Available online at www.sciencedirect.com

ScienceDirect

European Journal of Medicinal Chemistry 42 (2007) 966–976

EUROPEAN JOURNAL OF MEDICINAL CHEMISTRY

http://www.elsevier.com/locate/ejmech

Original article

In-silico drug screening method based on the protein–compound affinity matrix using the factor selection technique

Sukumaran Murali [a], Shinichi Hojo [b], Hideki Tsujishita [a], Haruki Nakamura [c,d], Yoshifumi Fukunishi [b,c,*]

[a] Japan Biological Information Research Center (JBIRC), Japan Biological Informatics Consortium (JBIC), 2-41-6, Aomi, Koto-ku, Tokyo 135-0064, Japan
[b] Department of Chemistry, Graduate School of Science and Engineering, Tokyo Metropolitan University,
1-1, Minamiosawa, Hachiouji, Tokyo 192-0397, Japan
[c] Biological Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST),
2-41-6, Aomi, Koto-ku, Tokyo 135-0064, Japan
[d] Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

Received 15 September 2006; received in revised form 14 December 2006; accepted 21 December 2006
Available online 13 January 2007

Abstract

We have developed a new in-silico drug screening method, a modified version of a docking score index (DSI) method, based on a protein–compound docking affinity matrix. By using this method, the docking scores are converted to the docking score indexes by the principal component analysis (PCA) method and each compound is projected into a PCA space. In this study, we propose a method to select a set of suitable principal component axes and evaluate the database enrichment for 12 target proteins. This method selects the new active compounds or hits, which are close to the known active compounds, thereby enhancing the database enrichment.

Keywords: Principal component analysis; Docking score; Protein–compound interaction matrix; Docking score index method

1. Introduction

The primary step in in-silico (virtual) screening of the modern drug discovery is to select the subsets or hits from millions of chemical compounds collected in either commercial or in-house databases. This screening is based on two different approaches in hit identification: (1) structure-based virtual screening [1,2], in which compounds are docked into the active site of a target and ranked; and (2) ligand-based virtual screening [3], in which the new hits are identified based on the similarity of known active compounds. The new active compounds derived from these methods will be further tested biologically using high-throughput screening (HTS) methods. As the number of compounds increases exponentially in the commercially available databases, it not only poses a major challenge in prioritizing the focused hits suitable for the target of particular interest but also would circumvent the refined selection of the hits, which will be considered for the further studies in the drug discovery hierarchy. Therefore, high enrichment in-silico screening method is needed to find the new active compounds for many targets associated with diseases.

There are many methods available for in-silico screening enrichment [4–12] of large chemical compound databases. A number of docking programs [13–23] have been developed to predict the protein–ligand binding free energies, but still the binding free energy estimation has an error of about 2–3 kcal/mol [17,23]. The low accuracy of binding free energy or the docking score will result in low database enrichment

* Corresponding author. Biological Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6, Aomi, Koto-ku, Tokyo 135-0064, Japan. Tel.: +81 3 3599 8290; fax: +81 3 3599 8099.

E-mail address: y-fukunishi@jbirc.aist.go.jp (Y. Fukunishi).

0223-5234/$ - see front matter © 2007 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2006.12.028

in in-silico screening. The problem is difficult to overcome by improving the docking score itself [24,25]. Some of the methods to overcome this problem are based on the protein–compound affinity matrix [23,26–28].

The protein–compound affinity matrix is obtained by thorough docking calculations between a set of many protein pockets and chemical compounds. Once the affinity matrix has been obtained, it can be reused for different target proteins. If the 3D structure of the target protein is available, we can apply the multiple active site correction (MASC) score [26], receptor selection (RS) method [25], and multiple target screening (MTS) methods [29]. If the 3D structure of the target protein is unavailable, we can apply the affinity fingerprint approach, which is a new type of similarity search method based on the multi-proteins–multi-compounds affinity matrix. In a study by Kauvar et al., who developed this approach, the IC_{50} value of the target protein was estimated from the IC_{50} values of many other proteins [29]. Later, the protein–compound docking score was used as the descriptor of the compound instead of the usual 1D or 2D descriptor, namely, the mass weight, number of rotatable bonds and number of hydrogen donors/acceptors of the compound, etc. [30–32]. To apply this approach, the compound library should include at least one active compound. On the basis of the docking score, if two compounds bind the same proteins they are identified as similar compounds. A compound which is similar to a known active compound of the target protein of interest could be a candidate-hit compound.

A chemometric-based approach has been widely used in virtual screening for compound selection and clustering [33–36]. The principal component analysis (PCA) method is widely used to reduce the noise and extract characteristic features of the multi-dimensional data. In our previous study [28], the docking score index (DSI) method, which utilizes the PCA to reduce the computational error of the protein–compound affinity matrix, was proposed to distinguish active compounds from negative compounds of the macrophage migration inhibitory factor (MIF). In that study, it was revealed that the active compounds were localized in the PCA space of compounds, while the negative compounds showed a wide distribution. In the PCA space, the compounds in a sphere whose center was set to the average position of the known compounds were selected as a focused library whose database enrichment was equivalent to or better than that obtained by a structure-based in-silico screening method.

By using the DSI method, the protein–compound docking scores are converted to the docking score indexes by the PCA method, and each compound is projected into a PCA space. The DSI method projects all the compounds onto a subspace of the PCA space to reduce the computational noise and selects the compound which is close to the known active compound. In the DSI method, usually 5–10 major principal components (PCs) are selected to form the subspace. It should be noted that the database enrichment by the DSI method depends on the number of principal PCs used. Also, in some cases, the distribution of the active compounds does not localize in the major principal component axis but in the minor principal component axis, because the major principal component carries the information about the absolute values of the docking scores, but does not necessarily carry information about how to divide between active and inactive compounds. Thus, it is essential to select suitable PCs, in which the active compounds are well separated from the inactive compounds. Here, we propose a method to select a set of suitable principal component axes automatically and evaluate the database enrichment for 12 target proteins. For all the targets investigated in this study, this modified DSI method outperformed the original DSI method. These results suggest that the current method is useful in finding new active compounds from the database containing the known active compounds.

2. Methods

2.1. Factor selection docking score index (FS-DSI) method

Our in-silico drug screening method is a sort of similarity search based on known active compounds. A measure to represent the distance between two compounds was determined based on the protein–ligand interaction matrix, each element of which is the corresponding docking score. From the covariance matrix of compounds, a PCA is performed to find similar clusters of compounds [28].

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

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

$$M_{ij}^{\mathrm{P}} = \frac{1}{N_c}\sum_{k=1}^{N_c}\left(s_i^k - \bar{s}_i\right)\left(s_j^k - \bar{s}_j\right) \quad (1)$$

and

$$\bar{s}_i = \frac{1}{N_c}\sum_{k}^{N_c} s_i^k, \quad (2)$$

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

$$X_k = \sum_{j=1}^{\mathrm{Nr}} c_j^k \phi_j. \quad (3)$$

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

The candidate-hit compounds were selected using the following method. In the PCA space, the compounds, which are close to the known active compounds, were selected as the candidate-hit compounds. In the original version of the DSI method, the distance from the k-th compound to the average position of the active compounds ($\overline{c_j}$) is defined as

$$D_k = \sqrt{\sum_{j=1}^{N_{\text{select}}} \left(c_j^k - \overline{c}_j\right)^2} \tag{4}$$

and

$$\overline{c}_j = \sum c_j^{\text{active}} / N_{\text{a}}, \tag{5}$$

where c_j^{active} and N_{a} are the DSI values of the active compounds and the total number of active compounds.

The standard deviations (σ) of the DSI values were calculated for each axis, and the DSI values, whose distance from the origin was more than 5σ, were removed from the analysis. We adopted a "standard Euclidian distance", namely, the DSI values were scaled to set the standard deviation of the distribution of compounds of each axis to 1. The method which is composed of these procedures is the so-called DSI method. In this study, N_{select} in Eq. (4) is set as 10.

In this study, the suffix j runs over the selected axes $\{\alpha_1, \alpha_2, \ldots, \alpha_{N_{\text{select}}}\}$ in Eq. (6), and the next modified distance D_k' is introduced.

$$D_k' = \sqrt{\sum_{j=\left\{\alpha_1, \alpha_2, \ldots, \alpha_{N_{\text{select}}}\right\}} \left(c_j^k - \overline{c}_j\right)^2}. \tag{6}$$

The principal component axes are selected in the manner described as follows. The contribution of each principal component is estimated by using the database enrichment curve. The surface area under the database enrichment curve q_α is evaluated for the α-th principal component axis, namely the suffix j in Eq. (6) is set as α and N_{select} is set as 1, and the database enrichment curve f_α is calculated for the α-th axis. The q_α value is calculated by

$$q_\alpha = \int_0^{100} f_\alpha(x)\text{d}x, \tag{7}$$

where x and f_α (x) are the number of compounds (%) selected from the total compound library and the database enrichment curve, respectively. The higher q_α value corresponds to the better database enrichment, and the q_α value is greater than zero and less than 100. For the random screening, $q_\alpha = 50$.

The axes are sorted in descendant order with respect to the q_α value. The surface area (q) under the total database enrichment curve (f) is a measure of the database enrichment as well as q_α in Eq. (7).

$$q = \int_0^{100} f(x)\text{d}x. \tag{8}$$

The q value is calculated by changing the number of used axes (N_{select}) in Eq. (6) to find the optimal N_{select} value, which gives the maximum q value.

Protein–compound docking simulation was performed by the program Sievgene [23], which is a protein–ligand flexible docking program for in-silico drug screening. This program generates many conformers (100 conformers by default) for each compound and keeps the target protein structure rigid, but with the soft interaction forces adapting its slight structural change to some extent [23]. This docking program was developed with a performance yielding of about 50% of the reconstructed complexes at a distance of less than 2 Å RMSD for the 132 complexed receptors with the compounds in PDB [23]. The predicted results by our program were almost the same as those by other docking programs; we expected that the results obtained by the other docking programs show the same trend as the results obtained by our docking program [37]. Our docking program Sievgene, which is part of the myPresto system, is available on the web site (http://www.jbic.or.jp/activity/st_pr_pj/mypresto/index_mypr.html) and is free for academic use.

3. Preparation of materials

To evaluate our method, we performed a protein–compound docking simulation based on the soluble protein structures registered in the Protein Data Bank (PDB). Since the ligand pockets were clearly determined in this database, the protein–ligand complex structures were suitable for the docking study. Of 180 protein–ligand complexes used in this evaluation, 142 complexes were selected from the database used in the evaluation of GOLD and FlexX [37] and the other 38 additional complexes were selected from the PDB. The former data set contains a rich variety of proteins and compounds whose structures were all determined by high quality experiments with a resolution of less than 2.5 Å. Almost the entire atom coordinates are supplied except the hydrogen atoms, and all of the atomic structures around the ligand pockets are quite reliable. Thus, this data set was used in the clustering analysis of proteins and in-silico screening. From the original data set, the complexes containing a covalent bond between the protein and ligand were removed, since our docking program cannot perform the protein–ligand docking when a covalent bond exists between the protein and the ligand. The set of other 38 structures contains the human immunodeficiency virus protease-1 (HIV), cyclooxygenase-2 (COX-2), and glutathione S-transferase (GST). The PDB identifiers are summarized in Appendix A. All water molecules and cofactors were removed from the proteins, and all missing hydrogen atoms were added to form the all-atom models of proteins.

Four subsets of proteins were selected from the whole 180 proteins by the clustering method [23]. The whole 180 proteins set was named protein set A. The four subsets were named protein sets B, C, D and E, and these sets consisted of 123, 93, 63 and 24 proteins, respectively. The lists of the PDB codes of the four subsets are summarized in Appendix A.

Our target proteins are the macrophage migration inhibitory factor (MIF), COX-2, HIV, thermolysin, GST, histamine H1 receptor, adrenaline beta receptor, serotonin receptor, and dopamine D2 receptor. These target proteins and the number of their active compounds are summarized in Table 1. The active compounds of the histamine H1 receptor [38], adrenaline beta receptor [39], serotonin receptor[40], and dopamine D2 receptor [41] were selected from the literature. The compound set consists of a total of 166 active compounds and 11 050 potential-negative compounds of the Coelacanth chemical compound library (Coelacanth Corporation, East Windsor, NJ, USA), which is a random library. Usually only one hit compound is found out of 10^4 randomly selected compounds, so we would expect that there is no or only a few hit compounds out of 11 050 compounds. The active compounds of MIF are depicted in Fig. 1 and the other 152 active compounds are listed in Appendix B. In Fig. 1, compounds **7** and **12** were selected from the PDB, and compounds **1**, **3**, **4**, **6** and **8** were reported in a previous study [42]. The others (compounds **2**, **5**, **9**, **10**, and **11**) were prepared in our previous study [28]. Compounds **13** and **14** are D-dopachrome and 5,6-dihydroxyindole-2-carboxylic acid (DHICA), which are the native ligand of MIF, respectively [42]. The size distributions of the compounds used are summarized in Table 2.

The 3D coordinates of the 11 050 random compounds mentioned above were generated by the Concord program (Tripos, St. Louis, MO, USA) from the 2D Sybyl SD files provided by the Coelacanth Chemical Corporation. The 3D coordinates of the inhibitors were generated by the Chem3D (Cambridge Software, Cambridge, MA, USA). The atomic charges of each ligand were determined by the Gasteiger method [43,44]. The atomic charges of the proteins were the same as the atomic charges of AMBER parm99 [45].

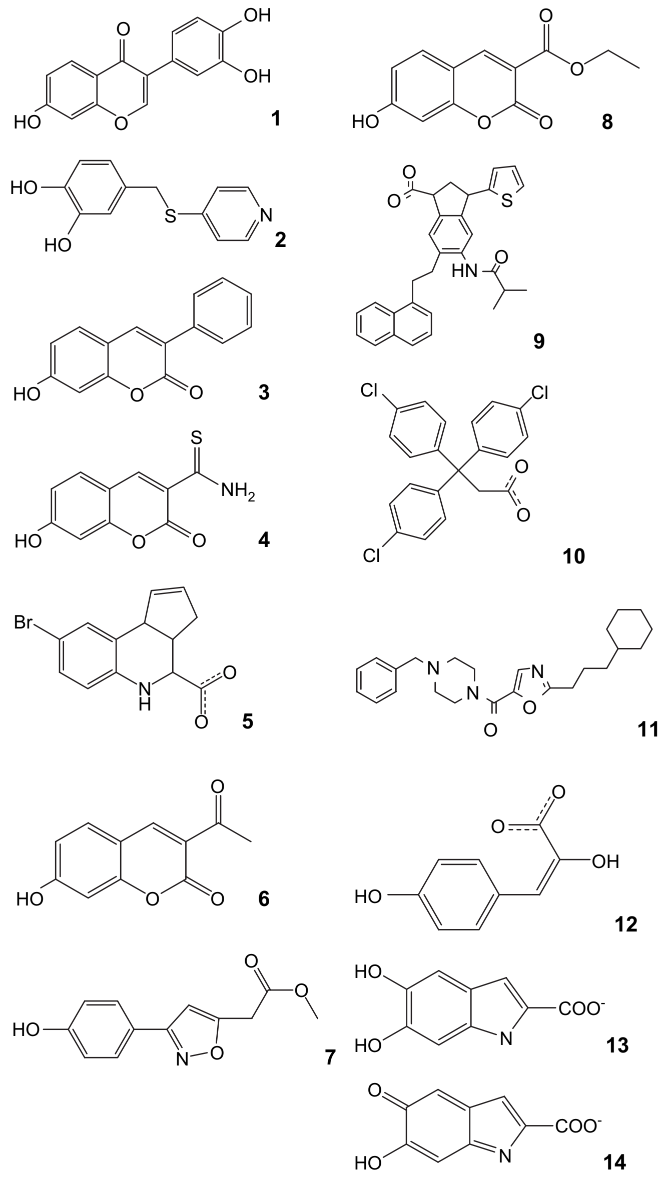

Fig. 1. MIF active compounds.

Table 1
List of target proteins and their active compounds

Target protein name	PDB codes	Number of active compounds
Macrophage migration inhibitory factor	1gcz	14
COX-2	1cx2, 1pxx, 3pgh, 4cox, 5cox and 6cox	14
HIV	1aid, 1hpx and 1ivp	20
Thermolysin	2tmn	28
GST	18gs, 2gss and 3pgt	12
Histamine H1 receptor (antagonist)	None	10
Adrenaline beta receptor (agonist)	None	12
Adrenaline beta receptor (antagonist)	None	13
Serotonin receptor (agonist)	None	8
Serotonin receptor (antagonist)	None	9
Dopamine D2 receptor (agonist)	None	6
Dopamine D2 receptor (antagonist)	None	15

4. Results

The docking score index (DSI) and factor selection-DSI (FS-DSI) methods were applied to 12 targets, which are listed in the previous section, using a different number of protein sets such as A (180 proteins), B (123 proteins), C (93 proteins), D (63 proteins) and E (24 proteins). The names of the protein sets are listed in Appendix A.

To evaluate the efficiency of this method, the Jack-knife test was applied: the active compounds of each target protein were divided into two sets, the set of known active compounds for factor selection and the set of hidden active compounds, which should be found by the software. For each target protein, the sets containing active compounds were divided into half. Ten pairs of these active compound sets were prepared for each target protein. Thus, a total of 120 (=12 targets × 10 trials) database enrichment curves were calculated for the 12 target proteins and the results were averaged.

The surface area of the database enrichment curve for each axis, q_α in Eq. (7), against the number of principal component (PC) axes is plotted in Fig. 2a. This figure enables information

Table 2
Molecular size distribution of the compound group

Size	Library	MIF	COX-2	Ther	HIV	GST	Hant	Aago	Aant	Sago	Sant	Dago	Dant
0–19	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
20–29	0.5	58.3	10.0	16.0	0.0	0.0	0.0	16.7	0.0	37.5	0.0	0.0	0.0
30–39	0.5	16.7	70.0	36.0	0.0	25.0	10.0	50.0	7.7	0.0	11.1	50.0	0.0
40–49	6.5	8.3	20.0	4.0	0.0	8.3	60.0	33.3	61.5	50.0	44.4	33.3	40.0
50–59	22.5	0.0	0.0	12.0	10.0	33.3	20.0	0.0	30.8	12.5	33.3	0.0	46.7
60–69	40.4	16.7	0.0	28.0	0.0	33.3	10.0	0.0	0.0	0.0	11.1	0.0	13.3
70–79	22.1	0.0	0.0	4.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
80	7.4	0.0	0.0	0.0	70.0	0.0	0.0	0.0	0.0	0.0	0.0	16.7	0.0

The number of atoms of the compound represents the molecular size. The numbers in this table are in %. Ther: inhibitors of thermolysin; Hant: histamine H1 receptor antagonists; Aago: adrenalin beta receptor agonists; Aant: adrenalin beta receptor antagonist; Sago: serotonin receptor agonist; Sant: serotonin receptor antagonist; Dago: dopamine D2 receptor agonist; Dant: dopamine D2 receptor antagonist.

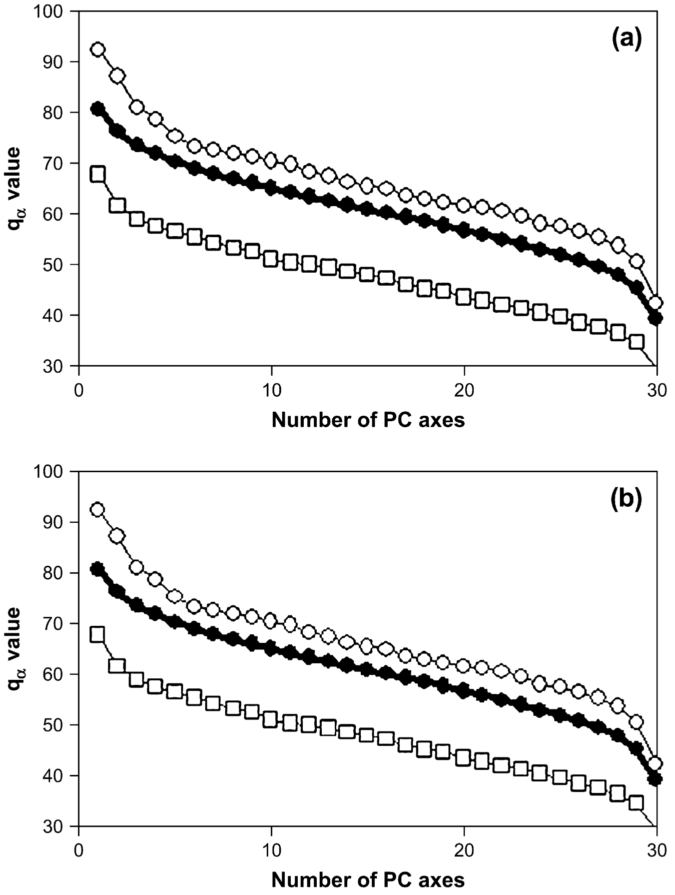

Fig. 2. The q_α and q values using the affinity matrix of 180 proteins (protein set A). (a) Averaged q_α value in Eq. (7) and q_α values for selected targets. Filled circles, open squares and open circles represent the averaged q_α value of the 12 targets, q_α value of GST and q_α value of histamine H1 receptor antagonist, respectively. (b) Averaged q value in Eq. (8) and q values for selected targets. Filled circles, open squares and open circles represent the average q value of the 12 targets, the q value of HIV protease-1 and the q value of the dopamine D2 receptor antagonist, respectively.

on the dependency of q_α on the target used. In Fig. 2, only the major top 30 PCs were used for the analysis, because the minor PCs carry only a little information due to computational noise. The q_α value was calculated for each PC, and the PCs and q_α values were sorted in descendant order in Fig. 2a. The q_α values were calculated for all the targets, and for each target the q_α value was averaged from the 10 q_α values obtained by the Jack-knife test.

Fig. 2a shows the averaged q_α value of 12 targets and the q_α value of GST with the lowest q_α value for $\alpha = 1$ among the 12 targets, while the histamine H1 receptor shows the highest q_α value for $\alpha = 1$ among the 12 targets. The trends of the q_α values of the other proteins were similar to these results. On average, among the 30 major PCs, 26 PCs show a q_α value greater than 50%. In addition, the number of PCs which show a q_α value greater than 50% depends on the target. For GST, 12 PCs show a q_α value greater than 50%, and for the histamine H1 receptor, 29 PCs show a q_α value greater than 50%.

The surface area (q) under the total database enrichment curve is plotted against the number of selected PC axes in Fig. 2b. In Fig. 2b, the averaged q value of 12 targets and the q values of the HIV inhibitor show a lower optimal q value, and the q value of the dopamine D2 receptor antagonist shows a higher optimal q value. On average, the q value reaches the maximum with the number of selected PCs to 8. The q value reaches the maximum when the number of selected PCs = 1 for HIV, and the q value reaches the maximum when the number of selected PCs = 13 for the dopamine D2 antagonist. This suggests that the optimal number of selected PCs depends on the target.

The q values for all the targets using different protein sets, obtained by both the DSI and FS-DSI methods, are compiled in Table 3, which also contains the average q values for both methods. These values suggest that the FS-DSI method outperforms the DSI method for all the targets regardless of the number of proteins used. The q values for GST and HIV were drastically improved by the FS-DSI method. The q values by the DSI method showed a wide distribution from 35.5 to 88.8 for protein set A, from 31.2 to 87.1 for set B, from 34.6 to 85.9 for set C, from 36.2 to 87.3 for set D, and from 31.3 to 90.1 for set E. In contrast, the q values by the FS-DSI method showed a narrow distribution from 71.3 to 98.5 for protein set A, from 70.9 to 99.3 for set B, from 70.8 to 98.8 for set C, from 73.4 to 98.8 for set D, and from 57.7 to 96.5 for set E. This result shows that the FS-DSI method is more robust than the DSI method.

Fig. 3a and b shows the PCA results for 20 HIV active compounds and the other 11 191 potential-negative compounds by the DSI and FS-DSI methods, respectively. In Fig. 3a, the active compounds are not localized in the PCA space. However, in Fig. 3b, the active compounds are localized in the PCA space. Thus the factor selection method worked well to select

Table 3
The q values of 12 targets and their average by using the DSI and FS-DSI methods

No. of proteins	180 (Protein set A)		123 (Protein set B)		93 (Protein set C)		63 (Protein set D)		24 (Protein set E)	
Target	DSI	FS-DSI	DSI	FS-DSI	DSI	FS-DSI	DSI	FS-DSI	DSI	FS-DSI
MIF	63.5	88.4	64.8	80.3	60.0	80.3	58.3	81.0	53.5	78.9
Ther	71.6	84.5	68.6	89.4	66.8	86.8	67.1	84.7	65.5	80.8
GST	35.5	71.3	31.2	70.9	34.6	70.8	36.2	73.4	31.3	57.7
HIV	37.9	78.6	35.5	79.1	38.4	78.5	38.2	76.1	34.1	67.8
COX-2	77.6	92.6	73.5	88.5	80.9	91.4	82.3	89.8	83.0	90.4
Hant	83.3	97.6	81.1	98.3	82.8	98.8	82.7	94.8	85.7	92.0
Aago	88.8	97.9	87.1	96.9	85.9	97.1	87.3	96.8	90.1	95.8
Aant	85.1	94.0	82.2	95.5	81.5	93.0	83.6	93.7	80.0	87.6
Sago	82.2	98.5	79.1	99.3	76.2	98.5	78.9	98.8	80.3	94.1
Sant	87.2	96.7	85.0	96.6	83.7	95.7	83.1	98.5	87.1	96.5
Dago	66.7	93.4	67.3	97.9	66.5	90.4	67.4	97.8	69.3	92.1
Dant	80.2	92.2	80.2	90.8	79.7	89.8	80.2	86.6	73.1	88.5
Average	71.6	90.5	69.6	90.3	69.8	89.3	70.4	89.3	69.4	85.2

MIF: macrophage migration inhibitory factor; Ther: thermolysin; GST: glutathione S-transferase; HIV: HIV protease-1; COX-2: cyclooxygenase-2; Hant: antagonists of histamine H1 receptor; Aago: agonists of adrenaline beta receptor; Aant: antagonists of adrenaline beta receptor; Sago: agonists of serotonin receptor; Sant: antagonists of serotonin receptor; Dago: agonists of dopamine D2 receptor; Dant: antagonists of dopamine D2 receptor.

the appropriate PC axes on which the active compounds are localized.

Figs. 4–8 show the average database enrichment curves, which are the average of the 10 database enrichment curves of 12 targets using the affinity matrix of different protein sets, the PDB codes of which are listed in Appendix A.

When using a larger number of protein containing sets such as set A (180 proteins), the database enrichment by the FS-DSI method was significantly higher than that obtained by the original DSI method. The important part of the database enrichment curve is the slope around the origin of the axis, since the purpose of the in-silico screening is to select a small number of compounds from the large number of compounds of the library. The slope by the FS-DSI method is much better than that by the DSI method. With the protein set A, 12.4% and 43.4% of the active compounds were found within the first 1% of the database by the DSI and by the FS-DSI method, respectively. The average q value by the FS-DSI method was 90.5, which is better than the q value of 71.6 by the DSI method.

With 123 proteins (set B), 93 proteins (set C) and 63 proteins (set D), the database enrichment results from both the methods are comparable to that of using 180 proteins (set A). A total of 10.8%, 10.9% and 10.3% of the active compounds were found within the first 1% of the database by the DSI method for the protein sets B, C and D, respectively. A total of 41.6%, 43.0% and 43.0% of the active compounds were found within the first 1% of the database by the FS-DSI method for the protein sets B, C and D, respectively. The average q values obtained by the FS-DSI method were 90.3, 89.3 and 89.3 for the protein sets B, C and D, respectively; and by the DSI method were 69.6, 69.8 and 70.4 for the protein sets B, C and D, respectively. In contrast, when using 24 proteins (set E), the database enrichment by the FS-DSI method was drastically decreased; that is, 14.2% and 26.5% of the active compounds were found within the first 1% of the database by the DSI and FS-DSI methods, respectively. However, the average q values 69.4 and 85.2, obtained by the DSI and FS-DSI methods, are slightly lower than that obtained for the other protein sets.

5. Discussion

As the matrix dimension indicates the number of used proteins, the total number of PCs was 180 for protein set A (180 proteins). Fig. 2b shows that one can project the PC space into a 10-dimensional subspace in order to achieve maximum database enrichment. There are only 10 important PCs among the 30 major PCs. As the protein sets A (180 proteins), B (123 proteins), C (93 proteins) and D (63 proteins) provide more than 30 PCs, only the 10 important PCs would be calculated. It is possible that some of the most important of the 10 PCs would not be obtained from the total 24 PCs with the protein set E (24 proteins), if the important PCs were between the 25th and 30th PC. These results are consistent with the fact that the database enrichment by set E is worse than that by sets A, B, C and D, and the database enrichments by sets A, B, C and D are almost equivalent.

Our previous work [28] showed that the first and second PCs represent the molecular size and solvation free energy of the compound, respectively. In the current study, the optimal number of selected PCs was more than four for 11 targets, but only one for HIV, as shown in Fig. 2b. This means that the important PCs carry additional information besides the molecular size and the solvation free energy of compounds. In Table 3, the q values of GST and HIV are very low with 35.5 and 37.9 by the DSI method; in contrast, the values are drastically improved by the FS-DSI method to 71.3 and 78.6. The DSI method adopted the 10 major PCs, and the inhibitors of GST and HIV were not localized in this 10-dimensional compound space, since the database enrichment represents the localizability of inhibitors. As shown in Fig. 3a, the active compounds of HIV formed two clusters and the average position of these active compounds was not located in these two clusters. In contrast, as shown in

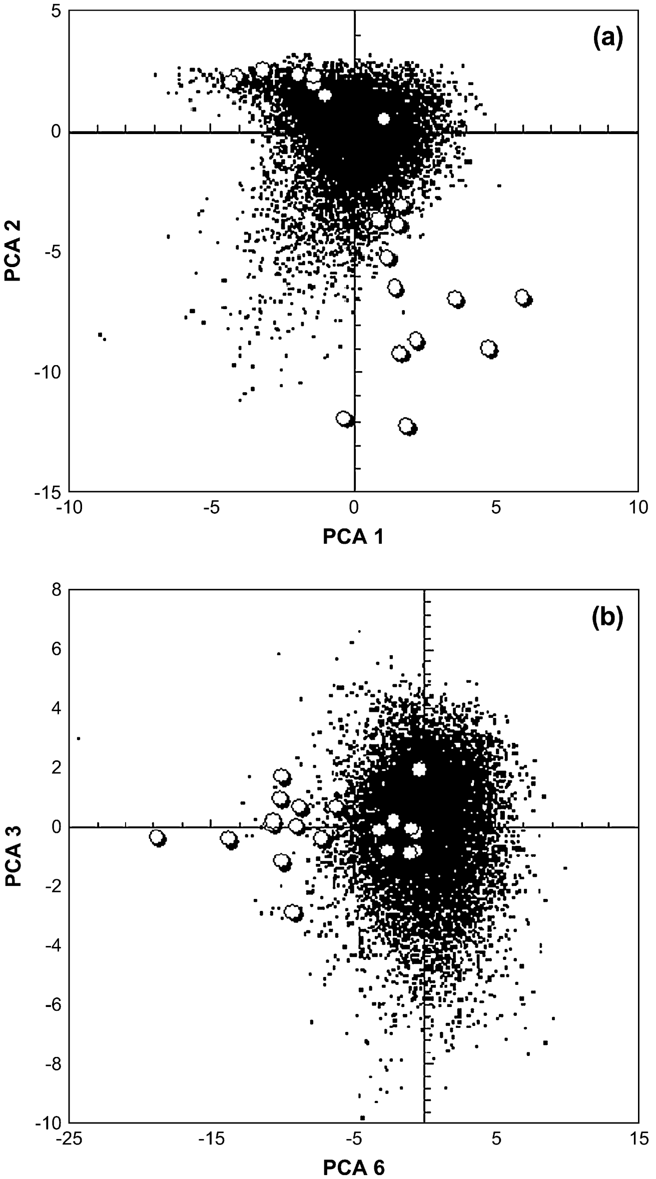

Fig. 3. PCA results of HIV active compounds and the other compounds. The open circles and black dots represent the HIV active compounds and the other compounds, respectively. (a) PCA result by the DSI method. “PCA 1” and “PCA 2” represent the first and second principal component axes. (b) PCA result by the FS-DSI method. “PCA 6” and “PCA 3” represent the sixth and third principal component axes, respectively, which gave the best and the second best q_α values.

Fig. 3b, in the subspace generated by the FS-DSI method, these inhibitors can form one cluster and the average position of the inhibitors was located in the cluster. This suggests the drastic improvement of the database enrichment by the FS-DSI method.

The docking score of the DSI carries information about the shape and physical complementarities between the protein pocket and compound, namely, a small or large pocket likely binds a small or large ligand, a hydrophobic or hydrophilic pocket likely binds a hydrophobic or hydrophilic ligand, a positively charged or negatively charged pocket binds a negatively charged or positively charged ligand, and so on. There are various properties of shape and physical properties of protein pockets, and because of these properties, many PCs carry information that could be important for drug screening.

In the current study, the PCA was used to eliminate redundant information about the protein–compound affinity matrix and the important variables (PCs) were selected. There are alternative methods, such as the unsupervised forward selection (UFS) used in QSAR [46]. The purpose of the UFS is similar to that of the FS-DSI method; the UFS could be applied to the protein–compound affinity matrix. Also, the genetic algorithm (GA) could be applied to select the combination of

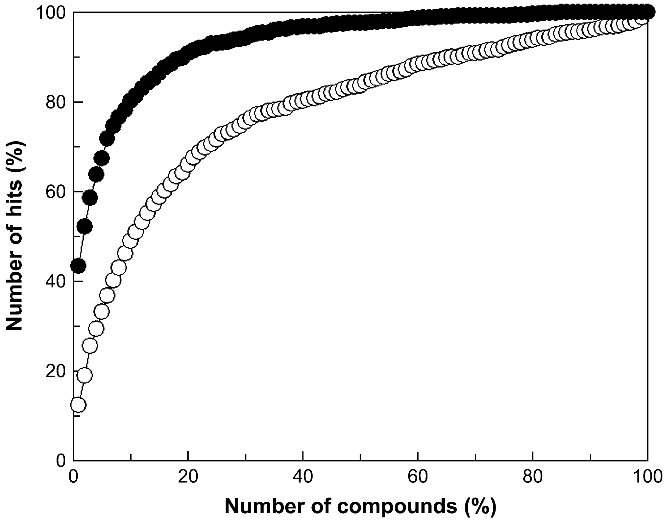

Fig. 4. Averaged database enrichment curves of 12 targets using the affinity matrix of 180 proteins (protein set A). Open circles and filled circles represent the averaged database enrichments by the DSI method and that by the FS-DSI method, respectively.

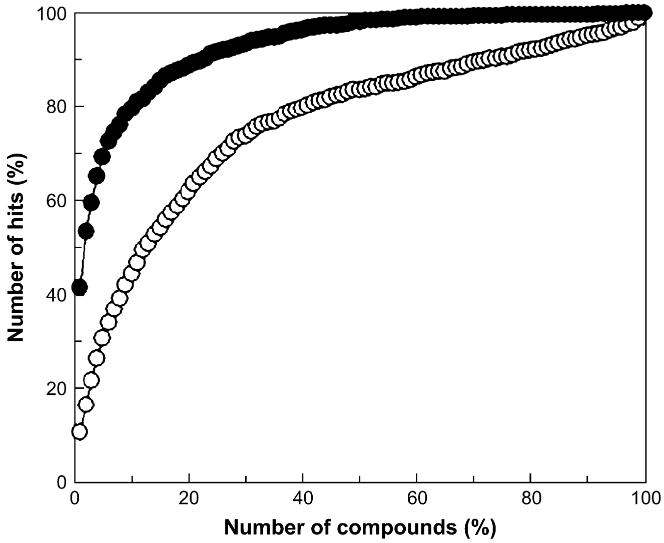

Fig. 5. Averaged database enrichment curves of 12 targets using the affinity matrix of 123 proteins (protein set B). Open circles and filled circles represent the averaged database enrichments by the DSI method and that by the FS-DSI method, respectively.

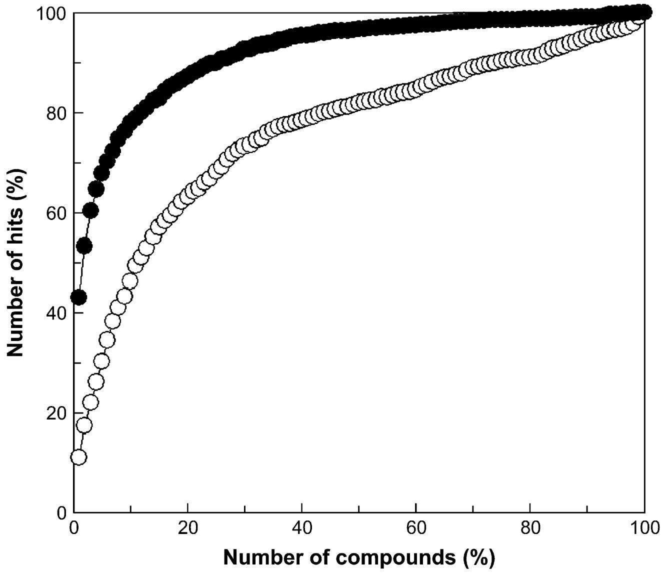

Fig. 6. Averaged database enrichment curves of 12 targets using the affinity matrix of 93 proteins (protein set C). Open circles and filled circles represent the averaged database enrichments by the DSI method and that by the FS-DSI method, respectively.

important PCs which give the best q value, because the q value in Eq. (8) is not a linear combination of q_α in Eq. (7), and the combination of PCs which give the best q value is not always the same combination of PCs which give the best q_α values.

The theoretical framework of the FS-DSI method, namely PCA followed by factor selection, can be applied to other chemometric methods, since the docking scores of each compound correspond to the compound descriptors. The most time consuming step is the diagonalization of the matrix M^P in Eq. (1). For example in this study, each screening took about 9 min on a 500 MHz SGI origin 300 computer. In addition, the FS-DSI method can be combined with the ordinary chemometric method by simply joining the score vector and the fingerprint vector by the other chemometric method of each compound. The overlaps of this information could be reduced by the PCA, and thus we can add any kind of descriptors to the score vector.

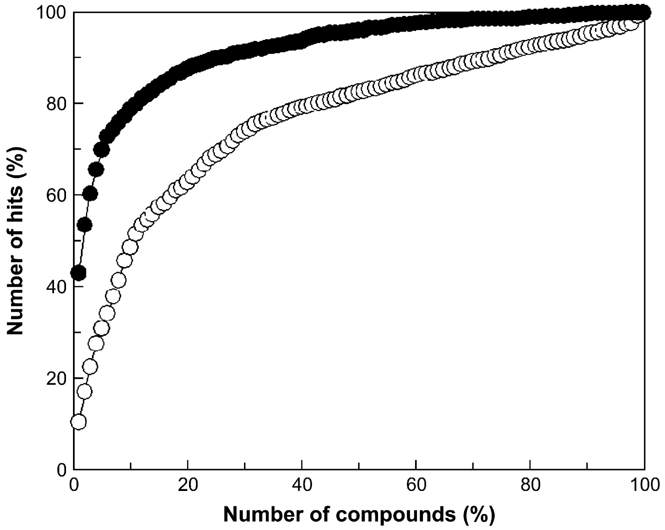

Fig. 7. Averaged database enrichment curves of 12 targets using the affinity matrix of 63 proteins (protein set D). Open circles and filled circles represent the averaged database enrichments by the DSI method and that by the FS-DSI method, respectively.

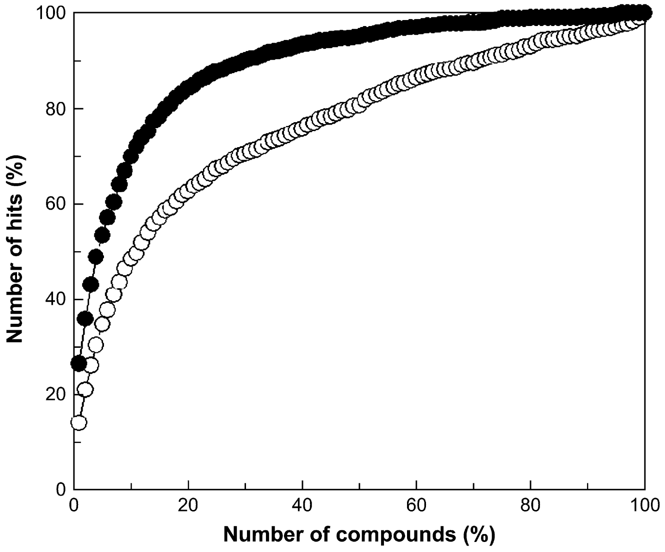

Fig. 8. Averaged database enrichment curves of 12 targets using the affinity matrix of 24 proteins (protein set E). Open circles and filled circles represent the averaged database enrichments by the DSI method and that by the FS-DSI method, respectively.

From Figs. 4–8, although the database enrichment for some targets can be enhanced by using a higher number of proteins, the database enrichment is saturated when using more than 63 proteins (set D) for all the targets investigated in this work, irrespective of whether the DSI or FS-DSI methods are used. Hence, the finding of hits can be achieved by using even a smaller number of proteins rather than using a higher number of proteins.

6. Conclusions

We have developed the FS-DSI method, a modified version of our previous DSI method, to enhance database enrichment based on the protein–ligand docking score matrix using known active compounds. The theoretical framework of the FS-DSI method is applicable to other chemometric methods. DSI and FS-DSI methods were applied to targets such as MIF, HIV, COX-2, thermolysin, GST, the histamine H1 receptor, the adrenaline beta receptor, the serotonin receptor and the dopamine D2 receptor. The new FS-DSI method outperformed the DSI method for all targets. The dependency of the number of proteins on the database enrichment was examined, and it was found that database enrichment was saturated when 63 proteins were used.

Although the FS-DSI method performs well for many of the targets studied here by selecting the major principal components with high database enrichment values, the total database enrichment value and optimal number of selected principal component axes depend on the target used, especially for HIV and GST. Thus, this method is useful for predicting the new active compounds from the database.

Acknowledgements

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

Appendix A

The selected 180 proteins (protein set A) are as follows: 12as, 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, 1cx2, 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, 1htf1, 1htf2, 1hyt, 1ida, 1ivb, 1jap, 1lah, 1lcp, 1ldm, 1lic, 1lna, 1lst, 1mbi, 1mdr, 1gc7, 1mld, 1mmq, 1mmu, 1mrg, 1mts, 1mup, 1nco, 1ngp, 1nis, 1nks, 1okl, 1pbd, 1pdz, 1phd, 1phg, 1poc, 1ppc, 1pph, 1pso, 1pxx, 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, 2tmn, 3cla, 3cpa, 3erd, 3ert, 3pgh, 3r1r, 3tpi, 4cox, 4est, 4lbd, 4phv, 5abp, 5cpp, 5er1, 6cox, 6rnt, 7tim, 1l3f, 3hvp, 5cox, 1aid, 1hpx, 1ivp, 18gs, 2gss, 3pgt and 16gs. For 1abe, 1abf, 5abp, and 1htf, two receptor pockets were prepared, since these proteins both bind two ligands each.

The selected 123 proteins (protein set B) are as follows: 3pgh, 4cox, 5cox, 6cox, 1pxx, 1cx2, 18gs, 3pgt, 2gss, 1aid, 1hpx, 1ivp, 4phv, 1hpv, 1hos, 1qbr, 1tnl, 1tng, 1tlp, 1lna, 1a4q, 1a4g, 1abf, 1glp, 1srj, 1d0l, 1r55, 1c1e, 2cht, 1fki, 2qwk, 2pk4, 1ets, 1aqw, 1hfc, 1f0s, 1hdc, 4est, 1bma, 2pk4, 1ets, 1aqw, 1hfc, 1f0s, 1hdc, 4est, 1bma, 3cla, 1ppc, 2ifb, 1tnh, 1htf, 1byg, 1ckp, 1aoe, 1pph, 1mts, 1a42, 1cdg, 1lic, 1f0r, 1dhf, 1f3d, 1qpq, 1tni, 1lst, 1dg5, 1nco, 1hsb, 1ejn, 1cvu, 1atl, 1cle, 1rne, 1mmq, 1bqq, 1rob, 1ivb, 1coy, 1byb, 3ert, 1dd7, 1bkc, 1ai5, 2tmn, 2fox, 1coy, 1byb, 3ert, 1dd7, 1snc, 1jap, 1hyt, 1epb, 1cbs, 2gbp, 1c5c, 1ngp, 1poc, 1yee, 1cps, 1pbd, 1mbi, 1com, 1xid, 1okl, 3erd, 1a28, 1xie, 1b58, 1d3h, 1fl3, 1hsl, 4lbd, 1fen, 1mdr, 1c83, 1ldm, 3tpi, 1lcp, 2ada, 1dog, 1gc7, 1pdz, 1lah, 3cpa, 4aah, 2ack, 1ebg, 1mrg, 1cbx, 1nis, 1aco, 2ctc and 1mup.

The selected 93 proteins (protein set C) are as follows: 3pgh, 4cox, 5cox, 6cox, 1pxx, 1cx2, 18gs, 3pgt, 2gss, 1aid, 1hpx, 1ivp, 1f3d, 1qpq, 1tnl, 1tni, 1glp, 1lst, 1aoe, 1dg5, 1srj, 1nco, 1a42, 1hsb, 1ets, 1ejn, 1r55, 1cvu, 1hfc, 1atl, 2cht, 1cle, 1hos, 1rne, 1mmq, 1bqq, 1ppc, 1rob, 1ivb, 1coy, 1byb, 3ert, 1dhf, 1dd7, 1bkc, 1ai5, 2tmn, 2fox, 1snc, 1jap, 1hyt, 1epb, 1cbs, 2gbp, 1c5c, 1ngp, 1poc, 1yee, 1cps, 1pbd, 1mbi, 1com, 1xid, 1okl, 3erd, 1a28, 1xie, 1b58, 1d3h, 1fl3, 1hsl, 4lbd, 1fen, 1mdr, 1c83, 1ldm, 3tpi, 1lcp, 2ada, 1dog, 1gc7, 1pdz, 1lah, 3cpa, 4aah, 2ack, 1ebg, 1mrg, 1cbx, 1nis, 1aco, 2ctc and 1mup.

The selected 63 proteins (protein set D) are as follows: 3pgh, 4cox, 5cox, 6cox, 1pxx, 1cx2, 18gs, 3pgt, 2gss, 1aid, 1hpx, 1ivp, 2tmn, 1a28, 1ai5, 1b58, 1bqq, 1c83, 1cbx, 1cdg, 1com, 1coy, 1cvu, 1d3h, 1dog, 1epb, 1fen, 1fki, 1fl3, 1hfc, 1hos, 1jap, 1lcp, 1ldm, 1mbi, 1mdr, 1gc7, 1mld, 1mmq, 1mrg, 1mup, 1ngp, 1okl, 1pbd, 1pdz, 1pso, 1qbu, 1qpq, 1tng, 1xie, 1yee, 2ack, 2ada, 2cmd, 2ctc, 2fox, 2ifb, 2pk4, 3cpa, 3ert, 3tpi, 4aah and 4lbd.

The selected 24 proteins (protein set E) are as follows: 3pgh, 4cox, 5cox, 6cox, 1pxx, 1cx2, 18gs, 3pgt, 2gss, 1aid, 1hpx, 1ivp, 1gc7, 2tmn, 2ada, 1ngp, 1hfc, 1mup, 1fl3, 2ctc, 4aah, 2cmd, 1pbd and 1d3h.

Appendix B

As COX-2 active compounds, 12 inhibitors and two natural ligands were selected. The two natural ligands were arachidonic acid and prostaglandin H2. The 12 inhibitors were diclofenac, etodolac, suprofen, diflunisal, piroxicam, sulindac, indomethacin, ketoprofen, naproxen, nimesulide, rofecoxib, and 1-phenylsulfonamide-3-trifluoromethyl-5-parabromophenylpyrazole.

The names of the thermolysin inhibitors used in the present study are the following, in which the PDB code in parentheses is the complex structure from which the compound originated: l-benzylsuccinate (1hyt), phenylalanine phosphinic acid − deamino-methyl-phenylalanine (1os0), (6-methyl-3,4-dihydro-2H-chromen-2-yl)methylphosphonate (1pe5), 2-(4-methylphenoxy) ethylphosphonate − 3-methylbutan-1-amine (1pe7), 2-ethoxyethylphosphonate − 3-methylbutan-1-amine (1pe8), (2-sulfanyl-3-phenylpropanoyl)-Phe-Tyr (1qf0), [2(*R*,*S*)-2-sulfanylheptanoyl]-Phe-Ala (1qf1), [(2*S*)-2-sulfanyl-3-phenylpropanoyl]-Gly-(5-phenylproline) (1qf2), *n*-(1-(2(*R*,*S*)-carboxy-4-phenylbutyl)cyclopentylcarbonyl)-(*S*)-tryptophan (1thl), (*R*)-retro-thiorphan (1z9g), (*S*)-thiorphan (1zdp), hydroxamic acid (4tln), phenylalanine phosphinic acid (4tmn), honh-benzylmalonyl-L-alanylglycine-P-nitroanilide (5tln), Cbz-Gly^P-Leu-Leu (Zg^{P}Ll) (5tmn), Cbz-Gly^P-(O)-Leu-Leu (Zg^P(O)Ll) (6tmn), CH_2CO(N-OH)Leu-OCH_3 (7tln), benzyloxycarbonyl-D-Ala (1kto), benzyloxycarbonyl-L-Ala (1kl6), benzyloxycarbonyl-D-Thr (1kro), benzyloxycarbonyl-L-Thr (1kj0), benzyloxycarbonyl-D-Asp (1ks7), benzyloxycarbonyl-L-Asp (1kkk), benzyloxycarbonyl-D-Glu (1kr6) and benzyloxycarbonyl-L-Glu (1kjp), aspartame, ASP and PHE.

The names of the GST inhibitors used in the present study are the following, in which the PDB code in parentheses is the complex structure from which the compound originated: benzylcysteine − phenylglycine (10gs), glutathione − [2,3-dichloro-4-(2-methylene-1-oxobutyl)phenoxyacetic acid (11gs), *S*-nonyl-cysteine (12gs), 1-(*S*-glutathionyl)-2,4-dinitrobenzene

(18gs), glutamyl group − *S*-(4-bromobenzyl)cystine (1aqv), glutamyl group − *S*-(2,3,6-trinitrophenyl)cysteine (1aqx), *S*-hexylglutathione (1pgt), cibacron blue (20gs), chlorambucil (21gs), ethacrynic acid (2gss), (9r,10r)-9-(*S*-glutathionyl)-10-hydroxy-9,10-dihydrophenanthrene (2pgt), 2-amino-4-[1-(carboxymethyl-carbamoyl)-2-(9-hydroxy-7,8-dioxo-7,8,9,10-tetrahydro-benzo [def]chrysen-10-ylsulfanyl)-ethylcarbamoyl]-butyric acid (3pgt).

The names or the SMILES of the HIV protease-1 inhibitors used in the present study are the following, in which the PDB code in parentheses is the complex structure from which the compound originated: C1(c2ccc(F)cc2)(SCCS1)CCCN3CCC(c4ccc(Cl)cc4)(O)CC3 (1aid), c1(OCC2N(S(N(C(C(C2O)O)COc3ccccc3)Cc4ccccc4)(=O)=O)Cc5ccccc5)ccccc1 (1ajv), C1(N(C(C(C(C(N1Cc2ccccc2)COc3ccccc3)O)O)COc4ccccc4)Cc5ccccc5)=O (1ajx), [4r-(4alpha,5alpha,6beta,7beta)]-3,3′-[[tetrahydro-5,6-dihydroxy-2-oxo-4,7-bis(phenylmethyl)-1h-1,3-diazepine-1,3(2h)-diyl]bis(methylene)]bis[*N*-2-thiazolylbenzamide (1bv7), C(N(Cc1ncccc1)C)(=O)NC(C(=O)NC(C(C(C(NC(=O)C(C(C)C)NC(N(Cc2ncccc2)C)=O)Cc3ccccc3)(O)O)(F)F)Cc4ccccc4)C(C)C (1dif), C(N1C(C(=O)NC(C)(C)C)CSC1)(=O)C(C(NC(=O)C(NC(=O)COc2[c]3[c](cncc3)ccc2)CSC)Cc4ccccc4)O (1hpx), C(=O)(C(NC(=O)C(CC(C)C)N)CCC(=O)N)NC(C(=O)NC(C(=O)O)CO)CCC(=O)O (1hte), C(=O)(C1C(SC(C(C(=O)NCc2ccccc2)NC(=O)Cc3ccccc3)N1)(C)C)NC(Cc4ccccc4)CO (1htf), c12c(cccc1)NC(=N2)CNC(=O)CC(C(NC(=O)C3C(SC(C(C(=O)NCc4ccccc4)NC(=O)Cc5ccccc5)N3)(C)C)Cc6ccccc6)O (1htg), 2-phosphoglycolic acid (1hvi), C1(N(C(C(C(C(N1Cc2c[c]3[c](cc2)cccc3)Cc4ccccc4)O)O)Cc5ccccc5)Cc6c[c]7[c](cc6)cccc7)=O (1hvr), 2-carbonylquinoline − phenylalaninol group − decahydro-1-methylisoquinoline-2-carbonyl − tertiary-butylamino group (1hxb), ritonavir (1hxw), naphthyloxyacetyl − cyclohexyl Ala-Psi(choh-choh)-Val-2-aminomethyl-pyridine (1ivp), 2-carbonylquinoline − phenylalanylmethane-3-(carboxyamide(2-carboxyamide-2-tertbutylethyl))penta (1jld), C1(N(C(C(C(C(N1Cc2ccc(cc2)CO)Cc3ccccc3)O)O)Cc4ccccc4)Cc5ccc(cc5)CO)=O (1mes), tertiary-butoxyformic acid − phenylalaninol group − dimethylamine − phenylalaninol group − tertiary-butoxyformic acid (1odw), (5r,6r)-2,4-bis-(4-hydroxy-3-methoxybenzyl)-1,5dibenzyl-3-oxo-6-hydroxy-1,2,4-triazacycloheptane (1pro), C1(C(=C(C=C(O1)C(Cc2ccccc2)CC)O)C(c3cc(ccc3)NC(=O)CCNC(=O)OC(C)(C)C)C4CC4)=O (2upj), *N*,*N*-bis-(2(*R*)-hydroxy-1-(*S*)-indanyl-2,6-(*R*,*R*)-diphenylmethyl-4-hydroxy-1,7-heptandiamide (4hpv).

The following compounds are the antagonists of the histamine H1 receptor: astemizole, cetirizine, chlorpheniramine, clemastine, cyprohrptadine, diphenhydramine, homochlorcyclizine, mequitazine, olopatadine, and promethazine.

The following compounds are the agonists of the adrenaline beta receptor: clenbuterol, dobutamine, epinephrine, fenoterol, isoprenaline, mabuterol, methylephedrine, norepinephrine, procatelol, salbutamol, terbutaline, and trimetoquinol.

The following compounds are the antagonists of the adrenaline beta receptor: alprenolol, arotinolol, atenolol, betaxolol, bisoprolol, bopindolol, carteolol, metoprolol, nadlol, pindolol, propranolol, tilisolol, and timolol.

The following compounds are the agonists of the serotonin receptor: 2-Me-5-HT, 8-OH-DPAT, LY334370, alpha-Me-5-HT, m-CPBG, ML10302, RS67506 and sumatriptan.

The following compounds are the antagonists of the serotonin receptor: azasetron, GR113808, granisetron, ketanserin, mesulergine, ondansetron, ramosetron, tropisetron, and WAY-100635.

The following compounds are the agonists of the dopamine D2 receptor: apomorphine, bromocriptine, denopamine, dobutamine, quinpirole, and SFK38393.

The following compounds are the antagonists of the dopamine D2 receptor: benperidol, chlorpromazine, clozapine, fluphenazine, haloperidol, metiapine, molindone, primozide, prochlorperazine, promazine, spiperone, sulpiride, thioproperazine, thioridazine, and trazodone.

References

[1] P.D. Lyne, Drug Discov. Today 7 (2002) 1047−1055.

[2] X. Barril, R.E. Hubbard, S.D. Morley, Mini Rev. Med. Chem. 4 (2004) 779−791.

[3] F.L. Stahura, J. Bajorath, Curr. Pharm. Des 11 (2005) 1189−1202.

[4] J.L. Jenkins, R.Y. Kao, R. Shapiro, Proteins 50 (2003) 81−93.

[5] M. Whittle, V.J. Gillet, P. Willett, A. Alex, J. Loesel, J. Chem. Inf. Comput. Sci. 44 (2004) 1840−1848.

[6] M.M. Ahlstrom, M. Ridderstrom, K. Luthman, I. Zamora, J. Chem. Inf. Model. 45 (2005) 1313−1323.

[7] A. Bocker, S. Derksen, E. Schmidt, A. Teckentrup, G. Schneider, J. Chem. Inf. Model. 45 (2005) 807−815.

[8] A. Evers, G. Hessler, H. Matter, T. Klabunde, J. Med. Chem. 48 (2005) 5448−5465.

[9] G. Hessler, M. Zimmermann, H. Matter, A. Evers, T. Naumann, T. Lengauer, M. Rarey, J. Med. Chem. 48 (2005) 6575−6584.

[10] R.N. Jorissen, M.K. Gilson, J. Chem. Inf. Model. 45 (2005) 549−561.

[11] H. Chen, P.D. Lyne, F. Giordanetto, T. Lovell, J. Li, J. Chem. Inf. Model. 46 (2006) 401−415.

[12] N. Huang, C. Kalyanaraman, J.J. Irwin, M.P. Jacobson, J. Chem. Inf. Model. 46 (2006) 243−253.

[13] I.D. Kuntz, J.M. Blaney, S.J. Oatley, R. Langridge, T.E. Ferrin, J. Mol. Biol. 161 (1982) 269−288.

[14] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, J. Mol. Biol. 261 (1996) 470−489.

[15] G. Jones, P. Willett, R.C. Glen, A.R. Leach, R. Taylor, J. Mol. Biol. 267 (1997) 727−748.

[16] N. Paul, D. Rognan, Proteins 47 (2002) 521−533.

[17] C.A. Baxter, C.W. Murray, D.E. Clark, D.R. Westhead, M.D. Eldridge, Proteins 33 (1998) 367−382.

[18] M.R. McGann, H.R. Almond, A. Nicholls, J.A. Grant, F.K. Brown, Biopolymers 68 (2003) 76−90.

[19] D.S. Goodsell, A.J. Olson, Proteins 8 (1990) 195−202.

[20] J.S. Taylor, R.M. Burnett, Proteins 41 (2000) 173−191.

[21] P.M. Colman, Curr. Opin. Struct. Biol. 4 (1994) 868−874.

[22] A. Krammer, P.D. Kirchhoff, X. Jiang, C.M. Venkatachalam, M. Waldman, J. Mol. Graph. Model. 23 (2005) 395−407.

[23] Y. Fukunishi, Y. Mikami, H. Nakamura, J. Mol. Graph. Model 24 (2005) 34−45.

[24] C. Zhang, S. Liu, Q. Zhu, Y. Zhou, J. Med. Chem. 48 (2005) 2325−2335.

[25] I. Muegge, Y.C. Martin, J. Med. Chem. 42 (1999) 791−804.

[26] G.P. Vigers, J.P. Rizzi, J. Med. Chem. 47 (2004) 80−89.

[27] Y. Fukunishi, Y. Mikami, S. Kubota, H. Nakamura, J. Mol. Graph. Model. (2005).

[28] Y. Fukunishi, Y. Mikami, K. Takedomi, M. Yamanouchi, H. Shima, H. Nakamura, J. Med. Chem. 49 (2006) 523−533.

[29] L.M. Kauvar, D.L. Higgins, H.O. Villar, J.R. Sportsman, A. Engqvist-Goldstein, R. Bukar, K.E. Bauer, H. Dilley, D.M. Rocke, Chem. Biol. 2 (1995) 107−118.

[30] H. Briem, I.D. Kuntz, J. Med. Chem. 39 (1996) 3401−3408.

[31] U.F. Lessel, H. Briem, J. Chem. Inf. Comput. Sci. 40 (2000) 246−253.

[32] N. Hsu, D. Cai, K. Damodaran, R.F. Gomez, J.G. Keck, E. Laborde, R.T. Lum, T.J. Macke, G. Martin, S.R. Schow, R.J. Simon, H.O. Villar, M.M. Wick, P. Beroza, J. Med. Chem. 47 (2004) 4875−4880.

[33] E. Myshkin, B. Wang, J. Chem. Inf. Comput. Sci. 43 (2003) 1004−1010.

[34] A.J. Knox, M.J. Meegan, D.G. Lloyd, Curr. Top. Med. Chem. 6 (2006) 217−243.

[35] O. Roche, G. Trube, J. Zuegge, P. Pflimlin, A. Alanine, G. Schneider, Chembiochem 3 (2002) 455−459.

[36] T.M. Steindl, C.E. Crump, F.G. Hayden, T. Langer, J. Med. Chem. 48 (2005) 6250−6260.

[37] J.W. Nissink, C. Murray, M. Hartshorn, M.L. Verdonk, J.C. Cole, R. Taylor, Proteins 49 (2002) 457−471.

[38] T. Watanabe, Y. Fukui, in: Takayanagi (Ed.), Saiboumaku no jyuyoutai, Nanzandou, Tokyo, 1998, pp. 121−131.

[39] K. Koike, T. Nagatomo, in: Takayanagi (Ed.), Saiboumaku no jyuyoutai, Nanzandou, Tokyo, 1998, pp. 103−118.

[40] M. Sasa, K. Ishihara, in: Takayanagi (Ed.), Saiboumaku no jyuyoutai, Nanzandou, Tokyo, 1998, pp. 135−147.

[41] Y. Nakata, A. Inoue, in: Takayanagi (Ed.), Saiboumaku no jyuyoutai, Nanzandou, Tokyo, 1998, pp. 169−182.

[42] M. Orita, S. Yamamoto, N. Katayama, M. Aoki, K. Takayama, Y. Yamagiwa, N. Seki, H. Suzuki, H. Kurihara, H. Sakashita, M. Takeuchi, S. Fujita, T. Yamada, A. Tanaka, J. Med. Chem. 44 (2001) 540−547.

[43] J. Gasteiger, M. Marsili, Tetrahedron 36 (1980) 3219−3228.

[44] J. Gasteiger, M. Marsili, Tetrahedron Lett. (1978) 3181−3184.

[45] D.A. Case, T.A. Darden, T.E. Cheatham, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, K.M. Merz, B. Wang, A. Pearlman, M. Crwoley, S. Brozell, V. Tsui, H. Gohlke, J. Mongan, V. Hornak, G. Cui, P. Beroza, C. schafmeister, J.W. Caldwell, W.S. Ross, P.A. Kollman, AMBER, 8, University of California, San Francisco, 2004.

[46] D.C. Whitly, M.G. Ford, D.J. Livingstone, J. Chem. Inf. Comput. Sci. 40 (2000) 1160−1168.