In Silico Fragment Screening by Replica Generation (FSRG) Method for Fragment-Based Drug Design

Yoshifumi Fukunishi,*[†,‡] Tadaaki Mashimo,[§,||] Masaya Orita,[||,⊥] Kazuki Ohno,[||,⊥] and Haruki Nakamura[†,#]

Biomedicinal 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, Information and Mathematical Science Laboratory Inc., Meikei Building, 1-5-21, Ohtsuka, Bunkyo-ku, Tokyo, 112-0012, Japan, Japan Biological Informatics Consortium (JBIC), 2-41-6, Aomi, Koto-ku, Tokyo 135-0064, Japan, Chemistry Research Laboratories, Drug Discovery Research, Astellas Pharma Inc., 21 Miyukigaoka, Tsukuba, Ibaraki, 305-8585, Japan, and Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

Received December 1, 2008

We developed a new in silico screening method, which is a structure-based virtual fragment screening with protein-compound docking. The structure-based in silico screening of small fragments is known to be difficult due to poor surface complementarity between protein surfaces and small compound (fragment) surfaces. In our method, several side chains were attached to the fragment in question to generate a set of replica molecules of different sizes. This chemical modification enabled us to select potentially active fragments more easily than basing the selection on the original form of the fragment. In addition, the Coulombic and hydrogen bonding interactions were ignored in the docking simulation to reduce the variety of chemical modifications. Namely, we focused on the sizes and the shapes of the side chains and could ignore the atomic charges and types of elements. This procedure was validated in the screenings of inhibitors of six target proteins using known active compounds, and the results revealed that our procedure was effective.

1. INTRODUCTION

Recently, fragment-based drug design (FBDD) has become popular; many successful lead compounds have been developed using this method.[1−7] In FBDD, drug screening is performed for a compound library of small compounds, which are so-called fragments (mass weight < cf. 300 Da); then the subsequent "fragment linking" or "fragment evolution" process generates more active and selective compounds than the original active fragments. A review of research using this method reported that the average mass weight of the active fragments was 270 Da and the average mass weight of the generated lead compounds was 430 Da.[1] The IC_{50} values of these fragments are around 10 mM, and the IC_{50} values of the lead compounds are around 10 nM.

In computer-aided drug design, one of the most difficult steps is the synthesis of the designed compound. The synthetic reaction process of the compound can be predicted by computer software,[8−10] but the actual synthesis is time-consuming and expensive. Many reagent venders provide the building blocks from which many new compounds can be easily generated; however, the hit ratio of the randomly generated compounds is only 0.01%, much lower than the hit ratio by in silico drug screening. Thus, it is advantageous to select the active small compounds by in silico screening from a large variety of fragments or building blocks, before actual chemical synthesis.

In FBDD, the active fragments are selected by experiments. For in silico screening, the fragments are too small to dock into the binding pocket of the target protein. The structure-based in silico screening of small fragments is difficult due to the poor surface complementarity between the protein surface and the narrow compound (fragment) surface. Any compound that perfectly fits the ligand-binding pocket should show strong affinity, detectable by a docking program. If the molecule is larger than the ligand-binding pocket, then the docking program cannot put this molecule into the ligand-binding pocket, allowing the program to easily eliminate this molecule from its pool of possibilities. On the contrary, if all the fragments are much smaller than the pocket, the docking program would put any fragment into the ligand-binding pocket. Such a program could not eliminate any fragments. In addition, the affinities of the fragments are generally too weak to find an active fragment among many nonactive fragments. Many docking programs have been developed,[11−19] and still the accuracy of the binding free energy estimation remains about 2−3 kcal/mol.[14,19−22]

In FBDD, structure-based in silico drug screening based on docking software is used after some active fragments are determined by experiment. in Silico drug screening has been

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

[†] National Institute of Advanced Industrial Science and Technology (AIST).

[‡] BioGrid Center Kansai.

[§] Information and Mathematical Science Laboratory Inc.

[||] Japan Biological Informatics Consortium (JBIC).

[⊥] Astellas Pharma Inc.

[#] Osaka University.

10.1021/ci800435x CCC: $40.75 © 2009 American Chemical Society
Published on Web 04/08/2009

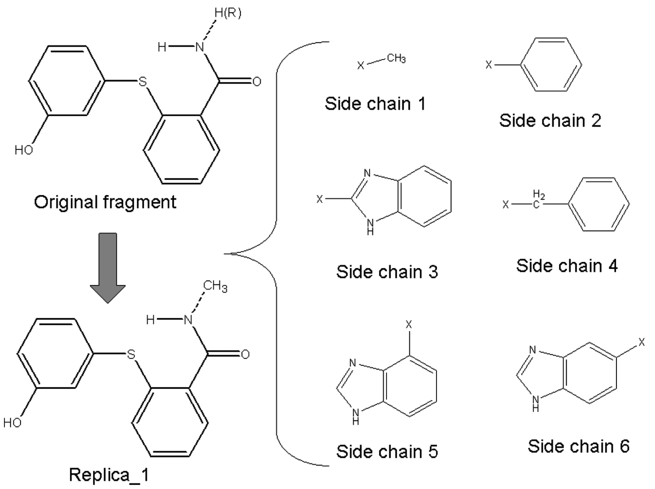

Figure 1. Basic side chains used for the chemical modification in the FSRG method.

applied to computationally generated compounds from the active fragments by a fragment evolution process.[1−7] Since even a small-scale random screening of the fragment library is difficult for us, we planned to select the active fragment by a protein-compound docking simulation. The required enrichment of in silico fragment screening should not be high compared to the usual in silico drug screening. A "hit" or an active fragment is defined as having an IC_{50} value of around 1−10 mM.[1−3] The hit ratio of fragment screening has been reported as about 1%, [1−3] which is much higher than the hit ratio of the usual drug screening, 0.01%. Thus, in silico fragment screening of low enrichment should be useful.

We developed a computer simulation procedure for FBDD based on fragments. We generated a set of larger molecules (replica molecules) than the original fragments by in silico chemical modification and applied the in silico screening to these replica molecules. The efficiency of this procedure was confirmed for six target proteins using the known active compounds.

2. METHOD

2.1. Fragment-Based in Silico Screening by the Replica Generation (FSRG) Method. We developed a new in silico screening method, consisting of a molecular generation step and subsequent in silico drug screening step. The first step was the new compound generation by a sort of fragment evolution. The second step was the in silico drug screening of the newly generated compounds obtained by the first step. Then, candidate hit compounds (fragments) were selected. The details of the procedure are described below.

Step 1. The fragment library is constructed. In the current study, the compound library consists of a set of fragments of known active compounds and a random library instead of the actual fragment database (or building blocks). Each known active compound is divided into two fragments by breaking only one chemical bond close to a heteroatom.

Step 2. A set of new compounds (replica molecules) is generated from the fragment library. For the fragments of the known active compound, the border atom, located at a site where the compound was divided, is replaced by a side chain from the side-chain database. Figure 1 shows these side chains. This process is performed by VCOL, the Virtual COmbinatorial Library generation program.

Step 3. The in silico screening is performed using Sievgene, a protein-compound docking program,[19] followed by the multiple target screening (MTS) method [22,23] based on the compound library generated in the previous step. In the protein-compound docking process, the Coulombic and the hydrogen-bonding interactions were ignored; only the van der Waals interaction and the accessible surface term were taken into account. The database enrichment curves and the hit ratios were calculated to evaluate the procedure.

2.2. Virtual Combinatorial Library Generation Program (VCOL). The VCOL program generates a set of new compounds from two sets of fragments. One of the atoms of a side chain is denoted by the virtual atom "X" and one of the atoms of the molecule of the other side is denoted by

Table 1. Average Number of Atoms, Average Number of Heavy Atoms, and Average Mass Weight (Da) of Decoy Set, Active Compounds, and Fragment Sets for Each Target Protein

		COX2/COX1	ACE	AMPC	FXA	THR
Coelacanth	no. of atoms	63.6	63.6	63.6	63.6	63.6
	no. of heavy atoms	30.9	30.9	30.9	30.9	30.9
	mass weight	423.0	423.0	423.0	423.0	423.0
DUD decoy	no. of atoms	40.2	40.6	30.2	54.5	57.6
	no. of heavy atoms	25.3	23.1	20.8	32.6	32.3
	mass weight	364.1	329.4	304.8	455.7	453.2
original ligand	no. of atoms	35.0	47.8	30.0	55.6	66.1
	no. of heavy atoms	22.0	25.7	20.5	33.4	34.7
	mass weight	316.9	370.7	315.6	465.9	495.4
set I	no. of atoms	18.6	245.0	17.0	30.5	35.5
	no. of heavy atoms	10.8	12.2	11.0	17.8	17.9
	mass weight	156.4	180.3	169.2	250.6	257.2
set II	no. of atoms	20.2	26.2	17.8	33.2	35.8
	no. of heavy atoms	12.4	13.8	12.2	19.8	18.9
	mass weight	175.5	203.0	189.0	277.8	271.1
set III	no. of atoms	33.1	37.5	31.3	42.5	52.7
	no. of heavy atoms	20.6	20.5	18.5	24.9	27.5
	mass weight	290.7	289.9	252.4	338.6	385.0
set IV	no. of atoms	33.4	38.3	31.4	43.7	49.1
	no. of heavy atoms	20.9	20.8	20.2	26.5	28.0
	mass weight	294.2	291.4	289.9	363.5	391.4

the virtual atom “R”. VCOL connects these two fragments by replacing the R atom of the molecule by the side chain (the X atom is removed). The intramolecule atomic conflict is reduced by rotating the chemical bonds of the newly generated compound. In some cases, the protonation state of the compound is changed. The new protonation state is calculated by the program, Hgene. The Hgene program generates the dominant ion form of a compound in pure water. The new atomic charge is calculated by the Gasteiger method using the Hgene program.[24,25] The VCOL and Hgene programs are available from the Web site (http://presto.protein.osaka-u.ac.jp/myPresto4/index_e.html).

2.3. Multiple Target Screening (MTS) Method. We used a structure-based drug screening method based on a protein-compound affinity matrix, called the MTS method.[22,23] This is also a sort of “affinity fingerprint” approach. The basic idea of the MTS method is that potentially active compounds are those compounds that show the strongest affinity with the target protein. Then, the selected compounds are sorted according to their docking scores. Thus, based on the protein-compound affinity matrix, the compounds that show the strongest affinities with the target protein are selected as the hit compounds. The protein set consists of 180 proteins listed in Appendix A, which were also used in our previous study.[23] To perform the docking simulation, the Sievgene/myPresto protein-compound docking program was used.[19] The docking program, the MTS screening tools, and the 3D structures of the used proteins are available on the Web site http://presto.protein.osaka-u.ac.jp/myPresto4/index_e.html.

The MTS method sorts the replicas of the fragments according to the selectivity and the docking scores of the replicas. We want to sort the original fragments instead of the replicas; each fragment is the source of several replicas. The best ranking-order of a replica among the several replicas is adopted as the ranking order of the original fragment. The database enrichment and the hit ratio calculations are based on this reranked list for the fragment.

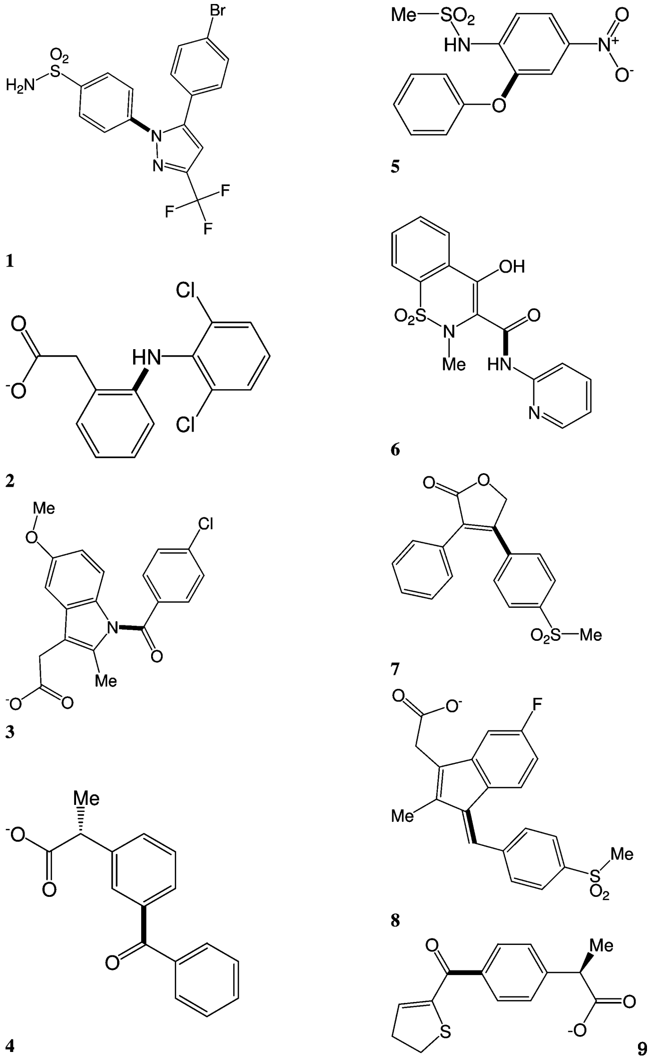

Figure 2. COX2 active compounds. These active compounds were divided into two fragments by breaking the bond shown in boldface. **1**: Sc-558 (1-phenylsulfonamide-3-trifluoromethyl-5-parabromophenylpyrazole). **2**: diclofenac. **3**: indomethacin. **4**: ketoprofen. **5**: nimesulide. **6**: piroxicam. **7**: rofecoxib. **8**: sulindac. **9**: suprofen.

3. RESULTS

3.1. Screening Procedure. The cyclooxygenase-2 (COX2), cyclooxygenase-1 (COX1), angiotensin-converting enzymes (ACE), AmpC beta-lactamase (AMPC), factor Xa (FXA), and thrombin (THR) were selected for the validation test of the FSRG method. Six target protein structures (PDB IDs: 1cx2, 1pxx, 3pgh, 4cox, 5cox, and 6cox) were selected for COX2. Two target protein structures were selected for each of the other five proteins. Namely, 1cqe and 1eqg for COX1, 1uze and 1uzf for ACE, 2pu2 and 2r9x for AMPC, 2w26 and 3ens for FXA, and 2pks and 2zgp for THR were used, respectively. The compound set consisted of inhibitors of a target protein and compounds of a decoy set. The numbers of prepared inhibitors (intact active compounds) for COX2, COX1, ACE, AMPC, FXA, and THR were 9, 9, 13, 10, 10, and 12, respectively.

We prepared 4 fragment sets, I, II, III, and IV, for each target protein. The average number of atoms, average number of heavy atoms, and the average mass weights of both the fragment sets and the original inhibitors are summarized in Table 1. For COX2, sets I and II consisted of 2 × N_{ligand} (where N_{ligand} is the number of original active compounds) fragments, which were obtained by dividing the original 9 active compounds into two different fragments as shown in Figure 2. Each active compound of COX2 was manually divided into two fragments around the amide group, when the compound had an amide group as shown in Figure 3. When the compound did not include an amide group, either a chemical bond close to a heteroatom was broken, or a chemical bond close to middle point of the compound was broken. The fragment sets for COX1 were exactly the same as those for COX2. The fragment sets for the other proteins were prepared in the same way as those for COX2. Most of the active compounds for ACE, AMPC, FXA, and THR included only one amide group; thus, the fragmentation could be uniquely defined. These fragments are listed in the Supporting Information.

First, we generated the set I and set II fragments. For set I fragments, the dangling bond, i.e., the broken chemical bond, was capped by a hydrogen atom, denoted as the R atom. This is the minimal chemical modification. In set II, the dangling bond was capped by an amide group. This amide group mimicked the building block. One of the amide hydrogens was denoted as the R atom. The N−C bonds of

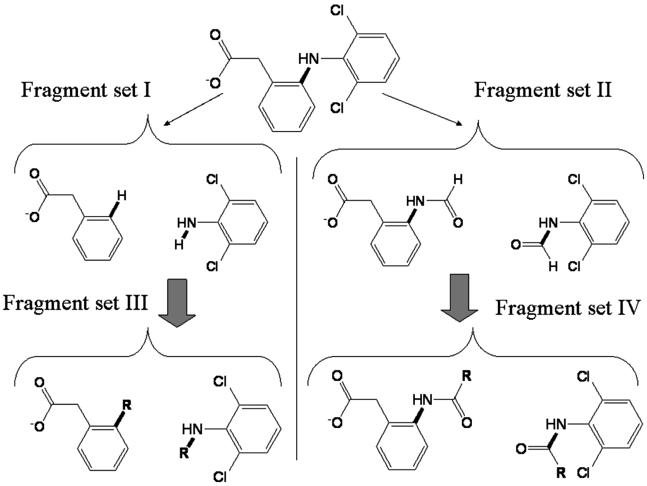

Figure 3. Fragment sets I, II, III, and IV generated from original (intact) active compounds.

the amide group were overlapped onto the thick bonds in Figure 2. The orientation of the amide group was arbitrarily chosen. The sp2 carbon (>C=) of molecule **8** was replaced by an sp3 carbon (>CH-), and the orientation of the additional H was also arbitrarily chosen. The molecular size of set II was 1−2 heavy atoms (about 20 Da in mass weight) larger than that of set I.

Then, we generated sets III and IV composed of 2 × 6 × N_{ligand} fragments, from sets I and II, respectively, as shown in Figure 2. The R atom of each fragment was replaced by the side chain. Figure 1 shows the 6 side chains attached to the original fragments (sets I and II). The molecular size of set IV is 1−2 heavy atoms (about 0−30 Da in mass weight) larger than that of set III.

We used two decoy sets for each target protein. One decoy set was the Coelacanth chemical compound library (Coelacanth Corporation, East Windsor, NJ, USA), which is a random library consisting of 11050 potential-negative compounds. The Coelacanth decoy set was used for all targets. The other decoy set was the decoy set of the directory of useful decoys (DUD) for each target protein.[27] Specific DUD decoy sets were prepared for each target. The decoy set for COX2 was used for COX1. The numbers of compounds in the DUD decoy sets for COX2, ACE, AMPC, FXA, and THR were 13289, 1797, 786, 5745, and 2456, respectively. The average number of atoms, average number of heavy atoms, and the average mass weights of these decoy sets are summarized in Table 1. Usually only one hit compound was found out of 10^4 randomly selected compounds; thus, we expected that there were no hit compounds, or only a few, among these 10^4 compounds.

The side chains shown in Figure 1 were not attached to the compound of the decoy set because the DUD decoy set was designed for the screening of the target protein. If the side chains were attached to the compounds of the DUD decoy set, the newly generated compounds were not suitable for the screening test. Also, the size of the fragment sets is always smaller than that of the decoy sets. When the basic side chains are attached to the compounds of the decoy sets, the difference increases between the fragment set and the decoy set. In the case of COX2, the average number of atoms of compounds of the Coelacanth decoy set was 63.6, and that of the DUD decoy set was 40.2. On the other hand, the average number of atoms of set III was 33.1, and that of set IV was 33.4. The fragments are much smaller than the compounds of the decoy sets.

The 3D coordinates of the 11,050 chemical compounds of the Coelacanth chemical compound library 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 known active compounds were generated by the Chem3D program (Cambridge Software, Cambridge, MA, USA). We used the general AMBER force field (GAFF),[26] and the molecular topology files were generated by tplgeneL/myPresto. The energy optimization of the coordinates of small molecules was performed by Cosgene/myPresto.[28] The atomic charges were calculated by the Gasteiger method of Hgene/myPresto.[24,25] Details about the DUD decoy set were given in an earlier paper.[27]

The protein-compound docking procedure was exactly the same as that reported in our previous work. 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,[29] and the other 38 complexes were selected from the PDB. The former 142-protein data set contained a rich variety of proteins and compounds whose structures had all been determined by high-quality experiments with a resolution of less than 2.5 Å. The coordinates of almost all of the atoms except the hydrogen atoms are supplied, and the atomic structures around the ligand pockets are reliable. The docking pocket of each protein was indicated by the coordinates of the original ligand. The atomic charges of the proteins were the same as the atomic charges of AMBER parm99.[30] For flexible docking, the Sievgene program generated up to 100 conformers for each compound.

3.2. Screening Results without Replicas with Ordinary Scoring Function. Let x and $f(x)$ be the numbers of compounds (%) selected from the total compound library and from the database enrichment curve, respectively. The surface area under the database enrichment curve (q) is a measure of the database enrichment.

$$q = \int_0^{100} f(x)dx \qquad (1)$$

Higher q values correspond to better database enrichment, and $0 < q < 100$. The q value by a random screening is 50. The q value is almost the same as the area under the receiver operating characteristic (ROC) curve (AUC), when the number of active compounds is much smaller than the number of the decoy compounds.

Figure 4a shows the averaged database enrichment curves of the intact active compounds (original active compounds) and the 18 original fragments of sets I and II with the usual docking score for COX2. The database enrichment curves of fragments were very close to those of a random screening, while the database enrichment curve of the intact active compounds showed good enrichment. It clearly reflects how the in silico screening of fragments is difficult. The enrichment curves with the Coelacanth decoy set were better than those with the DUD decoy set. The screening results depend on the compound library used.

The database enrichment curves of fragments with the DUD decoy set for COX2 were worse than those of a random

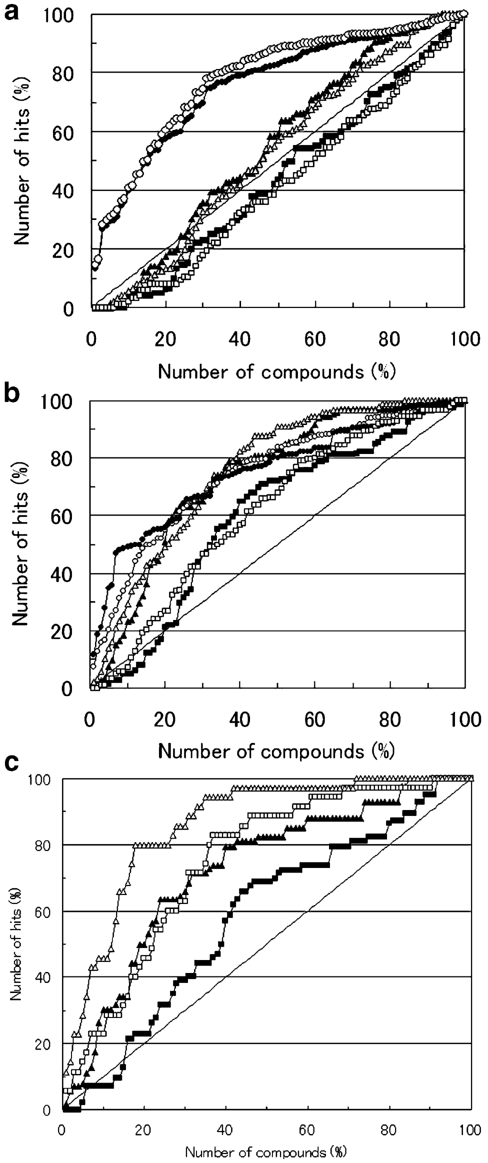

Figure 4. Database enrichment curves of intact active compounds for COX2 and fragments of sets I, II, III, and IV. Filled and open marks represent the database enrichment curves with the Coelacanth decoy set and that with the DUD decoy set. (a) Database enrichment curves of intact active compounds, set I and set II, using the ordinary scoring function. The circles, squares, and triangles represent the averaged database enrichment curves of the intact active compounds, set I molecules and set II molecules, respectively. (b) Database enrichment curves of intact active compounds and set I and set II fragments without the Coulombic and the hydrogen bonding interaction terms. The circles, squares, and triangles correspond to the averaged database enrichment curves of all active compounds, set I fragments and set II molecules, respectively. (c) Database enrichment curves of set III and set IV by the FSRG method. The squares and the triangles are the averaged database enrichment curves of set III and set IV molecules, respectively.

screening. The used DUD decoy set is designed for the COX2 screening test, and the fragments are smaller than the compounds of the DUD decoy set. Thus, the screening of fragments with the DUD decoy set was more difficult than the screening with the Coelacanth decoy set.

The database enrichment curves of set I were slightly worse than those of II. Since the fragments of set II are slightly larger than those of set I, the screening of set II should be slightly easier than the screening of set I.

Table 2. *q* Values of the Intact Active Compounds and Sets I, II, III, and IV Fragments for COX2

Case 1[a]

	intact active compound		set I		set II	
	Coelacanth decoy	DUD decoy	Coelacanth decoy	DUD decoy	Coelacanth decoy	DUD decoy
1cx2	84.8	75.5	53.3	52.9	52.4	50.2
1pxx	71.3	67.7	64.7	48.0	63.8	48.4
3pgh	70.7	70.1	74.9	64.9	69.4	58.4
4cox	64.0	66.7	31.2	14.5	28.4	13.6
5cox	81.6	80.7	54.8	48.7	46.9	39.3
6cox	84.1	83.3	41.7	19.7	44.2	23.3
average	76.1	74.0	53.4	41.4	50.8	38.8

Case 2[b]

	Intact active compound		set I		set II	
	Coelacanth decoy	DUD decoy	Coelacanth decoy	DUD decoy	Coelacanth decoy	DUD decoy
1cx2	81.6	76.7	77.7	70.9	81.2	74.5
1pxx	78.0	60.8	78.5	60.5	72.6	54.3
3pgh	93.0	62.5	91.8	80.4	92.9	82.0
4cox	50.4	68.4	53.3	28.5	63.1	38.5
5cox	79.9	75.9	74.0	62.9	71.6	58.8
6cox	69.4	81.2	58.2	38.4	64.9	44.5
average	75.4	70.9	72.2	56.9	74.4	58.8

Case 3[c] (FSRG)

	set III		set IV	
	Coelacanth decoy	DUD decoy	Coelacanth decoy	DUD decoy
1cx2	70.8	65.8	87.6	81.0
1pxx	76.6	58.1	80.6	65.4
3pgh	77.5	65.7	90.2	77.6
4cox	69.5	47.4	71.0	53.4
5cox	74.9	67.4	88.8	80.4
6cox	70.2	52.4	93.0	83.4
average	73.3	59.5	85.2	73.5

[a] Results with the ordinary scoring function. [b] Results without the Coulomb and the hydrogen bonding interaction terms. [c] Results by the FSRG method.

Tables 2, 3, and 4 show the *q* values obtained by the MTS method. The *q* values of the fragment sets are worse than the *q* values of the intact active compounds (see "case 1" of Tables 2, 3, and 4) for all 16 proteins. Namely, when the Coelacanth decoy set was used, 9 out of 16 intact active compounds had *q* values > 70. Four out of the 16 set I cases and only 1 out of the 16 set II cases had $q > 70$. When the DUD decoy set was used, 5 out of the 16 intact active compounds had $q > 70$. Only 2 out of the 16 set I cases and none of the 16 set II cases had $q > 70$.

3.3. Screening Results without Replicas Ignoring Coulombic and Hydrogen Bonding Interaction Energies. Figure 4b shows the averaged database enrichment curves of the intact active compounds and the 18 original fragments of sets I and II with the docking scores, which do not include the Coulombic or the hydrogen-bonding interactions for COX2. The docking score of the Sievgene program consists of five terms: van der Waals interaction, Coulombic interaction, hydrogen bonding interaction, accessible surface interaction, and an entropy term due to the number of rotational

Table 3. *q* Values of the Intact Active Compounds and Sets I, II, III, and IV Fragments for ACE, AMPC, FXA, and THR

Case 1[a]

	intact active compound		set I		set II	
	Coelacanth decoy	DUD decoy	Coelacanth decoy	DUD decoy	Coelacanth decoy	DUD decoy
1uze	65.6	70.4	62.0	39.7	52.4	31.1
1uzf	70.0	60.0	67.2	32.9	58.3	36.1
2pu2	58.4	42.6	37.1	39.0	48.6	42.1
2r9x	66.2	45.4	41.1	42.3	54.8	45.6
2w26	73.9	66.9	81.5	78.0	74.1	67.6
3ens	66.6	63.0	74.0	74.5	65.6	62.9
2pks	65.6	57.5	62.0	56.6	52.4	38.8
2zgp	70.8	64.4	67.2	63.5	58.3	47.6
average	67.1	58.8	61.5	53.3	58.1	46.5

Case 2[b]

	intact active compound		set I		set II	
	Coelacanth decoy	DUD decoy	Coelacanth decoy	DUD decoy	Coelacanth decoy	DUD decoy
1uze	65.4	55.2	52.2	4.3	47.1	6.7
1uzf	66.7	62.3	54.5	7.2	49.5	7.0
2pu2	21.7	14.9	1.9	3.3	2.9	3.0
2r9x	55.4	41.2	8.6	14.1	14.7	13.7
2w26	69.4	62.7	69.6	73.0	51.6	71.4
3ens	60.1	56.6	59.6	71.5	40.2	69.5
2pks	65.4	58.9	52.2	36.6	47.1	37.6
2zgp	66.7	63.3	54.5	42.8	49.5	43.4
average	58.9	51.9	44.1	31.6	37.8	31.5

Case 3[c] (FSRG)

	set III		set IV	
	Coelacanth decoy	DUD decoy	Coelacanth decoy	DUD decoy
1uze	25.5	55.6	27.9	55.6
1uzf	51.4	63.9	45.5	63.9
2pu2	23.4	63.9	42.5	84.6
2r9x	27.1	66.4	45.2	82.8
2w26	85.3	77.5	80.9	73.1
3ens	83.8	82.4	78.0	75.5
2pks	52.2	54.5	83.6	84.3
2zgp	36.6	42.8	76.0	79.2
average	48.2	63.4	60.0	74.9

[a] Results with the ordinary scoring function. [b] Results without the Coulomb and the hydrogen bonding interaction terms. [c] Results by the FSRG method.

Table 4. *q* Values of the Intact Active Compounds and Sets I, II, III, and IV Fragments for COX1

Case 1[a]

	intact active compound		set I		set II	
	Coelacanth decoy	DUD decoy	Coelacanth decoy	DUD decoy	Coelacanth decoy	DUD decoy
1cqe	75.8	57.6	70.5	48.2	16.2	19.4
1eqg	61.7	45.9	51.9	34.9	1.2	18.9
average	68.8	51.8	61.2	41.6	8.7	19.2

Case 2[b]

	intact active compound		set I		set II	
	Coelacanth decoy	DUD decoy	Coelacanth decoy	DUD decoy	Coelacanth decoy	DUD decoy
1cqe	83.6	66.9	88.1	70.5	88.6	69.1
1eqg	72.1	55.3	71.9	49.6	76.8	54.8
average	77.9	61.1	80.0	60.0	82.7	62.0

Case 3[c] (FSRG)

	set III		set IV	
	Coelacanth decoy	DUD decoy	Coelacanth decoy	DUD decoy
1cqe	93.4	76.7	92.0	70.5
1eqg	72.4	52.4	84.0	63.5
average	82.9	64.6	88.0	67.0

[a] Results with the ordinary scoring function. [b] Results without the Coulomb and the hydrogen bonding interaction terms. [c] Results by the FSRG method.

bonds of the compound. The weights of these terms were determined by the input file of the Sievgene program. The database enrichment curves of the fragments were slightly improved by ignoring the Coulombic and the hydrogen-bonding interactions, while the database enrichment curves of the intact active compounds became worse than the above results with the full interaction terms.

The enrichment curves with the Coelacanth decoy set were better than those with the DUD decoy set, as in Figure 4a. The database enrichment curves of set I were also slightly worse than those of set II.

Tables 2, 3 and 4 (see "Case 2") show the *q* values obtained by the MTS method. The *q* values without Coulombic or hydrogen bond interactions were lower than those with the ordinary potential function; however, the in silico screening by ignoring Coulombic and hydrogen bond interactions still worked in some cases. When the intact structures of active compounds were used, the *q* values were decreased by ignoring Coulombic and hydrogen bond interactions in many cases. For COX2, in 4 out of 6 cases, the *q* values were larger than 70 with the Coelacanth decoy set, and in 3 out of 6 cases the *q* values were larger than 70 with the DUD decoy set.

When the fragments were used, the *q* values increased by ignoring Coulombic and hydrogen bond interactions for COX2 and COX1. For the other proteins, the *q* values decreased by ignoring Coulombic and hydrogen bond interactions. For sets I and II, 4 out of 6 cases showed *q* values >70 with the Coelacanth decoy set, and 2 out of 6 cases showed *q* values >70 with the DUD decoy set for COX2. For the other targets (ACE, AMPC, FXA, and THR), 2 or 3 *q* values out of 8 were extremely small (less than 20).

3.4. Screening Results with Replicas Ignoring Coulombic and Hydrogen Bonding Interaction Energies: FSRG Method. Figure 4c shows the averaged database enrichment curves of the fragments by the FSRG method, in which the docking score does not include the Coulombic or the hydrogen-bonding interactions for COX2. The FSRG method was effective for COX2. The database enrichment curves of the fragments were drastically improved by using the replicas. The database enrichment curves of set IV were better than those of set III. When the Coelacanth decoy set was used, the database enrichment curves were better than those with the DUD decoy set as in Figure 4a,b. The hit ratio for the first 1% of the entries in the database of set IV with the Coelacanth decoy set and that with the DUD decoy set were 11.4% and 5.7%, respectively. The hit ratio for the first 1% of the entries in the database of set III with the Coelacanth decoy set and that with the DUD decoy set were 1.6% and 0.0%, respectively. The hit ratio at the first 1% of the entries in the database of the intact active compounds with the Coelacanth decoy set and that with the DUD decoy set were 14.2% and 13.2%, respectively. The hit ratio of set IV was lower than that of the intact active compounds, but still the hit ratio was obviously better than that of a random screening.

Tables 2, 3, and 4 (see "Case 3") show the *q* values obtained by the MTS method for COX2, COX1, ACE, AMPC, FXA, and THR. The *q* values obtained by the FSRG method were close to the values of the intact active compounds with the ordinary potential function for almost all the 16 proteins. Namely, when the Coelacanth decoy set was used, 9 out of 16 intact active compounds had $q > 70$ using the ordinary scoring function. Nine out of 16 set III cases and 12 out of 16 set IV cases had $q > 70$ with the FSRG method. When the DUD decoy set was used, 5 out of 16 intact active compounds had $q > 70$ using the ordinary scoring function. Three out of 16 set III cases and 11 out of 16 set IV cases had $q > 70$ with the FSRG method. The FSRG method with set IV worked well and showed high hit ratios in in silico fragment screening. For all 16 proteins, the average hit ratios for the first 1% compounds selected from the compound library were 1.2 and 5.4% for sets III and IV, respectively. The *q* value for set IV was close to the *q* value for the intact active compounds using the ordinary scoring function.

We also examined the scoring function dependence of the hit ratio of the FSRG method for COX2. When the weight of the Coulombic and hydrogen bonding interactions was set to 0.1, the average *q* value for set III with the Coelacanth decoy set and that with the DUD decoy set were 72.5 and 58.9, respectively. The average *q* value for set IV with the Coelacanth decoy set and that with the DUD decoy set were 84.3 and 72.6, respectively. These values were slightly worse than the values obtained by totally ignoring the Coulombic and hydrogen bonding interactions. Namely, the average *q* value for set III with the Coelacanth decoy set and that with the DUD decoy set were 73.3 and 59.5, respectively. The average *q* value for set IV with the Coelacanth decoy set and that with the DUD decoy set were 85.2 and 73.5, respectively.

When the weight of the Coulombic and hydrogen bonding interactions was set to 0.5, the hit ratio became worse than in the above cases. The average *q* value for set III with the Coelacanth decoy set and that with the DUD decoy set were 68.4 and 56.8, respectively. The average *q* value for set IV with the Coelacanth decoy set and that with the DUD decoy set were 78.4 and 66.4, respectively. Thus, ignoring the Coulombic and hydrogen bonding interactions could give the best results for sets III and IV.

3.5. Screening Results of COX2 and COX1. We compared the screening results for COX2 and COX1. COX1 and COX2 belong to the same protein family. The active compounds are the COX2 inhibitors, which can bind COX1 and it cause its side effects. The results for COX2 and COX1 are summarized in Tables 2 and 4, respectively. The trends of *q* values of COX1 are similar to those of COX2. Namely, the *q* values of the original active compounds are high, the *q* values of sets I and II are small, and the *q* values obtained by the FSRG method are high. The average *q* values for COX1 are slightly smaller than those for COX2, but the *q* values for COX1 obtained by the FSRG method are almost equivalent to the values for COX2.

The 3D structure of COX2 is very similar to that of COX1; their sequence identity is 64.26%. The difference in affinities between the COX1 selective inhibitor and the COX2 selective inhibitor[31] is so little that our docking program was unable to distinguish between them. A more precise score function or careful investigation of protein−ligand complex structures will be necessary to distinguish and develop a COX2 selective inhibitor.[32−34]

3.6. Molecular Size Dependency of Docking Accuracy. We investigated the relationship between the docking accuracy and the ligand size. We applied the Sievgene program to the self-docking test of 132 protein−ligand complex structures. The selected protein−ligand complex structures are listed in Appendix B. The data set and the procedure of this self-docking test were exactly the same as in the previous study.[19] The 3D coordinates of the inhibitors were generated by Chem3D (Cambridge Software, Cambridge, MA, USA). The conformations of the ligands, which were extracted from the protein−ligand complexes, were randomized before the current docking study. For flexible docking, up to 100 conformers were generated for each ligand. The average number of atoms of ligands is 44.5; the smallest ligand consists of only 12 atoms and the largest of 114 atoms. Fifty-six percent of predicted structures showed an rmsd < 2 Å. The correlation coefficient between the number of atoms of ligands and the rmsd values of the predicted structures was 0.011. This result showed that the Sievgene program could dock small ligands to its target protein as well as larger ligands. On the contrary, the database enrichment of fragments with ordinary interaction was not good. From the binding poses of these 132 protein−ligand complex structures, the binding pockets for small ligands were small, and the surface of the ligand was matched to the surface of the binding pocket. In such cases, the Sievgene program works well. For fragments, the surface of the binding pocket is much larger than the surface of each fragment. In this case, the Sievgene program cannot work well, since the surface complementarity is lost for fragments that are much smaller than the binding pocket.

4. DISCUSSION

The surface complementarity between the protein surface and compound surface is generally important in docking programs. In fact, in the Sievgene program, using 180 proteins and the Coelacanth decoy set, the contribution of the accessible surface interaction term to the docking score was, on average, 86.2%, while the contributions of the van der Waals interaction term, Coulombic interaction term, and the hydrogen bonding interaction were 1.2%, 0.1%, and 12.5%, respectively. That is why the in silico screening still worked when we ignored both the Coulombic and the hydrogen bonding interactions. The contribution of the van der Waals interaction was small, but it is important to avoid the atomic confliction between the protein and the compound. Thus, the FSRG method could work with the docking program, which underestimates the Coulombic and the hydrogen bonding interactions. If the Coulombic and the hydrogen bonding interactions are essential to a docking program, the FSRG method would not work.

When set IV was used, 11 out of 16 cases showed *q* values >70 with the DUD decoy set. In some cases, the FSRG method did not work well. However, the FSRG method is well suited for practical use. A previous work showed that the current in silico screening methods work well for screening known active compounds in roughly half of the cases, while these methods failed in the other half.[35] Compared to results in this previous report, the results obtained by the FSRG method for set IV were not bad.

The screening results with set IV were better than those with set III. The average number of atoms of set IV was only 0.3-2 atoms greater than that of set III. This difference was small, but the 3D structures of compounds of set IV had more similarity to the original (intact) active compounds. Thus, we must be careful to treat the linker part between the fragment and its side chains.

5. CONCLUSION

We developed a new structure-based in silico fragment screening based on protein-compound docking simulation. Our new FSRG method performs a virtual fragment screening for the first fragment selection of the FBDD. Several replica molecules were generated from each fragment by adding side chains to the fragment, and the FSRG method evaluates the activity of these replicas instead of the original fragment. In the FSRG method, the Coulombic and the hydrogen bonding interaction were ignored. Thus, only surface complementarity between protein and compound surfaces was evaluated in the protein-compound docking simulation and its score calculation. This score worked well to find active fragments among the decoy compounds.

We applied the FSRG method to the structure-based in silico fragment screenings of six target proteins. The known active compounds of these proteins were divided into two fragments manually. The compound library consisted of the fragments of the known active compounds and about 10^4 decoy compounds. The FSRG method worked well. The average hit ratio of 1% of compounds from the compound library for set IV obtained by the FSRG method was close to that for the original active compounds obtained by the conventional MTS method with the intact scoring function. The average area under the database enrichment curve (*q*) was 60−88% for the six target proteins. These values were close to the values of the intact active compounds.

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 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, 1 pso, 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.

APPENDIX B

The selected 132 proteins 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, 1 pso, 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.

Supporting Information Available: The original (intact) active compounds and fragments of sets I, II, III, and IV. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Orita, M.; Ohno, K.; Niimi, T. Two "Golden ratio" indices in fragment-based drug discovery. *Drug Discovery Today* , . in press.

(2) Hajduk, P. J.; Huth, J. R.; Fesik, S. W. Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.* **2005**, *48*, 2518–2525.

(3) Albert, J. S.; Blomberg, N.; Breeze, A. L.; Brown, A. J. H.; Burrows, J. N.; Edwards, P. D.; Folmer, R. H. A.; Geschwindner, S.; Griffen, E. J.; Kenny, P. W.; Nowak, T.; Olsson, L. L.; Sanganess, H.; Shapiro, A. B. An integrated approach to fragment-based lead generation: philosophy, strategy and case studies from AstraZeneca's drug discovery programmes. *Curr. Top. Med. Chem.* **2007**, *7*, 1600–1629.

(4) Erlanson, D. A.; McDowell, R. S.; O'Brien, T. Fragment-based drug discovery. *J. Med. Chem.* **2004**, *47*, 3463–3482.

(5) Alex, A. A.; Flocco, M. M. Fragment-based drug discovery: What has it achieved so far. *Curr. Top. Med. Chem.* **2007**, *7*, 1544–1567.

(6) Hajduk, P. J.; Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug Discovery* **2007**, *6*, 211–219.

(7) Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. Recent developments in fragment-based drug discovery. *J. Med. Chem.* **2008**, *51*, 3661–3680.

(8) Corey, E. J.; Wipke, W. T. Computer-assisted design of complex organic syntheses. *Science* **1969**, *166*, 178–192.

(9) Timothy, D.; Salatin, T. D.; Jorgensen, W. L. Computer-assisted mechanistic evaluation of organic reactions. 1. overview. *J. Org. Chem.* **1980**, *45*, 2043–2057.

(10) Funatsu, K.; Sasaki, S. Computer-assisted organic synthesis design and reaction prediction system, "AIPHOS". *Tetrahedron Comput. Methodol.* **1988**, *1*, 27–38.

(11) 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.

(12) 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.

(13) 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.

(14) 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.

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

(16) 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.

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

(18) 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.

(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) 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.

(23) 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.

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

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

(26) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. "Development and testing of a general amber force field". *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(27) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(28) Fukunishi, Y.; Mikami, Y.; Nakamura, H. The filling potential method: A method for estimating the free energy surface for protein-ligand docking. *J. Phys. Chem. B* **2003**, *107*, 13201–13210.

(29) 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.

(30) 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.

(31) Warner, T. D.; Giuliano, F.; Vojnovic, I.; Bukasa, A.; Mitchell, J. A.; Vane, J. R. Nonsteroid drug selectivities for cyclo-oxygenase-1 rather than cyclo-oxygenase-2 are associated with human gastrointestinal toxicity: A full *in vitro* analysis. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 7563–7568.

(32) Luong, C.; Miller, A.; Barnett, J.; Chow, J.; Ramesha, C.; Browner, M. F. Flexibility of the NSAID binding site in the structure of human cyclooxygenase-2. *Nat. Struct. Biol.* **1996**, *3*, 927–933.

(33) Leval, X.; Delarge, J.; Somers, F.; Tullio, P.; Henrotin, Y.; Pirotte, B.; Dogne, J. M. Recent advances in inducible cyclooxygenase (COX-2) inhibition. *Curr. Med. Chem.* **2000**, *7*, 1041–1062.

(34) Rao, P. N. P.; Uddin, M. J.; Knaus, E. E. Design, synthesis, and structure-activity relationship studies of 3,4,6-triphenylpyran-2-ones as selective cyclooxygenase-2 inhibitors. *J. Med. Chem.* **2004**, *47*, 3972–3990.

(35) Warren, G. L.; Webster Andrews, C.; 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.

CI800435X