening results in many cases, but these methods could not overcome the problem of induced fitting of the target protein. The protein structure change due to induced fitting is usually larger than the experimental error of the atomic coordinates of the protein. Thus, the choice of the suitable target structure is important for in silico screening rather than the resolution of atomic coordinates of the target protein.

The DSM and MSM methods would reduce the noise of the data of the general interaction matrix, since these methods do not use any particular information about the protein−compound interaction except the docking score. These methods could be applied to other affinity matrixes: namely, the protein−protein interaction matrix and the DNA binding protein−DNA promoter sequence interaction matrix. Also, the matrix element could be evaluated by theoretical model calculation or wet experiment. Thus, these methods could contribute to the improvement of knowledge about protein−protein and protein−DNA interactions given by genome-wide research.

For in silico target protein screening, the MTS method gives a worse result than that by the ordinary screening method, since the MTS method for target-protein screening means a comparison among the docking scores of the compounds for a protein. Usually, a comparison among the docking scores of proteins for a compound is more precise than a comparison among the docking scores of compounds for a protein. The database enrichment of in silico target protein screening became worse by the use of the DSM method than that by the raw docking score. The DSM method with eqs 4 and 5 or eqs 12 and 13 means that this modification is a sort of averaging of the scores. Thus, the DSM method improves the accuracy of the docking score for drug screening but decreases the differences among the scores. The decrease of differences among the scores may decrease the accuracy of the in silico target protein screening.

6. CONCLUSION

We developed DSM and MSM methods, which improve the docking score based on the protein−compound affinity matrix to achieve high database enrichment. The DSM method does not require any information about the active compound. The new docking scores of the target protein are modified by the docking scores of the other proteins. The MSM method requires a set of active compounds of the target protein. The docking scores are optimized to maximize the database enrichment, which is calculated based on the known active compounds.

The result by the DSM method strongly depends on the number of proteins that are used in the protein−compound affinity matrix. In this study, when the number of proteins is more than 93, the database enrichment result by the combined MTS and MASC scoring method can give a better result than that by the raw docking score. When the number of proteins was small, the improvement of the database enrichment by the direct DSM method became quite small. Namely, the result of 63 proteins was almost equivalent to that of 24 proteins.

The MSM method gave the best database enrichment among three methods (MSM, DSM, and raw scoring); thus, the MSM method is recommended when the active compounds are known. The result by the MSM method is not as dependent on the number of proteins, and even a set of 24 proteins could achieve high database enrichment, while a larger set of proteins can give better database enrichment. If any active compound is not available, the DSM method is recommended. To apply the DSM method, the number of proteins must be large, at least 93.

In silico target protein screening was tried to find target proteins for known active compounds. This screening worked with the raw docking score, but the DSM method decreased the database enrichment of this screening. Thus, the database enrichment of in silico target screening remained low compared to in silico drug screening.

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.

APPENDIX A

The selected 180 proteins (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 (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, 11nco, 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 (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 (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 (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.

The selected 132 proteins (protein set F) were as follows: 1a28, 1a42, 1a4g, 1a4q, 1abe, 1abf, 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, 1htf, 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, 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.

APPENDIX B

HIV protease-1 inhibitors used in this study (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)Cc3 ccccc3)N1)(C)C)NC(Cc4ccccc4)CO (1htf), compound **36**: c12c(cccc1)NC(=N2)CNC(=O)CC(C(NC(=O)C3C(SC(C(C(=O)NCc4 ccccc4)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)Cc5 ccccc5)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 ala-psi-(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(Cc2 ccccc2)CC)O)C(c3cc(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 (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-(carboxymethylcarbamoyl)-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 (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).

REFERENCES AND NOTES

(1) 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.

(2) 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.

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

(4) 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.

(5) 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.

(6) 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

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

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

(9) 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.

(10) 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.

(11) 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.

(12) 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.

(13) 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.

(14) 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.

(15) Paul, N.; Rognan, D. ConsDock: A new program for the consensus analysis of protein−ligand interactions. *Proteins* **2002**, *47*, 521−533.

(16) 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.

(17) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian Docking Functions. *Biopolymers* **2003**, *68*, 76−90.

(18) Goodsell, D. S.; Olson, A. J. Automated Docking of Substrates to Proteins by Simulated Annealing. *Proteins* **1990**, *8*, 195−202.

(19) Taylor, J. S.; Burnett, R. M. DARWIN: A program for docking flexible molecules. *Proteins* **2000**, *41*, 173−191.

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

(21) Colman, P. M.Structure-based drug design. *Curr. Opin. Struct. Biol.* **1994**, *4*, 868−874.

(22) Kramer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M. LigScore: a novel scoring function for predicting binding affinities. *J. Mol. Graphics Modell.* **2005**, *23*, 395−407.

(23) 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.

(24) 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.

(25) 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.

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

(27) 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.

(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) 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.

(31) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(32) Gasteiger, J.; Marsili, M. A new model for calculating atomic charges in molecules. *Tetrahedron Lett.* **1978**, 3181−3184.

(33) 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, UCSF 2004.

(34) Wang, J.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* **2000**, *21*, 1049−1074.

(35) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. The general atomic and molecular electronic structure system. *J. Comput. Chem.* **1993**, *14*, 1347−1363.

(36) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A.; Stratmann, R. E. Jr.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98, Revision A.9*; Gaussian, Inc.: Pittsburgh, PA, 1998.

CI060152Z