# Classification of Chemical Compounds by Protein−Compound Docking for Use in Designing a Focused Library

Yoshifumi Fukunishi,*,† Yoshiaki Mikami,‡ Kei Takedomi,‡ Masaya Yamanouchi,‡ Hideaki Shima,‡ and Haruki Nakamura†,§

*Biological Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6, Aomi, Koto-ku, Tokyo 135-0064, Japan, 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 May 23, 2005*

We developed a new method for the classification of chemical compounds and protein pockets and applied it to a random screening experiment for macrophage migration inhibitory factor (MIF). The principal component analysis (PCA) method was applied to the protein−compound interaction matrix, which was given by thorough docking calculations between a set of many protein pockets and chemical compounds. Each compound and protein pocket was depicted as a point in the PCA spaces of compounds and proteins, respectively. This method was applied to distinguish active compounds from negative compounds of MIF. A random screening experiment for MIF was performed, and our method revealed that the active compounds were localized in the PCA space of compounds, while the negative compounds showed a wide distribution. Furthermore, protein pockets, which bind similar compounds, were classified and were found to form a cluster in the PCA space.

## Introduction

When a measure of similarity among many kinds of substances is provided, it should, in general, offer a number of applications. The classification of proteins and chemical compounds is one of the primary applications of such similarity measures in pharmaceutical science. The classification of proteins is important in examining biological functions and evolution; additionally, new drugs can be developed to recognize target proteins among similar proteins with high specificity. The classification of compounds aids in the identification of new active compounds which are similar to known active compounds and also in selecting a limited number of candidate active compounds, known as a focused library, from a large number of chemical compounds in a database. The molecular similarity facilitates the design of mimetics of an active compound, while providing a measure of diversity in the chemical compound library. Similarity searching and the evaluation of the chemical compound library are closely related techniques.

Many methods have been proposed for similarity searching of chemical compounds,[1] such as the overlapping of chemical structure, the CATS descriptor method developed by Schneider et al.,[2] the BCUT descriptor method,[3] etc. In the CATS descriptor method, for each pair of pharmacophoric features (donor, acceptor, acid, base, etc.) in the molecule, the frequency of occurrence as a function of the number of bonds separating the features is accumulated in a pharmacophore pair vector. The bond distances from 1 to 10 are considered over all 15 feature combinations to give a vector size of 150. The Euclidian distance between two pharmacophore pair vectors is used as the similarity.

BCUT is one of the most widely used descriptor methods to evaluate the similarity of chemical compounds and the diversity of a given library. BCUT is a set of several descriptors, which are eigenvalues of matrixes. The diagonal parts of the matrixes represent the atomic charge, polarizability, and hydrogen donors and acceptors, and the off-diagonal parts of the matrixes represent the structure of the compound.

Although the classification of proteins has been well studied by analyses of amino acid sequences, several recent studies have classified proteins based on their 3D structures,[4−15] and some recent studies have focused on the local structure around a protein pocket. Kinoshita and Nakamura,[5] for example, compared the molecular surfaces of proteins using a topological graph method, and Schmitt et al.[7] compared the distributions of functional groups in the pockets. These approaches have succeeded in the functional classification of non- or low-homologous proteins.

The conventional methods for the classification of compounds and proteins are based on the independent information of compounds and proteins, respectively. To date, many protein−compound docking programs have been developed[16−25] and large-scale computing allows us to calculate a protein−compound affinity panel. We here propose a new method for the classification of compounds and proteins based on the information provided by protein−compound docking. Our method was applied to distinguish active and negative compounds of macrophage migration inhibitory factor (MIF), which were observed by our in vitro assay.

## Methods

**Analysis.** A measure to represent the distance between two compounds is determined based on the protein−ligand interaction matrix, each element of which is the corresponding docking score. From the covariance matrix of compounds, a principle component analysis (PCA) is performed to find similar clusters of compounds. The same method can be applied to protein pockets as well as to compounds.

We prepare a set of pockets $P = \{p_1, p_2, p_3, \ldots p_{Nr}\}$, where $p_i$ represents the *i*-th pocket and Nr is the total number of pockets, and a set of compounds $X = \{x^1, x^2, \ldots x^{Nc}\}$, where $x^k$ represents the *k*-th compound and Nc is the total number of compounds. For

* 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 (AIST).

‡ Japan Biological Information Research Center (JBIRC).

§ Osaka University.

10.1021/jm050480a CCC: $33.50 
Published on Web 12/16/2005

each pocket $p_i$, all compounds of the set X are docked to the pocket $p_i$ with score $s_i^k$ between the *i*-th pocket and the *k*-th compound. Here, $s_i^k$ corresponds to the binding free energy.

Similarity (distance) between the *k*-th compound and the *l*-th compound is defined as follows:

$$D^{C}_{kl} = \sqrt{\sum_{i=1}^{Nr}(s_i^k - s_i^l)^2} \quad (1)$$

$D^{C}_{kl}$ satisfies the following three conditions, which are sufficient for a definition of distance: $D^{C}_{kk} = 0$, $D^{C}_{kl} = D^{C}_{lk}$ and $D^{C}_{kh} + D^{C}_{hl} \geq D^{C}_{kl}$. This definition was used for cluster analysis in our previous work.[26] To define the distance between two protein pockets $D^{P}_{ij}$, the same analysis can be applied, changing the suffix:

$$D^{P}_{ij} = \sqrt{\sum_{k=1}^{Nc}(s_i^k - s_j^k)^2} \quad (2)$$

The covariance matrix $M^C$ of compounds is defined as,

$$M^{C}_{kl} = \frac{1}{Nr}\sum_{i=1}^{Nr}(s_i^k - \bar{s^k})(s_i^l - \bar{s^l}) \quad (3)$$

and

$$\bar{s^k} = \frac{1}{Nr}\sum_{i}^{Nr} s_i^k \quad (4)$$

where the upper bar represents the average. Let $\phi_k$ be a *k*-th eigenvector of $M^C$ with eigenvalue $\epsilon_k$, and the order of $\epsilon_k$ is descendant. The vector of docking scores for the *k*-th compound $X_k = (s_1^k, s_2^k, \ldots s_{Nr}^k)$ is represented by the linear combination of $\phi_k$ as

$$X_k = \sum_{j=1}^{Nc} c_j \phi_j \quad (5)$$

The coefficient $\{c_j\}$ represents the coordinate of the PCA space of compounds. To calculate the PCA space of protein pockets, the same analysis can be applied, changing the suffix. The covariance matrix $M^P$ of compounds is defined as,

$$M^{P}_{ij} = \frac{1}{Nc}\sum_{k=1}^{Nc}(s_i^k - \bar{s_i})(s_j^k - \bar{s_j}) \quad (6)$$

and

$$\bar{s_i} = \frac{1}{Nc}\sum_{k=1}^{Nc} s_i^k \quad (7)$$

Let $\phi^i$ be an *i*-th eigenvector of $M^P$, and the vector of docking scores for the *i*-th pocket is then represented by the linear combination of $\phi^i$, whose coefficient represents the coordinate of the PCA space of pockets.

Protein−compound docking simulation was performed by our in-house program named Sievgene, which is a protein−ligand flexible docking program for in silico drug screening.[26] The scoring function of this method is based on the rough shape of a protein surface to reduce structural noise. The conventional potential function is applied to the outer region of the protein, while in contrast, a smooth virtual function is applied to the inner region of the protein. Assuming that at least three ligand atoms come into contact with the protein surface, a geometrical hash method is used for protein−ligand conformation searching. This method was applied to the 132 known protein−ligand complexes and correctly predicted ~50% of these complex conformations within the 2 Å RMSD,[26] attaining a similar performance to that achieved by popular docking programs.[27] In the present study, the number of conformers for flexible docking was limited to 100 for each compound.

**Preparation of Materials.** Two datasets were prepared. The first, which consisted of a total of 138 proteins and 1012 compounds, was a dataset to evaluate the localizability of active compounds in the PCA spaces. All of these proteins were extracted from complexes that were selected from the database used in the evaluation of the DOCK, FlexX and GOLD methods.[27] This set of 138 proteins provided a rich variety of proteins and compounds whose structures were all determined by high quality experiments with resolution of less than 2.5 Å. The lack of atom coordination was almost zero, and the atomic structure around the ligand pocket was quite reliable. We removed some complexes, which contained a covalent bond between the protein and ligand from the original data set, since our docking program cannot perform the protein−ligand docking when a covalent bond exists between the protein and the ligand. The protein databank (PDB) identifiers of the used complexes are listed in Appendix A.

Two types of subsets of protein pockets were prepared and the current classification method was applied to these two subsets for examination of the dependence on the choice of the proteins. One protein set consists of diverse proteins, and another protein set consists of similar proteins to the target protein. Since we will examine which protein set can divide well the active compounds from the negative compounds, both protein sets must consist of almost the same number of proteins.

First, a preliminary docking study was performed with the 138 proteins vs the 138 compounds. Here, the 138 compounds were the ligands of the 138 complexes extracted from the PDB as described in our previous report.[26] Cluster analysis based on the definition of the distances given by eq 2 was applied to the 138 proteins vs the score panel of the 138 compounds. The group average method divided 138 proteins into two kinds of clusters, namely 7 clusters and 23 clusters. These seven clusters showed a good correspondence to the conventional functional classification. When two clusters out of the seven were adopted, the multiple active site correction (MASC) scoring method showed good database enrichment for these two clusters, which consist of total 23 proteins. The members of each cluster would be similar to each other in point of ligand-binding function. As the diverse protein set, we selected the representative proteins from the 23 clusters. In each cluster, one of the two proteins between which the distance is shortest was selected to be the representative protein. These 23 proteins will be identified hereafter as "protein data set A" and they are listed in Appendix A.

Next, we tried to make a cluster, whose members are similar to the MIF, and the number of proteins of the cluster is almost the same as the number of proteins of protein set A. One cluster, which consists of 26 proteins including the MIF, was selected from the above 7 clusters dividing the 138 proteins. We adopted all members of the cluster except the MIF and these 25 proteins will be identified hereafter as "protein data set B" and they are listed in Appendix A. Neither protein data set A nor B includes the target protein MIF.

Two compound datasets were prepared. One dataset (compound set C) consists of 1012 compounds originated from a random chemical library, which includes active compounds and negative compounds of a target protein. The current classification method can be used as an in silico screening method. This dataset was used to evaluate the localization of the active compounds and the database enrichment of the current method. Another dataset (compound set D) consists of 1006 compounds, which are mainly some series of similar compounds, namely, amino acids, dipeptides, tripeptides, etc. This dataset was used to evaluate the localization of the series of compounds and classification of the proteins.

The compound set (compound set C) consisted of 12 active compounds and 1000 negative compounds extracted from the Coelacanth chemical compound library (Coelacanth Corporation, East Windsor, NJ). The 12 active compounds of MIF are listed in Figure 1, and their $IC_{50}$ values are listed in Table 1. Compounds **7**

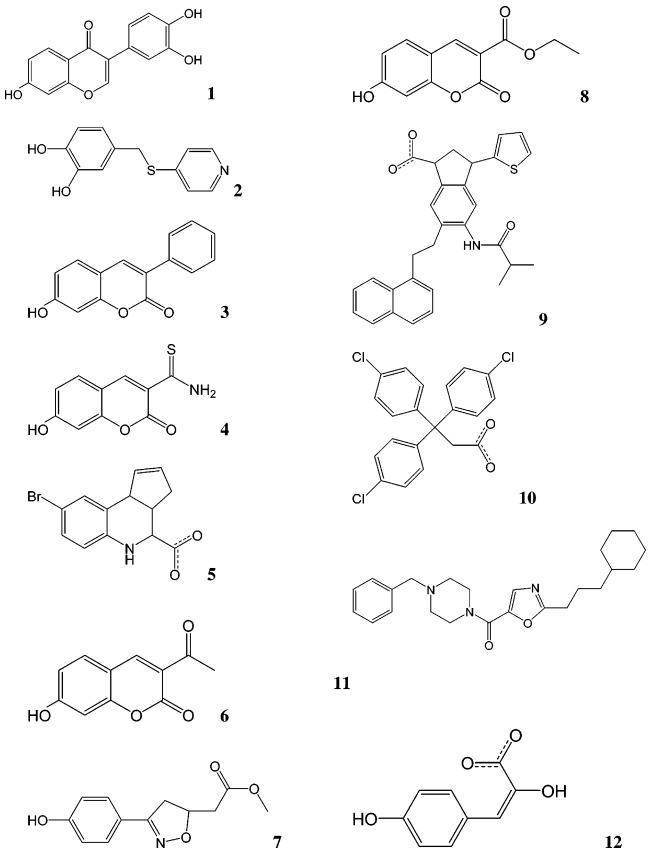

**Figure 1.** MIF active compounds.

**Table 1.** List of MIF Active Compounds

| compound[a] | $IC_{50}(\mu M)$ | source[b] |
|---|---|---|
| **1** | 0.038 | Y |
| **2** | 0.4 | T |
| **3** | 0.47 | Y |
| **4** | 0.55 | Y |
| **5** | 3.4 | T |
| **6** | 4.3 | Y |
| **7** | 7 | 1LJT |
| **8** | 7.4 | Y |
| **9** | 8 | T |
| **10** | 8.1 | T |
| **11** | 30 | T |
| **12** | no data | 1CA7 |

[a] The compound serial number is consistent with the number in Figure 1. [b] "Y" indicates data originally presented by Orita et al.[28] "T" represents data from the current study, and "1LJT" and "1CA7" represent the PDB IDs which were the source of the data in question.

and **12** were selected from the PDB, and compounds **1**, **3**, **4**, **6** and **8** were reported in a previous study.[28] The others (compounds **2**, **5**, **9**, **10**, **11**) were prepared in the present study. The definition of the active compound is described at the last of this section. The 11047 compounds of the original Coelacanth chemical compound library, which is a random library, were put in alphabetical order, and the top 1000 negative compounds were selected. The correlation coefficient between the order and the number of atoms of compound is only 0.12; thus, these 1000 compounds will form a random library.

The size distribution of ligands is as follows: ratio of 0−19 atoms, 0.0%; ratio of 20−29 atoms, 0.5%; ratio of 30−39 atoms, 0.5%; ratio of 40−49 atoms, 6.5%; ratio of 50−59 atoms, 22.5%; ratio of 60−69 atoms, 40.4%; ratio of 70−79 atoms, 22.1%; and ratio of more than 80 atoms, 7.4%. The average ligand size was 64.3 atoms.

The 3D coordinates of chemical compounds were generated by the Concord program (Tripos, St. Louis, MO) from the 2D Sybyl SD files provided by the Coelacanth Chemical Corporation. The atomic charges of each ligand were determined by the Gasteiger method.[29,30] The atomic charges of proteins were the same as the atomic charges of AMBER parm99.[31]

**Table 2.** List of AMP/ADP/ATP Binding Proteins

| PDB ID | name of ligand[a] | residue A | PA | residue B | PB | residue C | PC |
|---|---|---|---|---|---|---|---|
| 12as | AMP | 100R | 1 | | | | |
| 1ady | HAM | 259R | 1 | | | | |
| 1aer | AMP | 458R | 1 | | | | |
| 1asz | ATP | 325R | 1 | 531R | 3 | | |
| 1aux | ADT | Ca ion | 1,2,3 | | | | |
| 1b76 | ATP | 186R | 1 | 220R,231R, 366R | 3 | | |
| 1csn | ATP | Mg ion | 1,2,3 | | | | |
| 1efv | AMP | 126A main chain | 1 | | | | |
| 1gol | ATP | Mg ion | 2,3 | | | | |
| 1gtr | ATP | 43H | 1 | 270K | 2,3 | | |
| 1hck | ATP | Mg ion | 1,2,3 | | | | |
| 1nks | ADP | 54R | 1 | | | | |
| 1pyg | AMP | 81R,193R, 310R | 1 | | | | |
| 1ses | AMP | 528R | 1 | | | | |
| 2tmd | ADP | none | none | | | | |
| 3r1r | ATP | 9K | 1 | 91K | 2 | 21K | 3 |

[a] The abbreviations HAM and ADT indicate histidyl adenosine monophosphate, and adenosine diphosphate monothiophosphate, respectively. "Residue A" and "PA" are the binding residue or atom of the protein and the serial number of bonded phospate, where 1, 2 and 3 represent the first, second and third phosphates of AMP/ADP/ATP, respectively; the same applies to the terms "Residue B", "PB", "Residue C" and "PC". "Residue A", "Residue B" and "Residue C" bind "PA", "PB" and "PC", respectively.

**Table 3.** List of Sugar Binding Proteins

| PDB ID | name of ligand |
|---|---|
| 1abe1 | α-L-arabinose |
| 1abe2 | β-L-arabinose |
| 1abf1 | α-D-fucose |
| 1abf2 | β-D-fucose |
| 5abp1 | α-D-galactose |
| 5abp2 | β-D-galactose |
| 1bdg | glucose |
| 1glg | D-galactose |
| 2gbp | β-D-glucose |
| 1mmu | α-D-glucose |
| 1qh7 | β-D-xylopyranose |
| 1so0 | D-galactose |
| 2aac | β-D-fucose |

The name of the ligand follows those in the PDB file.

Another data set (compound set D) was also prepared to evaluate the localizability of similar compounds and similar proteins in the PCA spaces. This dataset consisted of a total of 169 proteins and 1006 compounds. Of the proteins, 138 were the basic protein sets described above. In addition, 5 sugar binding proteins, 16 adenosine-phosphate binding proteins, and 10 enzymes of prostaglandin were included in the dataset. Tables 2−4 show the PDB IDs of these proteins and ligands. In the present study, the sequence identities and similarities are global sequence identities and global sequence similarities calculated by FASTA with the Blosum50 matrix.[32,33] Tables 5 and 6 show the names and structural information of the 16 AMP/ADP/ATP binding proteins and the 5 sugar binding proteins, respectively; note that the homology of these proteins is quite low except in several protein pairs.[34] With the exception of the pairs 12as-1asz, 1csn-1gtr, 1csn-1nks, 1gol-1nks, 1gol-1ses, 1hck-1ses and 1gh7-1so0, the amino acid sequence identities are only 20−30%. Tables 5 and 6 also show the CATH classification (Class, Architecture, Topology, Homologous Superfamily) of the protein 3D structures.[35,36] The architectures of the proteins also differ, so the dataset includes a variety of 3D structures of proteins.

The compound data set includes the ligands of the 138 basic proteins, the additional 6 monosaccharide binding proteins and the 16 AMP/ADP/ATP binding proteins described above. In addition, 20 amino acid monomers, 400 dipeptides, and 400 tripeptides are included. For the tripeptides, since the total number of tripeptides

**Table 4.** List of Prostaglandin Binding Proteins and Inhibitors

| PDB ID | protein | name of ligand[a] | selectivity |
|---|---|---|---|
| none | none | suprofen | COX1 selective |
| 1eqg | COX-1 | ibuprofen | COX1 selective |
| 1eqh | COX-1 | flurbiprofen | COX1 selective |
| 3pgh | COX-2 | flurbiprofen | COX1 selective |
| 4cox | COX-2 | indomethacin | COX1 selective |
| 1s2a | prostaglandin D2 11-keto reductase | indomethacin | COX1 selective |
| none | none | ketoprofen | COX1 selective |
| none | none | naproxen | COX1 selective |
| 1cx2 | COX-2 | Sc-558[b] | COX2 selective |
| 6cox | COX-2 | Sc-558[b] | COX2 selective |
| 1s2c | prostaglandin D2 11-keto reductase | flufenamic acid | no data |
| none | none | celecoxib | COX2 selective |
| none | none | etodolac | COX2 selective |
| none | none | nimesulide | COX2 selective |
| none | none | rofecoxib | COX2 selective |
| 1pxx | COX-2 | diclofenac | COX2 selective |
| 1cqe | COX-1 | flurbiprofen | COX1 selective |

[a] The name of the ligand follows those in the PDB file. [b] 1-Phenylsulfonamide-3-trifluoromethyl-5-*p*-bromophenylpyrazole.

**Table 5.** CATH Classification and Protein Names

| PDB code | CATH name[a] | CATH no.[b] | name |
|---|---|---|---|
| 12as | TLS | 30.930.10 | asparagine synthetase |
| 1ady | ABA | 40.50.800 | histidyl-tRNA synthetase |
| 1aer | CMP | 90.175.10 | exotoxin |
| 1asz | TLS | 30.930.10 | aspartyl-tRNA synthetase |
| 1aux | TLS | 30.470.20 | synapsin Ia |
| 1b76 | no data | no data | glycyl-tRNA synthetase |
| 1csn | TLS | 30.200.20 | casein kinase-1 |
| 1efv | ABA | 40.50.1120 | electron-transfer flavoprotein |
| 1gol | TLS | 30.200.20 | Map kinase Erk2 |
| 1gtr | ABA | 40.50.620 | glutaminyl-tRNA synthetase |
| 1hck | TLS | 30.200.20 | cyclin-dependent kinase 2 |
| 1nks | ABA | 40.50.300 | adenylate kinase |
| 1pyg | CMP | 90.270.10 | glycogen phosphorylase b |
| 1ses | TLS | 30.930.10 | seryl-tRNA synthetase |
| 2tmd | ABA | 40.50.720 | trimethylamine dehydrogenase |
| 3r1r | CMP | 90.188.10 | ribonucleotide reductase R1 |

[a] CATH classification name. TLS, ABA and CMP correspond to two-layer sandwich, $\alpha\beta\alpha$ sandwich and complex structures, respectively. [b] CATH classification number. The first digits of all proteins are "3".

**Table 6.** CATH Classification and Protein Names

| PDB code | CATH name[a] | CATH number[b] | name |
|---|---|---|---|
| 1abe,1abf,5abp[c] | $\alpha\beta\alpha$ sandwich | 3.40.50.2300 | L-arabinose-binding protein |
| 1bdg | $\alpha\beta\alpha$ sandwich | 3.40.367.20 | hexokinase |
| 1glg,2gbp[c] | $\alpha\beta\alpha$ sandwich | 3.40.50.2300 | chemotactic protein receptor |
| 1mmu | no data | 8.1.176.1 | galactose mutarotase |
| 1qh7 | $\beta$ sandwich | 2.60.120.180 | xylanase |
| 1so0 | no data | no data | galactose mutarotase |
| 2aac | $\beta$ sandwich | 2.60.120.280 | regulatory protein Arac |

[a] CATH classification name. [b] CATH classification number. [c] The ligands of these proteins are different from each other. The names of these ligands are summarized in Table 3.

is as large as 8000 ($=20^3$), all 20 amino acids were used as the first and last residues, and the second residues were randomly selected from 20 amino acids to generate 400 tripeptides. Additionally, 9 disaccharides in the Cambridge Structural Database (CSD)[37] were added to the compound data set: $\alpha$-cellobiose, $\beta$-cellobiose, $\alpha$-lactose, $\beta$-lactose, $\alpha$-maltose, galabiose, D-mannose, $\alpha$-D-talose, and trehalose. Finally, the redundant 17 COX-2 inhibitors listed in Table 4 were included as examples of drugs. The total number of compounds was 1006.

The size distribution of the ligands is as follows: ratio of 0−9 atoms, 0.1%; ratio of 10−19 atoms, 1.7%; ratio of 20−29 atoms, 14.1%; ratio of 30−39 atoms, 29.0%; ratio of 40−49 atoms, 26.2%; ratio of 50−59 atoms, 19.6%; ratio of 60−69 atoms, 6.5%; and ratio of more than 70 atoms, 1.5%. The average ligand size was 42.7 atoms.

The atomic charges of each ligand were determined by the restricted electrostatic potential (RESP) procedure using HF/6-31G* level quantum chemical calculations[31] which were performed using the programs GAMESS and Gaussian98.[38,39] The atomic charges of proteins, dipeptides and tripeptides were the same as those of AMBER parm99.[31]

The MIF active compounds were found by the following experimental procedure. Binding affinity was experimentally observed between MIF and 11046 chemical compounds selected from the Coelacanth random library and their related compounds. The first in vitro screening was a surface plasmon resonance assay done by Biacore 3000 (Biacore International AB, Neuchâtel, Switzerland). The second screening was carried out by tautomerase enzymatic assay, which determined the $IC_{50}$ values of the active compounds and verified the results of the first screening. The experimental conditions were the same as those previously reported by Orita et al.:[28] pH 6.0 with a buffer containing 25 mM potassium phosphate, 0.5 mM EDTA, 0.01% tween20, and 0.25 mM L-dopachrome methyl ester as a ligand. The density of MIF was 125 ng/mL. All experiments were performed at room temperature.

The definition of "active compound" was determined as follows; the compound, which showed the reaction unit (RU) < 20 before injection and the RU > 20 after injection by the Biacore, was selected as an active compound. Then, the $IC_{50}$ values of the active compounds found by the Biacore were determined as shown in Table 1, then a compound whose $IC_{50}$ value < 30 $\mu$M was determined as an active compound. The $IC_{50}$ value of one compound (compound **12**) was unclear, but the compound was selected as the active compound to increase the diversity of the set of active compound.

## Results and Discussion

**Applications to the Focused Library for MIF.** We applied the current method to the analysis of the active and negative compounds of MIF given by our in vitro screening assay. All 138 protein pockets were analyzed against the 1012 compounds (compound set C). The average CPU time for a docking procedure per pair of a pocket and a ligand was 149.4 s on a Compaq Alpha ES40 workstation. Two types of subsets of protein pockets, protein sets A and B, were then prepared and the current method was applied to these two subsets for examination of the dependence on the choice of the proteins. The 23 representative protein pockets of protein set A were analyzed against the 1012 compounds. Figure 2 shows the corresponding PCA plots. Then the 25 protein pockets of protein set B were analyzed against the 1012 compounds. The PCA plots are shown in Figure 3.

The PCA results based on the full set of 138 proteins carry the information of all 138 × 1012 docking scores, while the PCA results based on protein data set A and B carries the information of only 23 × 1012 and 25 × 1012 docking scores, respectively. Thus, the PCA result based on the full set of 138 proteins is thought to be one of the best classification results. The PCA plots are shown in Figure 4.

The 7 known active compounds (compounds **1**, **3**, **4**, **6**, **7**, **8** and **12**) are depicted as red circles in Figures 2−4, and the 5 active compounds found in the in vitro screening assay (compounds **2**, **5**, **9**, **10** and **11**) are depicted as green circles. Two or three of the newly found active compounds are close to the known active compounds in Figures 2−4. In Figure 2, the first, second, third, fourth, fifth and sixth eigenvalues of $M^C$ in eq 3 were 50.24%, 14.67%, 6.50%, 3.35%, 2.88% and 2.21%, respectively. Total 80.04% information was depicted. In Figure 3, the first, second, third, fourth, fifth and sixth

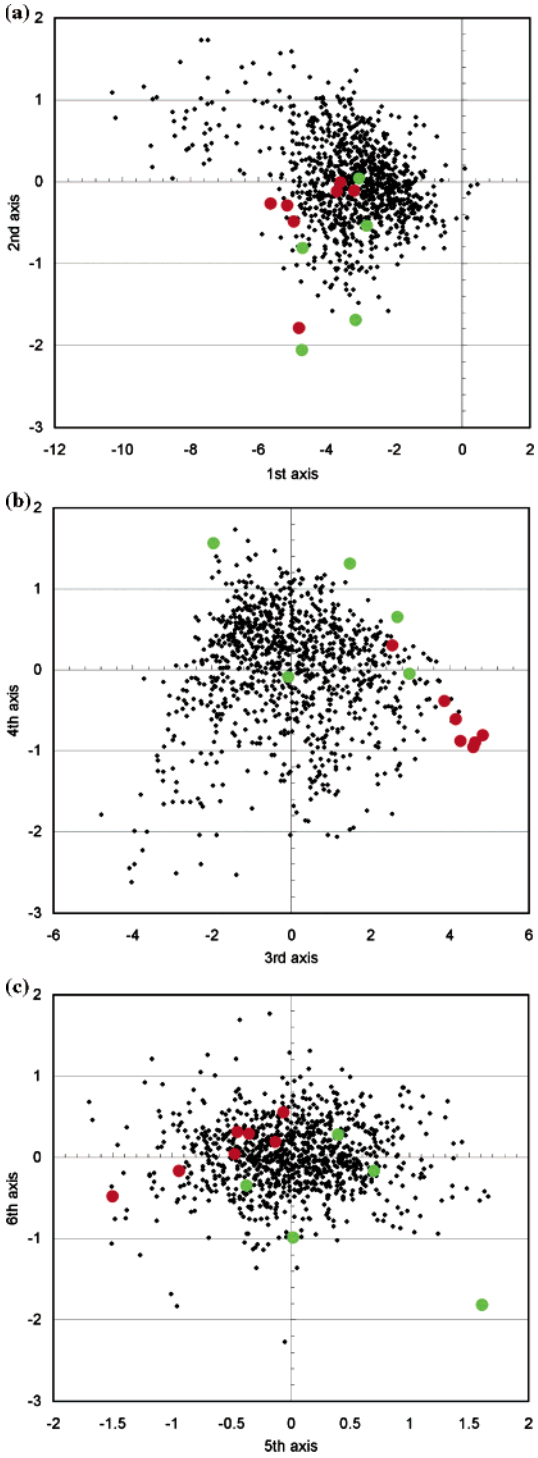


**Figure 2.** PCA plots of MIF active and negative compounds. Twenty-three representative proteins (protein data set A) were used for the analysis. The red circles represent the 7 known active compounds, and the green circles represent the 5 newly found active compounds. (a) PCA plot with the first and second major coordinates; (b) PCA plot with the third and fourth major coordinates; (c) PCA plot with the fifth and sixth major coordinates.

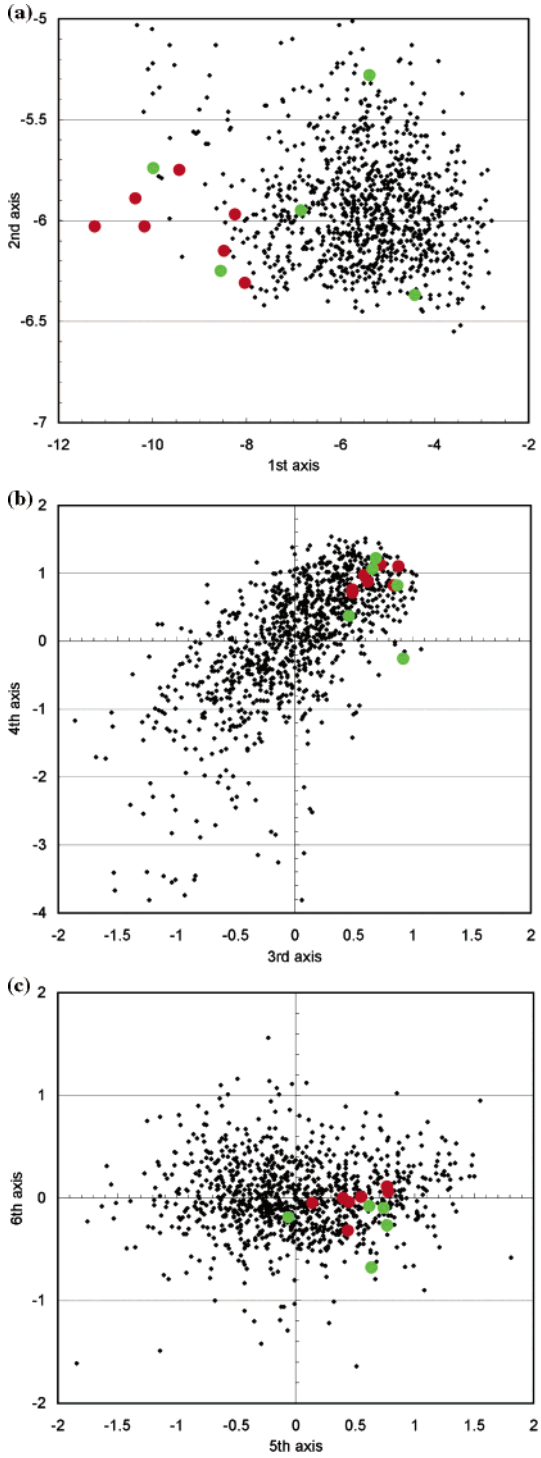


**Figure 3.** PCA plots of MIF active and negative compounds. Twenty-five proteins of a cluster, which includes MIF were used for the analysis. MIF itself was excluded from the protein data set (protein data set B). The red circles represent the 7 known active compounds, and the green filled circles represent the 5 newly found active compounds. (a) PCA plot with the first and second major coordinates; (b) PCA plot with the third and fourth major coordinates; (c) PCA plot with the fifth and sixth major coordinates.

eigenvalues of $M^C$ in eq 3 were 53.17%, 14.82%, 6.91%, 2.55%, 2.42% and 2.41%, respectively. Total 82.28% information was depicted. In Figure 4, the first, second, third, fourth, fifth and sixth eigenvalues of $M^C$ in eq 3 were 38.59%, 22.24%, 5.08%, 2.63%, 1.62% and 1.28%, respectively.

In Figures 2−4, the distribution of the active compounds is localized comparing to the distribution of the negative compounds in almost all cases. These results show that it is possible to select a set of candidate active compounds, which is a so-called focused library, even if the 3D structure of the target protein is not available. When one or more active compounds are known a priori, the compounds, which are close to the known active compound(s) in the PCA space could be candidate active compounds. In Figures 3a and 4, the degree of localization of active compounds is similar, and the distributions of the active

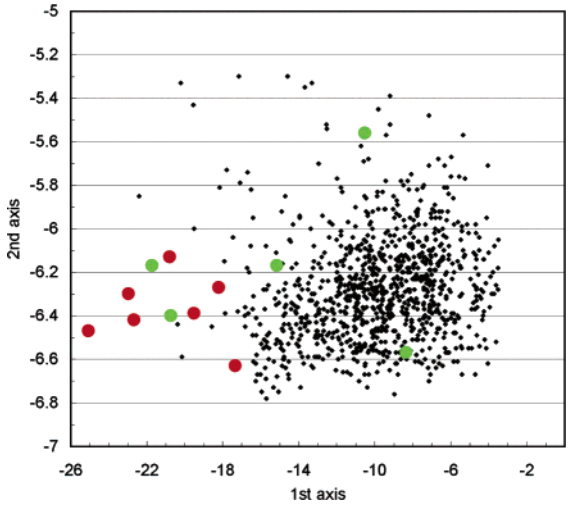


**Figure 4.** PCA plot of MIF active and negative compounds with the first and second major coordinates. All 138 proteins were used for the analysis. The red circles represent the 7 known active compounds, and the green circles represent the 5 newly found active compounds.

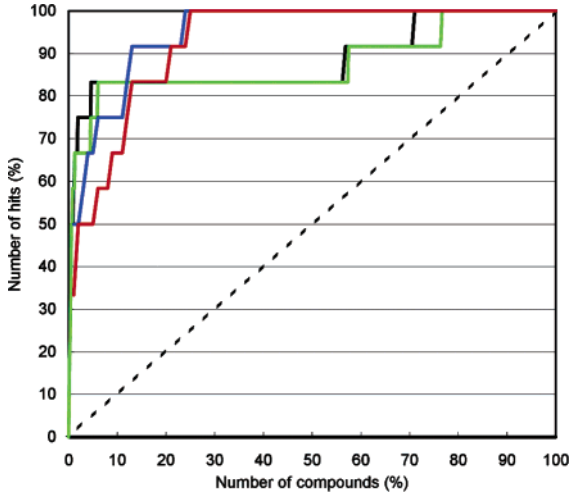


**Figure 5.** Database enrichment for MIF. The black, green, blue and red lines correspond to the database enrichment curves by the current PCA study with all 138 proteins, with the protein data set B, by the MASC scoring method with all 138 proteins, and by the MASC scoring method with protein data set B and the target MIF, respectively.

compounds localize better than those shown in Figure 2a. Thus, when the proteins are suitably selected, as in the present case in which the proteins are close to the target protein, a data set of only some proteins (25 proteins in the present case) gives results similar to those obtained using the data set of all proteins (138 proteins). Additionally, the present results show that the distribution does not change even if the target protein is not used in the analysis.

Figures 2b and 2c show the PCA results of protein data set A projected onto the third and fourth axes, and the fifth and sixth axes, respectively. The active compounds localize in Figure 2a, but do not so localize in Figures 2b and 2c. The projection of the active compounds onto the seventh and eighth axes localizes again. Figures 3b and 3c show the PCA results of protein data set B projected onto the third and fourth axes, and the fifth and sixth axes, respectively; the active compounds localize in both figures. Additionally, the projection of the active compounds onto the seventh and eighth axes also localizes. Thus, to make a focused library, 4 or more axes should be considered.

In Figures 2b and 2c, only one or two of the newly found active compounds are close to the known active compounds, but in Figures 3b and 3c, almost all of the newly found active compounds are close to the known active compounds. These results suggest that the current analysis is useful to predict new active compound(s) based on the known active compound(s) when the protein data set is suitably selected.

The volume of the spatial distribution of the active compounds was estimated. In N-dimensional PCA space, the average radii of distribution of active and all compounds were calculated assuming that they are spherical distributions. Then, the volumes and overlapping volumes among these spheres were calculated in the N-dimensional space. The larger volume means the wider distribution. The overlapping volumes were scaled in %, namely, the overlapping volume between the sphere of active compounds and the sphere of the all compounds was divided by the volume of the sphere of the all compounds. For protein set A, the volumes of active compounds in the whole compound's space were 26.0%, 15.0%, 20.2%, 42.9%, 30.4%, 27.0% and 29.5% for 1, 2, 3, 4, 5, 6 and 10 dimensional spaces, respectively. For protein set B, the volumes of active compounds in the whole compound's space were 59.9%, 4.4%, 5.1%, 4.9%, 4.6%, 8.1% and 1.3% for 1, 2, 3, 4, 5, 6 and 10 dimensional spaces, respectively. These results are consistent with the Figures 2 and 3. The active compounds localized when protein set B is used rather than protein set A.

Protein data set B gives better results than protein data set A; thus, the choice of a suitable set of proteins is important to design a focused library. One way to make a good choices is to adopt proteins in a cluster, which includes the target protein.

In Figures 2−4, the centers of distribution are different from the origin of the axes. The docking score corresponds to the binding free energy so that the range of score values is −5.5−0.0 in the present study. In some cases, the docking procedure fails to generate a protein−ligand complex structure, and the score is then set to +1. These points, which correspond to misdocking, are located far away from the major cluster of points in the PCA space so that the center of the distribution of the major cluster is different from the origin of the axes.

**Generation of the Focused Library.** A focused library was generated using the following method. In the PCA space, the compounds in a sphere whose center was set to the mass center of known compounds were selected as a focused library. The principal component axes were scaled to set the standard deviation of the distribution of compounds of each axis to 1. Ten major principal components were used. Figure 5 shows the database enrichment of our focused library, changing the radius of the sphere. The results obtained by the Multiple Active Site Correction (MASC) scoring method are also depicted for comparison.[26,40]

The increase of number of the used principal components does not mean the increase of the enrichment. The enrichment changes due to the number of the used principal components as follows: 16.7%, 83.3%, 75.0%, 75.0% and 75.0% of the ligands were found among the first 10% of the database with 1, 5, 10, 15, 20 principal components, respectively, when the protein set B was used. The first principal component, the first 5 principal components, the first 10 principal components, and the first 20 principal components have 53.17%, 79.87%, 89.98%, and 99.48% of the total information. The minor principal components will have noise, which is a computational error, we do not need the all information of the protein−compound affinity matrix to achieve a good enrichment.

The database enrichment of our focused library is better than that obtained by the MASC scoring method, until when the first

15% of compounds are selected for the database. Even if all proteins are used for screening, the present system remains superior, until the first 15% of compounds are selected for the database, after which the MASC system is better. The result of the protein data set without the target protein MIF is quite close to that of the protein data set with MIF, suggesting that our approach is effective even if the precise 3D structure of the target protein is unknown. The present method measures the distances among compounds by the docking scores. The protein pockets were used as probes to examine the chemical structure, which could be the partial structure or the whole structure, of each compound. Thus, when we adopt enough number of protein pockets to distinguish the all compounds of the compound library, the present method could divide the active compounds from the negative compounds even without the target protein.

In a previous study, a cell-based analysis was applied to prepare a focused library, and this analysis was adopted in the BCUT method.[41] The compound space was divided into cells by mesh, and the compounds included in each cell in which an active compound was included were adopted into the focused library. The cell-based analysis is suitable for designing a random library, but the present method was found to be more effective for generating a focused library in the current study. The distribution of active compounds is more circular (sphere) than square (cube), and the volume of the circle (sphere) is smaller than that of the square (cube) covering the distribution of the active compounds.

The most conventional indexes are the mass weight and Log P of each compound. The principal components by the present study include the information about the number of atoms and the solvation free energy of each compound as mentioned later in "classification of compounds" section. Since the number of atoms and the solvation free energy are closely related to the mass weight and Log P, the present study considers these indexes. We examined the efficiency of classifications by the present method and the conventional method. The 1012 compounds in compound set C were plotted onto 2D space, namely number of atoms vs solvation free energy. Then the same screening method as the present study was applied to the 2D space. The conventional method yielded a 6.7−13.3-fold enrichment, with 66.7% and 66.7% of the ligands found among the first 5% and 10% of the database, respectively. On the contrary, the present method based on protein set B yielded a 8.3−15-fold enrichment, with 75.0% and 83.3% of the ligands found among the first 5% and 10% of the database, respectively. The result by the present study was better than that of the conventional method.

The known seven active compounds consist of only 20−31 atoms, and they are much smaller than the average size of the compound set C, 64.3 atoms. But the newly found five active compounds consist of 28−65 atoms, and they are a little bigger than these seven active compounds. Thus, the conventional method could find the seven known active compounds out of 12, but it was difficult to find the rest newly found actives. On the contrary, the present method could overcome this defect.

**Classification of Proteins.** A total of 169 proteins were analyzed using 1006 compounds (compound set D) as stated above. The average CPU time for a docking procedure by Sievgene was 68.1 s on a Compaq Alpha ES40 workstation.

Tables 5 and 6 show the structural information of the AMP/ADP/ATP binding proteins and sugar binding proteins. Most of the global amino acid sequence identities are less than 25%,[34] and also the structure classification proves that these 3D folds are different to each other. The present structure classification

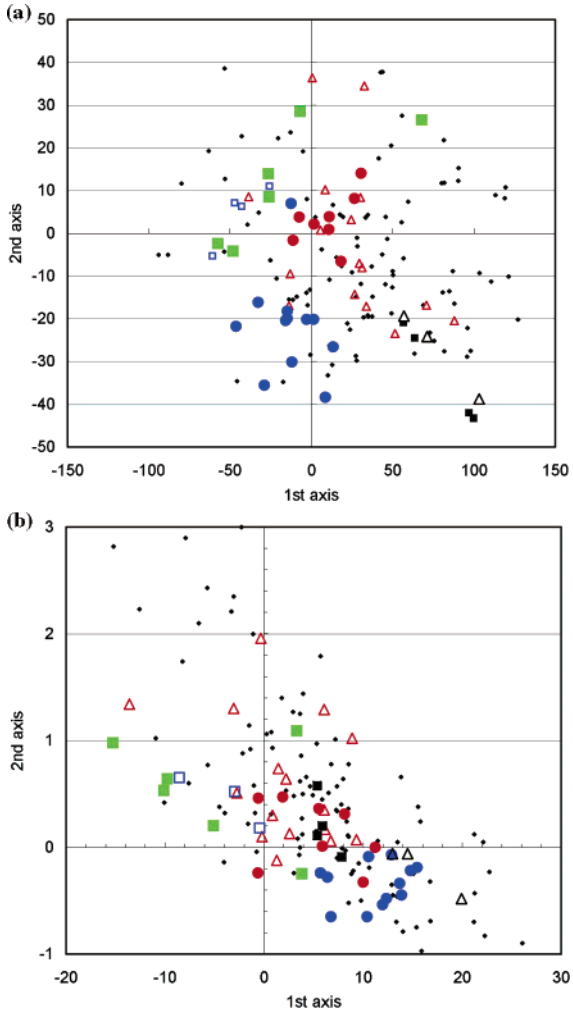


**Figure 6.** PCA plots of protein space. The red circles, blue circles, green squares, black squares, blue open squares, red open triangles, black open triangles and dots represent metalloproteases, acid proteases, sugar binding proteins, neuraminidases, endoribonucleases, AMP/ADP/ATP binding proteins, catalytic antibodies and other proteins, respectively. (a) Using 1006 compounds; (b) using 1000 compounds extracted from the Coelacanth random library.

follows the CATH classification, and the first, second, third and fourth digits represent the class, architecture, topology and homology classification numbers.[35,36]

Figure 6a shows the results of PCA plotting of protein pockets with the first and second major coordinates. The first, second, third and fourth eigenvalues of $M^P$ in eq 6 were 62.27%, 12.57%, 3.35% and 2.63%, respectively.

The present classification method successfully classified the protein binding pockets. As shown in Tables 5 and 6, some proteins, which bind the similar ligands, are in the different family or have different 3D fold. Proteins, which bind similar ligands, could form clusters in the protein space, even if the amino acid sequences were not homologous.

The distributions of endoribonucleases, metalloproteases, sugar binding proteins and acid proteases are localized near each other, indicating that the PCA of the docking score matrix works well as a functional analysis. Endoribonuclease binds RNA and hydrolyzes its phosphodiester chemical bond. The RNA is composed of only four kinds of RNAs (A, U, C and G), and the diversity of the RNA binding pocket is therefore expected to be small.

The acid proteases in the present study were aspartic protease (5er1, 1epo, 1eed), pepstatin (1apt, 1apu), rennin (1rne) and HIV

**Table 7.** List of Endoribonucleases

| PDB code | CATH name[a] | CATH no.[b] | name |
|---|---|---|---|
| 2aad | single sheet | 2.20.25.50 | ribonuclease T1 |
| 1rnt,6rnt[c] | single sheet | 2.20.25.50 | ribonuclease T1 |
| 1rds | single sheet | 2.20.25.50 | ribonuclease Ms |

[a] CATH classification name. [b] CATH classification number. [c] The ligands of 1rnt and 6rnt are guanosine-2′-monophosphate and adenosine-2′-monophosphate, respectively.

**Table 8.** List of Metalloproteases

| PDB code | CATH name[a] | CATH no.[b] | name |
|---|---|---|---|
| 1hfc | αβα sandwich | 3.40.390.10 | fibroblast collagenase |
| 1atl | αβα sandwich | 3.40.390.10 | atrolysin C |
| 1hyt,1lna, 1tlp,1tmn[c] | roll | 3.10.170.10 | thermolysin |
| 1jap | αβα sandwich | 3.40.390.10 | matrix metalloproteinase-8 |

[a] CATH classification name. [b] CATH classification number. [c] The ligands of 1hyt, 1lna, 1tlp and 1tmn are benzylsuccinic acid, dimethyl sulfoxide, phosphoramidon and *N*-(1-carboxy-3-phenylpropyl)-L-leucyl-L-tryptophan, respectively.

**Table 9.** List of Acid Proteases

| PDB code | CATH name[a] | CATH no.[b] | name |
|---|---|---|---|
| 5er1,1epo,1eed[c] | barrel | 2.40.70.10 | endothiapepsin |
| 1apt,1apu[d] | barrel | 2.40.70.10 | penicillopepsin |
| 1rne | barrel | 2.40.70.10 | renin |
| 1htf,1hos, 1hpv[e] | barrel | 2.40.70.10 | HIV-1 protease |
| 1ida | barrel | 2.40.70.10 | HIV-2 protease |
| 1qbu | barrel | 2.40.70.10 | HIV-1 protease |

[a] CATH classification name. [b] CATH classification number. [c] The ligands of 5er1, 1epo and 1eed are leucinol, n-carbonylmorpholine and Pd125754 (*tert*-butyloxycarbonyl-1-hydroxy-3-phenylalanine-propylene-1-hydroxy-3-phenylalanine-ethylene), respectively. [d] The ligands of 1apt and 1apu are isovaleryl (Iva)-Val-Val-Lysta-O-Et and isovaleryl (Iva)-Val-Val-Sta-O-Et, respectively. [e] The ligands of 1htf, 1hos and 1hpv are Gr126045 (2-(benzylcarbamoyl-phenylacetylamino-methyl)-5,5-dimethyl-thiazolidine-4-carboxylic acid (hydrozymethyl-2-phenylethyl)amide)), Sb204144 ((2-phenyl-1-carbobenzyl-oxyvalyl-amino)ethyl-phosphinic acid) and Vx-478 (3(*S*)-*N*-(3-tetrahydrofuranyloxycarbonyl)amino-1-(*N*,*N*-isobutyl,4-aminobenzenesulfonyl)amino-2-(*S*)-hydroxy-4-phenylbutane), respectively.

protease (1htf, 1hos, 1hpv, 1ida, 1qbu) as listed in Table 9. With the exception of the pairs 5er1−1apt, 1rne-1htf, 1rne-1qbu, 1htf-1ida, 1htf-1qbu, 1ida-1qbu, the amino acid sequence identities are only 20−30%. The sequence identity between these acid proteases is not high, but the local structures of pockets are similar in that all are at the clefts between the two units of the protein dimers, and these tunnel-like pockets are covered by loop regions. Furthermore, the sizes of these pockets are similar. Indeed, the inhibitors of HIV protease were developed based on rennin inhibitors.[42]

The binding pockets of metalloproteases are also very similar. Each binding pocket is at the cleft between a $\beta$-strand and an α-helix that is in parallel with the $\beta$ strand, and a metal ion that binds the ligand directly also binds the α-helix. However the amino acid sequence identities among them are not so high, namely, with the exception of the pairs 1hfc-1atl and 1htc-1jap, the amino acid sequence identities are only 20−24%.

The CATH classification numbers support these observations. As shown in Tables 7−9, the CATH classification numbers of most of the proteins examined in the present study are the same in each group.

The proteins with the same functions are plotted as neighboring points, even if the sequence identities between them are low. The distributions of some proteins with different functions overlap. Most of the proteins used in this analysis are proteases and peptidases, which bind peptide-like molecules, and the variety of proteins is thus limited. This poor variety of proteins may cause the overlap in the distributions of different proteins.

There are a certain number of redundant proteins in the protein data set. Specifically, there are 1, 3 and 5 redundant proteins for endoribonucleases, metalloproteases and acid proteases, respectively, and 4 endoribonucleases, 7 metalloproteases and 11 acid proteases were used in the present study. Tables 7−9 show the global amino acid sequence identities and similarities of these proteins. Exactly identical proteins appear as different points in the PCA space because the local pocket structures differ depending on the ligand binding states.

The distribution of AMP/ADP/ATP binding proteins is delocalized. The CATH classification numbers shown in Table 5 indicate that the global 3D structures of these proteins are different from each other. Also, the local structures of the binding pockets are different and, as shown in Table 2, the binding mode of each complex is also different. For example, the $Mg^{2+}$ ion of casein kinase-1 (1csn) or cyclin-dependent kinase 2 (1hck) binds all three phosphates, but the $Mg^{2+}$ ion of MAP kinase ERK2 (1gol) binds the second and third phosphates but does not bind the first. In many cases, ARG or LYS residues bind phosphates, but in trimethylamine dehydrogenase (2tmd), the phosphate is not bonded by the protein but is exposed to the solvent. There are a variety of distances between the hydrophobic pocket, which binds the adenosine scaffold of AMP/ADP/ATP, and the phosphate-binding site composed of $Mg^{2+}$ ion, ARG, LYS or HIS. The chemical structures of AMP, ADP and ATP are similar to each other and differ only in the number of phosphates. However, the pharmacophores of the AMP/ADP/ATP binding proteins are different; therefore, the distribution of these proteins could be delocalized.

The robustness of this analysis was evaluated by changing the compound data set. In this case, the atomic charges of compounds were calculated by the Gasteiger method.[29,30] The PCA results did not change qualitatively due to changes in the quality of the atomic charges between the RHF/6-31G* level and the Gasteiger charges (data not shown). The compound data set was also replaced by the Coelacanth random library, which was used for screening MIF active compounds as discussed above. Figure 6b shows the PCA results obtained using 1000 compounds extracted from this random library (compound set C without the 12 active compounds). Although we expected no true binder for each protein in this random library, the PCA results did not change qualitatively. Specifically, the distributions of endoribonucleases, metalloproteases, sugar binding proteins and acid proteases are localized with each other, and the distribution of AMP/ADP/ATP binding proteins is delocalized, while the principal component axes are rotated.

**Classification of Compounds.** A total of 1006 compounds (compound set D) were analyzed using 169 proteins; the set of proteins and compounds is the same set discussed in the previous section. To analyze diversity changes by oligomerization, the data set included 20 amino acids, 400 dipeptides, 400 tripeptideps, 13 monosaccharides, and nine disaccharides. Redundant 17 COX-2 inhibitors were also included as examples of drugs. Figure 7 shows the PCA results of compounds with the first and second major coordinates. The first, second, third and fourth eigenvalues of $M^C$ in eq 3 were 73.99%, 13.87%, 1.87% and 1.30%, respectively.

The distributions of amino acids, pepetides, saccharides and COX-2 inhibitors are localized as shown in Figure 7; this localization suggests that our compound classification system works well. The distributions of AMP/ADP/ATPs are restricted to a small region, while the distribution of AMP/ADP/ATP

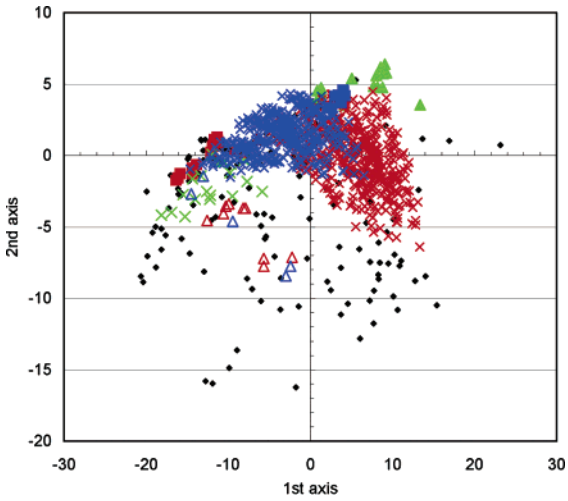


**Figure 7.** PCA plot of compound space. The green crosses, blue crosses, and red crosses represent mono-, di- and tripeptides, respectively. The red squares, blue squares, green triangles, red open triangles, blue open triangles, and dots represent monosaccharides, disaccharides, adenosine phosphates (AMP, ADP and ATP), COX-1 selective inhibitors, COX-2 selective inhibitors and other compounds, respectively.

binding proteins is wide, as shown in Figure 6. The variety of the binding pocket, which was discussed in the previous section, does not indicate variety of the ligand.

The distributions of amino acids, dipeptides and tripepetides are localized and the overlaps between these distributions are poor. Minor overlap is observed between the distributions of amino acids and dipeptides, while no overlap is observed between the distributions of amino acids and tripeptides. Furthermore, the distribution of dipeptides shows only a little overlap with that of tripeptides. Part of the chemical structure of amino acids is included in the dipeptides and tripeptides, but the pharmacophore could be different, showing the spatial distribution of the essential functional groups for protein−ligand binding. Also, the volumes of amino acids, dipeptides and tripeptides are different. The protein, which binds only a small molecule, cannot bind a larger molecule than its binding pocket. Small compounds, such as amino acids and monosaccharides, locate on the left side of PCA space while larger compounds, such as dipeptides, tripeptides and disaccharides, locate on the right. Thus, similarity of partial chemical structure does not correspond to the first or second principal axes in Figure 7. The same is true of dipeptides and tripeptides.

The physical meaning of the principal component axis was investigated, and the correlation coefficients between the total number of atoms of each compound and the principal component values were calculated. The coefficients were 0.837 for the first axis, 0.465 for the second axis, 0.176 for the third axis, 0.041 for the fourth axis, 0.085 for the fifth axis, and 0.014 for the sixth axis. Thus, the first principal component axis corresponds to the total number of atoms of each compound. The total number of hydrogen donors, the total number of hydrogen acceptors, and the solvation free energy of a given compound depends strongly on the total number of atoms of the compound, so these values are not good parameters. The correlation coefficients between the solvation free energy per atom of the compound and the principal component values were calculated and were found to be 0.132 for the first axis, 0.311 for the second axis, 0.010 for the third axis, 0.494 for the fourth axis, 0.157 for the fifth axis, and 0.084 for the sixth axis. Thus, the second and fourth principal component axes represent the solvation free energy per atom of a compound.

The distribution of tripeptides is wider than that of dipeptides or amino acid monomers, and the distribution of dipeptides is wider than that of amino acid monomers. The oligomerization of amino acids increases the diversity of the compounds.

In contrast, the diversity of disaccharides is almost the same as that of monosaccharides. These distributions are separated well, and there is no overlap between them. It is difficult to distinguish between axial and equatorial OH groups. As shown in Table 2, some proteins bind more than two kinds of saccharides. The difference among saccharides is the difference in the position of the OH groups of the saccharide. Thus, individual monosaccharides could not be distinguished well, and individual disaccharides could not be distinguished as well as monosaccharide. Thus, the diversities of mono- and disaccharides tend to be underestimated. In contrast, the volume difference is clearly distinguished, and the distributions of mono- and disaccharides are thus well separated.

The distributions of COX-1 selective inhibitors and COX-2 selective inhibitors are wide and these distributions overlap to the extent that no difference between them can be identified in the PCA plot. The compounds listed in Table 4 are all COX-2 inhibitors but some of these inhibitors inhibit COX-1 rather than COX-2, since the 3D structure of COX-2 is quite similar to that of COX-1; the sequence identity between them is 64.26%. The difference in activities between COX-1 selective inhibitor and COX-2 selective inhibitor[43] is so little that our docking software 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 COX-2 selective inhibitor.[44−46]

The robustness of the present analysis was evaluated by replacing the compound data set from charges by the RHF/6-31G* level with Gasteiger charges.[29,30] The PCA results did not change either qualitatively or qualitatively with the change in the quality of the atomic charges (data not shown).

## Conclusion

We developed a new classification method for proteins and compounds, providing the PCA of a protein−compound affinity map, which is constructed by a protein−compound docking program. Analysis of protein−ligand binding affinity could be used as a similarity search of the active compounds and the lead optimization. The first and second (or fourth) principal component values represent the total number of atoms and the solvation free energy per atom of the compound, respectively. A focused library for MIF generated by our method showed good database enrichment, which is equivalent to that obtained by an in silico screening method. If one or more active compounds are found for a target protein, the compounds that are close to the active compound in the compound space could be candidate active compounds for the target protein.

The BCUT descriptor is useful to evaluate the diversity of a compound based on the information of the compound itself. This method requires neither the 3D structure of the target protein nor the protein−ligand docking; however, to apply the BCUT method, many active compounds must be known.[3] On the other hand, the present classification method is based on the information of protein−ligand docking. Thus, when the 3D structure of the target protein is available and a conventional in silico screening method could predict one or more candidate active compound(s), this method could provide a focused library even if no active compound is known. Inversely, when some active compounds are known, this method can provide a focused library without any known 3D structure of the target protein. If

the compounds that are generated by combinatorial synthesis were plotted in the compound space, our method could evaluate the diversity of the compound library.

The drawback of our method is that it requires a large-scale protein−ligand docking simulation. This problem may be solved by the usage of the recent grid computing. However, the applications of a protein−compound affinity matrix are being developed and extended. For example, experimental observations indicate that precise $IC_{50}$ values can be estimated from the protein−compound affinity matrix.[47]

The present classification method successfully classified the protein binding pockets. Proteins having similar ligand-binding functions could form clusters in the protein space, even if the amino acid sequences were not homologous or the entire 3D folds were not similar. When the protein−compound affinity matrix is available, our method is useful in providing a new classification of proteins from a different point of view than the conventional fold classifications. The present method projected the classification result into 2-D space, while the conventional sequence-based classification methods shows the result as a dendrogram. It is easier to understand the diversity of a set of proteins in the 2-D plot rather than the dendrogram. The similarities among proteins are measured by the protein−compound affinities by the present method. Thus, the similar proteins could bind the similar ligands. If two protein clusters have similar or related functions, the distributions of these two clusters are expected to be close to each other.

The present method would not provide the optimal classifications of compounds and proteins. Other methods such as the factor rotation method can maximize the variance of the plotted data and would provide more diverse classification, in which the overlaps among protein clusters are decreased, rather than the current results.[48,49]

**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. Dr. Taku Nakahara kindly provided a list of PDB IDs of sugar binding proteins. We thank Dr. Masaya Orita (Yamanouchi Pharmaceutical Co., Ltd., Tokyo, Japan), Dr. Shigeo Fujita (Yamanouchi Pharmaceutical Co., Ltd., Tokyo, Japan) and Hideki Tsujishita (Shionogi Pharmaceutical Co., Ltd., Osaka, Japan) for their helpful suggestions.

## Appendix A

Basic protein set: The protein databank (PDB) identifier list of the basic protein set is: 1a28, 1a42, 1a4 g, 1a4q, 1abe, 1abf, 1aco, 1ai5, 1aoe, 1apt, 1apu, 1aqw, 1atl, 1b58, 1b9v, 1bma, 1byb, 1byg, 1c1e, 1c5c, 1c83, 1cbs, 1cbx, 1cdg, 1ckp, 1cle, 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, 1ldm, 1lic, 1lna, 1lst, 1mbi, 1mdr, 1 mld, 1 mmq, 1mrg, 1mts, 1mup, 1nco, 1ngp, 1nis, 1okl, 1pbd, 1pdz, 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, 4aah, 4est, 4lbd, 4phv, 5abp, 5cpp, 5er1, 6rnt and 7tim. For 1abe, 1abf, 5abp and 1htf, two protein pockets were prepared, since these proteins each bind two kinds of ligands. In addition, 1gc7, which was our target protein MIF, was also included.

Protein set A: The PDB IDs of the representative proteins of the protein set A are 1nis, 1aco, 1mdr, 1cbx, 2fox, 1mup, 1qpq, 1c83, 2ada, 1mrg, 1d3h, 4lbd, 1abf, 1lst, 1ets, 2ctc, 1pbd, 1rds, 2cmd, 2gbp, 1hsl, 1lah and 1ebg.

Protein set B: The PDB IDs of the proteins of the protein set B are 7tim, 1r55, 1okl, 1ivb, 1bqq, 2tmn, 1snc, 2qwk, 1tace, 1hsb, 1yee, 1mdr, 1fl3, 3tpi, 2ack, 1pdz, 1cbx, 2cmd, 1mld, 3cpa, 1lcp, 1qpq, 4aah, 1ldm and 1pbd.

## References

(1) Ghose, A. K., Viswanadhan, V. N., Eds. *Combinatorial library design and evaluation − principle, software tools, and applications in drug discovery*; Marcel Dekker: New York, 2001; pp 337−362.
(2) Pickett, S. In *Protein−ligand interactions from molecular recognition to drug design − methods and principles in medicinal chemistry*; Boehm, H. J., Schneider, G., Mannhold, R., Kubinyi, H., Folkers, G., Eds.; Wiley-VCH: Weinheim, 2003; pp 88−91.
(3) Pearlman, R. S.; Smith, K. M. Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Compt. Sci.* **1999**, *39*, 28−35.
(4) Kinoshita, K.; Nakamura, H. Protein informatics towards function identification. *Curr. Opin. Struct. Biol.* **2003**, *13*, 396−400.
(5) Kinoshita, K.; Nakamura, H. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci.* **2003**, *12*, 1589−1595.
(6) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of functional sites in protein structures. *J. Mol. Biol.* **2004**, *339*, 607−633.
(7) Schmitt, A.; Kuhn, D.; Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **2002**, *323*, 387−406.
(8) Schmitt, S.; Hendlich, M.; Klebe, G. Development of a database for protein cavities and its usage for similarity searches in binding sites. Rational Approaches to Drug Design.; Höltje, H.-D., Sippl, M., Eds.; Prous Science 2001; 135−141.
(9) Porter, C. T.; Bartlett, G. J.; Thornton, J. M. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.* **2004**, *32*, D129−D133.
(10) Holm, L.; Sander, C. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.* **1997**, *25*, 231−234.
(11) Heger, A.; Wilton, C. A.; Sivakumar, A.; Holm, L. ADDA: a domain database with global coverage of the protein universe. *Nucleic Acids Res.* **2005**, *33*, D188−D191.
(12) George, R. A.; Spriggs, R. V.; Thornton, J. M.; Al-Lazikani, B.; Swindells, M. B. SCOPEC: a database of protein catalytic domains. *Bioinformatics* **2004**, *20*, i130−i136.
(13) Jones, S.; Thornton, J. M. Searching for functional sites in protein structures. *Curr. Opin. Chem. Biol.* **2004**, *8*, 3−7.
(14) Choi, I. G.; Kwon, J.; Kim, S. H. Local feature frequency profile: A method to measure structural similarity in proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *16*, 3797−3802.
(15) Hou, J.; Sims, G. E.; Zhang, C.; Kim, S. H. A global representation of the protein fold space. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 2386-2390.
(16) Paul, N.; Rognan, D. ConsDock: A new program for the consensus analysis of protein−ligand interactions. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 521−533.
(17) 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: Struct., Funct., Genet.* **1998**, *33*, 367−382.
(18) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian Docking Functions. *Biopolymers* **2003**, *68*, 76−90.
(19) Goodsell, D. S.; Olson, A. J. Automated docking of substrates to proteins by simulated annealing. *Proteins: Struct., Funct., Genet.* **1990**, *8*, 195−202.
(20) Taylor, J. S.; Burnett, R. M. DARWIN: A program for docking flexible molecules. *Proteins: Struct., Funct., Genet.* **2000**, *41*, 173−191.
(21) 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.
(22) Colman, P. M. Structure-based drug design. *Curr. Opin. Struct. Biol.* **1994**, *4*, 868−874.
(23) 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.
(24) 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.

(25) 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.
(26) Fukunishi, Y.; Mikami, Y.; Nakamura, H. Similarities among receptor pockets and among compounds: analysis and application to in silico ligand screening. *J. Mol. Graphics Modeling* **2005**, in press.
(27) 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: Struct., Funct., Genet.* **2002**, *49*, 457−471.
(28) 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.
(29) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(30) Gasteiger, J.; Marsili, M. A new model for calculating atomic charges in molecules. *Tetrahedron Lett.* **1978**, 3181−3184.
(31) 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.
(32) Pearson, W. R.; Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 2444−2448.
(33) Pearson, W. R. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **1990**, *183*, 63−98.
(34) Dennesiouk, K. A.; Johnson, M. When fold is not important: a common structural framework for adenine and AMP binding in 12 unrelated protein families. *Proteins: Struct., Funct., Genet.* **2000**, *38*, 310−326.
(35) Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. CATH − a hierarchic classification of protein domain structures. *Structure* **1997**, *5*, 1093−1108.
(36) Pearl, F. M. G.; Lee, D.; Bray, J. E.; Sillitoe, I.; Todd, A. E.; Harrison, A. P.; Thornton, J. M.; Orengo, C. A. Assigning genomic sequences to CATH. *Nucleic Acids Res.* **2000**, *28*, 277−282.
(37) Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr.* **2002**, *B58*, 380−388.
(38) 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.
(39) 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.
(40) Vigers, G. P. A.; Rizzi, J. P. Multiple active site corrections for docking and virtual screening. *J. Med. Chem.* **2004**, *47*, 80−89.
(41) Ghose, A. K., Viswanadhan, V. N., Eds. *Combinatorial library design and evaluation − principle, software tools, and applications in drug discovery*. Marcel Dekker: New York, 2001; pp 478−497.
(42) Miller, M.; Schneider, J.; Sathyanarayna, B. K.; Toth, M. V.; Marshall, G. R.; Clawson, L.; Selk, L.; Kent, S. B.; Wlodawer, A. Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 Å resolution. *Science* **1989**, *246*, 1149−1152.
(43) 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.
(44) 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. *Nature Struct. Biol.* **1996**, *3*, 927−933.
(45) 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.
(46) 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.
(47) 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.
(48) Cattel, R. B. The scree test for the number of factors. *Multivariate Behavioral Res.* **1966**, *1*, 245−276.
(49) Abdi, H. In *Encyclopedia for research methods for the social science*; Lewis-Beck, M., Futing, T., Eds; Sage: Thousand Oaks, CA, 2003; pp 978−982.

JM050480A