Available online at www.sciencedirect.com

ScienceDirect

Journal of Molecular Graphics and Modelling 25 (2007) 633–643

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Finding ligands for G protein-coupled receptors based on the protein–compound affinity matrix

Yoshifumi Fukunishi [a,*], Satoru Kubota [b], Haruki Nakamura [a,c]

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

[b] *Japan Biological Information Research Center (JBIRC), Japan Biological Informatics Consortium (JBIC), 2-41-6, Aomi, Koto-ku, Tokyo 135-0064, Japan*

[c] *Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan*

Received 12 March 2006; received in revised form 27 April 2006; accepted 2 May 2006
Available online 10 May 2006

Abstract

We developed a novel method of identifying new active ligands based on information related to known active compounds using protein–compound docking simulations, even when the tertiary structure of the actual target receptor protein is unknown. This method was used to find ligands of G protein-coupled receptors (GPCRs), i.e., agonists and antagonists of histamine, adrenaline, serotonin and dopamine receptors. The principal component analysis (PCA) method was applied to the protein–compound affinity matrix, which was given by thorough docking calculations between sets of many protein pockets and chemical compounds. The set of protein pockets did not necessary include the target protein. Each compound was depicted as a point in the PCA space. Compounds in a sphere, whose center was set to the known active compound in the multi-dimensional PCA space or to the average position of several known active compounds, were selected as candidate-hit compounds. Our method was found to be effective for finding the ligands of GPCRs based on known native ligands, even when only the soluble protein structures were used in the docking simulations.

Keywords: Receptor–ligand docking; GPCR; Protein–compound interaction matrix; Principal component analysis; Database enrichment

1. Introduction

Protein–ligand docking is a key technology for in silico screening, and many protein–ligand docking programs have been reported [1–12]. This approach is successful when the target protein structure is properly prepared and the size of the ligand is small [13–20]. The hit ratio of the conventional single-target in silico screening is low, and several multiple-target screening methods have been proposed to improve it [12,21,22]. These approaches require the 3D atomic structure of the target protein, and are successful, based on the rapid increase in the number of entries in the Protein Data Bank (PDB). Some of the most important targets are the G protein-coupled receptors (GPCRs), whose 3D atomic structures are unknown with one exception, rhodopsin, which is not interesting in terms of potential medical applications.

When the atomic structures of targeted GPCRs are unknown, one of two approaches is usually applied: either a compound similarity search based on the known active compounds, or homology modeling of the GPCR structure using conventional in silico screening. The compound similarity search is one of the most widely used methods [23,24], however, it has the disadvantage that it finds only chemical derivatives of the known compounds, and it is difficult to find new antagonists when the given known ligands are agonists.

Homology modeling has been applied to GPCRs using the rhodopsin structure as a template. When the known active compounds are available and the model of the complex structure is carefully prepared, in silico screening shows a good hit ratio [25–28]. This approach selects antagonists or agonists only by means of careful preparation of the homology modeled structure.

Pharmacophore modeling is an intermediate approach between the ligand-based similarity search and structure-based drug screening. The pharmacophore model is built based on the

[*] Corresponding author. Tel.: +81 3 3599 8290; fax: +81 3 3599 8099.
E-mail address: y-fukunishi@jbirc.aist.go.jp (Y. Fukunishi).

1093-3263/$ – see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2006.05.001

known active compounds, and it can represent the 3D structure of the ligand-binding pocket. Thus, this method is useful in finding ligands which are not similar to the known active compounds [29].

The affinity fingerprint approach 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 [30]. Later, the protein–compound docking score came to be used as the descriptor of the compound instead of the usual 1D or 2D descriptor, namely, the mass weight, the number of rotatable bonds and the number of hydrogen donors/acceptors of the compound, etc. [31–33]. This approach does not require the 3D structure of the target protein.

Recently, we reported a new method for the classification of chemical compounds, and applied it to a random screening experiment for macrophage migration inhibitory factor (MIF) [34]. In this method, which is related to the affinity fingerprint approach, the principal component analysis (PCA) method was applied to the protein–compound interaction matrix, which was given by means of thorough docking calculations between sets of many protein pockets and chemical compounds. This method was applied to distinguish the active compounds of MIF from its negative compounds. A random screening experiment for MIF was performed, and our method revealed that the active compounds were localized in the PCA space of the compounds, while the negative compounds showed a wide distribution. 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 whose database enrichment was equivalent to or better than that obtained by an in silico screening method [34].

Since our new PCA method does not require the 3D atomic structure of the target protein, it can be applied to the GPCRs. In the present study, we applied our method to the GPCRs, having selected the native ligands of the GPCRs as the known active compounds.

2. Methods

2.1. Simple application of principal component 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, PCA is performed to find similar clusters of compounds. The same method can be applied to both protein pockets and compounds [34].

We prepare a set of pockets $P = \{p_1, p_2, p_3, \ldots, p_{N_r}\}$, where p_i represents the ith pocket and N_r 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 kth compound and N_c is the total number of compounds. For each pocket p_i, all compounds of the set X are docked to pocket p_i with a score of s_i^k between the ith pocket and the kth compound. Here, s_i^k corresponds to the binding free energy.

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

$$M_{ij}^{P} = \frac{1}{N_c}\sum_{k=1}^{N_c}(s_i^k - \overline{s_i})(s_j^k - \overline{s_j}), \quad (1)$$

and

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

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

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

The coefficient $\{c_j^k\}$ represents the jth coordinate of the PCA space of the kth compound. In the present study, we refer to this coefficient $\{c_j^k\}$ as the "docking score index (DSI)".

The candidate-hit compounds were selected using the following method. In the PCA space, the compounds in a sphere whose center was set to the native GPCR ligand were selected as candidate-hit compounds. If several native ligands were available, the average position of these ligands was adopted as the center of the sphere. The priority of the compound was determined based on the Euclidean distance between the compound and the native ligand, with the closer compound having the higher priority. The standard deviations (σ) of the DSI values were calculated for each axis, and any DSI values whose distance from the origin was more than 5σ were removed from the analysis. The DSI values were scaled to set the standard deviation of the distribution of compounds of each axis to 1; thus, the distance between two compounds is measured by the standard Euclidean distance in the PCA space. We refer to this procedure as the "DSI method". Fig. 1 is a schematic representation of this method, showing the PCA results of adrenalin, adrenaline beta-receptor agonists, adrenaline beta-receptor antagonists and 1000 other compounds selected randomly. The details of these data are described in the following sections. The blue square represents the adrenaline that is the native ligand. The compounds in the orange circles were selected as the candidate-hit compounds, and the adrenaline beta-receptor agonists/antagonists were efficiently selected.

2.2. Target-oriented rotation

After the PCA, we attempted one of the factor rotations, which rotate the principal component axes. One of the most popular methods of factor rotation is varimax rotation, which maximizes the variance of the data points in the PCA space and is useful for extracting information from various data [35,36]. In the present study, we adopted a slightly different rotation. Suppose one or more active compounds are given, and the first major N_t components are used in the DSI method. Let x and y be

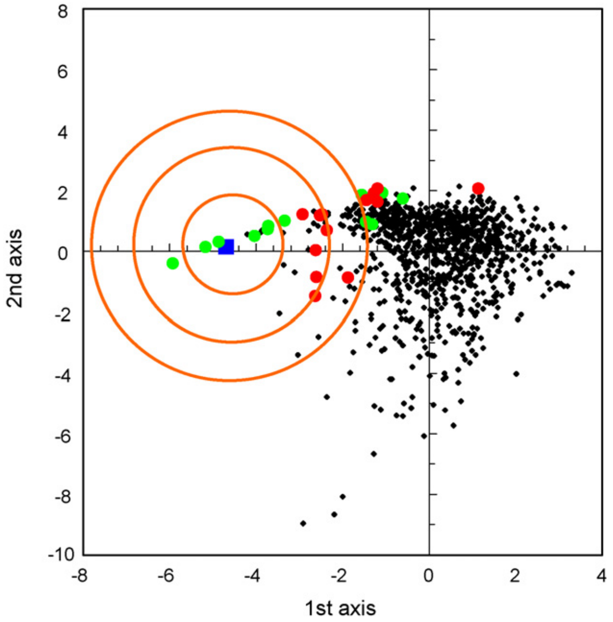

Fig. 1. PCA results of adrenalin, adrenaline beta-receptor agonists, adrenaline beta-receptor antagonists and other compounds. The blue square, green circles and red circles represent adrenalin, adrenaline beta-receptor agonists, and adrenaline beta-receptor antagonists, respectively, the black dots represent the other compounds. The center of the orange open circles is set to the adrenaline. The radii of the orange circles are arbitral.

the ith ($i < N_t + 1$) and jth ($N_t < j$) coefficients of the active compound in the N-dimensional PCA space. The average position of all of the compounds is at the origin (0, 0). On the ith principal component axis, the separation between the active compound and the average position of all of the compounds is equal to x. The following rotation can increase the separation, because $r \geq x$ for any x and y:

$$\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r \\ 0 \end{pmatrix} \tag{4}$$

where $\theta = \arctan(y/x)$ and $r = \sqrt{x^2 + y^2}$.

In the present study, N_t was set to 10, i was set to 1, and the rotation of Eq. (4) was applied to all j; $j > N_t$. We call this procedure the "target orientation rotation (TO)" method.

2.3. Protein–compound docking simulation

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 [12]. The scoring function of this method is based on the rough shape of a protein surface in order 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 geometric hashing 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 root-mean-square deviation (RMSD), attaining a similar performance to that achieved by popular docking programs [12]. In the present study, the number of conformers for flexible docking was limited to 100 for each compound.

3. Preparation of materials

In order to evaluate our screening method, we performed a protein–ligand docking simulation based on the soluble protein structures registered in the PDB and the known GPCR ligands. Here, the protein–ligand complex structures were suitable for the docking study since the ligand pockets were clearly determined. Thus, 142 complexes were selected from the database used in the evaluation of the GOLD and FlexX docking programs [37]. This 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. Since almost all of the atom coordinates are supplied, the atomic structures around the ligand pockets are reliable. Thus, this data set was used in the clustering analysis of proteins and compounds, and in silico screening. From the original data set, the complexes containing a covalent bond between the protein and ligand were removed, since our docking program cannot perform protein–ligand docking when such a covalent bond exists. The complexes are summarized according to their PDB identifiers 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.

The compound set consisted of 10 antagonists of histamine H1 receptor [38], 12 agonists and 13 antagonists of adrenaline beta-receptor [39], 8 agonists and 9 antagonists of serotonin receptor [40], 6 agonists and 15 antagonists of dopamine D2 receptor [41], and 1000 potential-negative compounds extracted from the Coelacanth Chemical Compound Library (Coelacanth Corporation, East Windsor, NJ, USA), which is a random library. Usually only 1 hit compound is found out of 10,000 randomly selected compounds. We would therefore expect that there would be no or very few hit compounds in the 1000 compounds extracted for the present compound set. In total, 73 ligands of GPCRs are listed in Tables 1–7. In addition, the compound

Table 1
Antagonists of histamine H1 receptor

Antagonist	Number of atoms	Distance from the native ligand
Histamine	18	0.0
Diphenhydramine	40	3.8
Promethazine	41	4.2
Chlorpheniramine	39	4.2
Homochlorcyclizine	45	4.4
Olopatadine	47	4.5
Cetirizine	52	4.7
Clemastine	50	4.8
Mequitazine	45	5.3
Cyproheptadine	43	5.3
Astemizole	65	6.8

Table 2
Agonists of adrenaline beta-receptor

Agonist	Number of atoms	Distance from the native ligand
Adrenaline	27	0.0
Clenbuterol	36	1.8
Salbutamol	39	2.0
Trimetoquinol	49	2.1
Procaterol	44	2.1
Mabuterol	39	2.3
Terbutaline	36	2.4
Isoprenaline	33	2.4
Dobutamine	46	2.4
Fenoterol	44	2.5
Methylephedrine	31	2.7
Epinephrine	27	2.9
Norepinephrine	24	3.5

Table 3
Antagonists of adrenaline beta-receptor

Antagonist	Number of atoms	Distance from the native ligand
Nadlol	50	1.5
Bopindolol	57	1.6
Metoprolol	45	1.8
Timolol	47	1.8
Tilisolol	47	1.8
Propranolol	41	1.8
Carteolol	46	1.9
Alprenolol	42	2.0
Pindolol	39	2.0
Atenolol	42	2.2
Bisoprolol	55	2.3
Arotinolol	45	3.0
Betaxolol	52	3.0

library includes the native ligands of these GPCRs, histamine, adrenaline, serotonin and dopamine.

The size distribution of the ligands was as follows: ratio of 0–19 atoms, 0.1%; ratio of 20–29 atoms, 1.2%; ratio of 30–39 atoms, 1.6%; ratio of 40–49 atoms, 9.3%; ratio of 50–59 atoms, 22.5%; ratio of 60–69 atoms, 37.9%; ratio of 70–79 atoms, 20.5%; and ratio of more than 80 atoms, 7.0%. The average ligand size was 64.3 atoms. The molecular size was almost equivalent to the size of the agonists and antagonists listed in Tables 1–7.

The 3D coordinates of the more than 1000 random compounds 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 GPCR ligands were generated by the Chem3D program (Cambridge Software, Cambridge, MA, USA). The atomic charges of each ligand were determined by the Gasteiger method [42,43]. The atomic charges of proteins were the same as the atomic charges of AMBER parm99 [44].

Table 4
Agonists of serotonin receptor

Agonist	Number of atoms	Receptor subtype	Distance from the native ligand
Serotonin	26		0.0
Alpha-Me-5-HT	29	5-HT$_{2A}$, 5-HT$_{2B}$, 5-HT$_{2C}$	0.7
2-Me-5-HT	29	5-HT$_3$	1.0
ML10302	43	5-HT$_4$	2.7
Sumatriptan	43	5-HT$_{1B}$, 5-HT$_{1D}$	2.9
8-OH-DPAT	44	5-HT$_{1A}$	3.5
LY334370	49	5-HT$_{1F}$	3.9
m-CPBG	26	5-HT$_3$	3.9
RS67506	57	5-HT$_4$	4.3

Table 5
Antagonists of serotonin receptor

Antagonist	Number of atoms	Receptor subtype	Distance from the native ligand
Tropisetron	42	5-HT$_3$	3.2
Ramosetron	38	–	3.6
WAY-100635	67	5-HT$_{1A}$	3.6
Ketanserin	53	5-HT$_{2A}$	3.6
Granisetron	48	5-HT$_3$	3.8
Mesulergine	54	5-HT$_{2C}$	3.9
Ondansetron	41	5-HT$_3$	4.4
GR113808	56	5-HT$_4$	4.5
Azasetron	45	–	4.8

Table 6
Agonists of dopamine D2 receptor

Agonist	Number of atoms	Distance from the native ligand
Dopamine	23	0.0
Quinpirole	38	2.2
Apomorphine	38	3.2
Dobutamine	46	3.5
Denopamine	47	3.5
SFK-38393	37	3.6
Bromocriptine	84	5.3

Table 7
Antagonists of dopamine D2 receptor

Antagonist	Number of atoms	Distance from the native ligand
Sulpiride	48	3.2
Fluphenazine	58	3.3
Prochlorperazine	51	3.4
Chlorpromazine	41	3.5
Molindone	45	3.6
Spiperone	56	3.7
Thioproperazine	63	3.7
Promazine	41	3.7
Clozapine	44	3.9
Haloperidol	50	4.0
Metiapine	45	4.1
Benperidol	53	4.5
Thioridazine	52	4.6
Trazodone	50	4.6
Primozide	64	5.6

4. Results

First, a preliminary docking study was performed with the 142 proteins versus the 142 compounds. Here, the 142 compounds were the ligands of the 142 complexes extracted from the PDB as described above. We removed 18 protein pockets for which the Sievgene program was unable to dock more than 0.1% compounds. The names of the removed proteins are listed in Appendix A. Cluster analysis based on the protein–compound score panel was applied to the 124 remaining proteins versus the 142 compounds, following the procedure described in our previous report [12]. Depending on the cluster levels, we prepared four cluster sets of 10, 20, 30 and 50 clusters. In each cluster, we selected one representative protein pocket whose average distance to any other protein pocket in the same cluster was the shortest of all the protein pockets in that cluster. The names of the representative proteins for the clusters are listed in Appendix B.

Fig. 1 shows the PCA results for adrenalin, adrenaline beta-receptor agonists, adrenaline beta-receptor antagonists and the other 1000 compounds. One hundred twenty-four proteins were used in this analysis. The original docking scores of the adrenaline beta-receptor agonists and antagonists are very similar to the values of other compounds, and it is very difficult to find a specific trend in their docking scores. The PCA is so effective in extracting specific features that the adrenalin, adrenaline beta-receptor agonists and adrenaline beta-receptor antagonists are all localized in the PCA space. Additionally, the point which corresponds to adrenaline that is a native ligand is close to the distributions of the adrenaline beta-receptor agonists and antagonists.

Figs. 2–8 show the results of database enrichments made by applying the DSI method to the 10, 20, 30 and 50 representative protein pockets for the clusters and to all 124 pockets for a total of 1077 compounds. Fig. 2a and b shows the database enrichments made by the DSI method for histamine H1 receptor antagonists, whose sphere centers were set to the native ligand histamine and the average position of the known ligands listed in Table 1, respectively. In Fig. 2a, the enrichments with the 10 and 50 representative proteins were good, while those with the 20, 30 and 124 proteins were poor. In contrast, in Fig. 2b, the enrichments were improved as the number of proteins increased.

Figs. 3a and b and 4a and b show the database enrichments made by the DSI method for the ligands of adrenaline beta-receptor. The sphere center was set to the native ligand adrenaline for Figs. 3a and 4a. In Fig. 3b, the sphere center was set to the average position of the known agonists, which are listed in Table 2, and the sphere center in Fig. 4b was set to the average position of the known antagonists, which are listed in Table 3. Fig. 3a and b represents the screening results of the agonists, and Fig. 4a and b represents the screening results of the antagonists. In Fig. 3a and b, the enrichments were quite high and were found to improve as the number of proteins increased. In Fig. 4a, the enrichments with the 20, 30 and 50 proteins were good, but those with the 10 and 124 proteins were poor. In contrast, in Fig. 4b, the enrichment was good and it improved as the number of proteins increased, as shown in Fig. 3a and b.

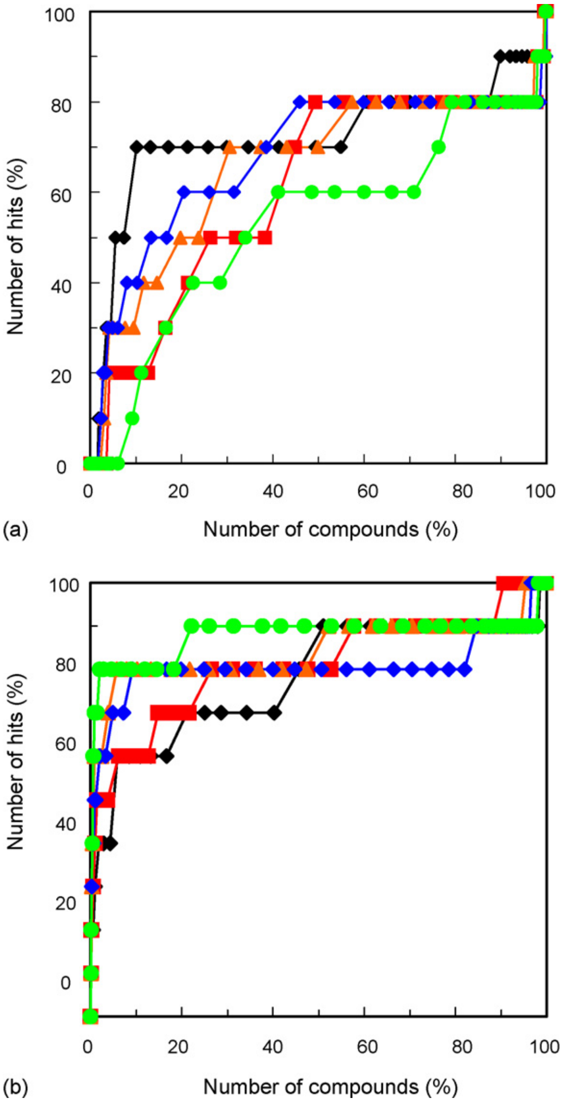

Fig. 2. Database enrichment results of histamine H1 receptor–ligand. The black diamonds, red squares, orange triangles, blue diamonds and green diamonds represent the results obtained for the 10, 20, 30, 50 and 124 proteins, respectively. (a) The sphere center is set to the native agonist histamine; (b) the sphere center is set to the mass center of the known agonists listed in Table 1.

Figs. 5a and b and 6a and b show the database enrichments made by the DSI method for the ligands of the serotonin receptor. Here, the sphere center of the ligands was set to the native ligand serotonin for the data shown in Figs. 5a and 6a. In Fig. 5b, the sphere center was set to the average position of the known agonists, which are listed in Table 4, and the sphere center in Fig. 6b was set to the average position of the known antagonists, which are listed in Table 5. Fig. 5a and b represents the screening results for the agonists, and Fig. 6a and b represents those for the antagonists. In Fig. 5a and b, the enrichments were high, and they improved as the number of proteins increased, except when the number of proteins was 124. The enrichment in Fig. 6a was poor, and did not improve with an increasing number of proteins. In Fig. 6b, however, the

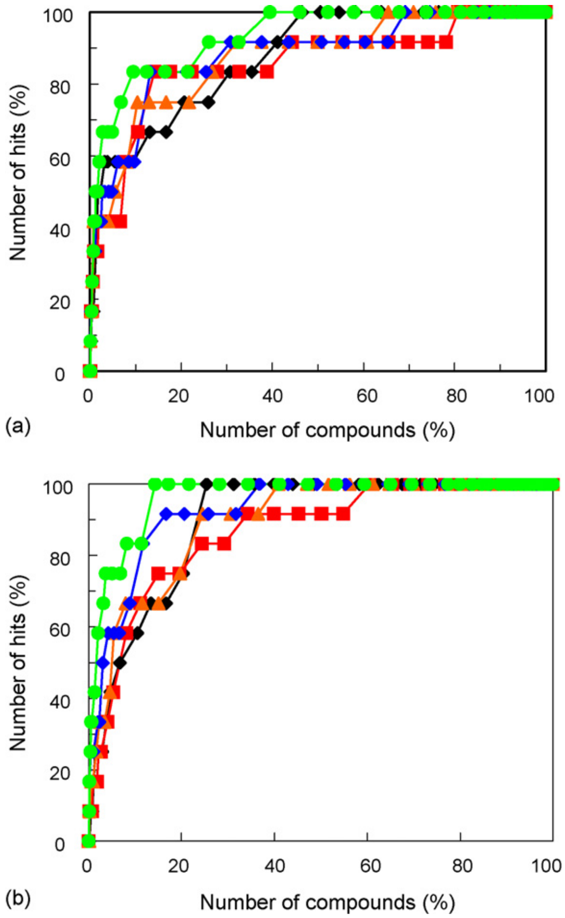

Fig. 3. Database enrichment results of adrenaline beta-receptor agonists. The black diamonds, red squares, orange triangles, blue diamonds and green diamonds represent the results obtained for the 10, 20, 30, 50 and 124 proteins, respectively. (a) The sphere center is set to the native agonist adrenaline; (b) the sphere center is set to the mass center of the known agonists listed in Table 2.

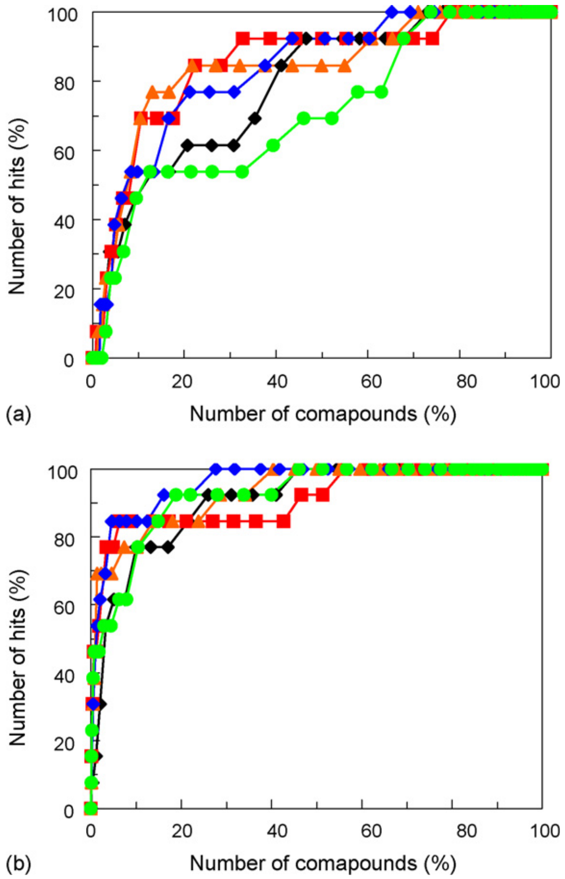

Fig. 4. Database enrichment results of adrenaline beta-receptor antagonists. The black diamonds, red squares, orange triangles, blue diamonds and green diamonds represent the results obtained for the 10, 20, 30, 50 and 124 proteins, respectively. (a) The sphere center is set to the native agonist adrenaline; (b) the sphere center is set to the mass center of the known antagonists listed in Table 3.

enrichment was good and improved with higher numbers of proteins, including the case of 124 proteins.

Figs. 7a and b and 8a and b show the database enrichments made by the DSI method for the ligands of the dopamine D2 receptor. Their sphere center was set to the native ligand dopamine for the data shown in Figs. 7a and 8a. The sphere center was set to the average positions of the known agonists listed in Tables 6 and 7 for Figs. 7b and 8b, respectively. Fig. 7a and b represents the screening results of the agonists, and Fig. 8a and b represents those of the antagonists. The enrichments in all four of these experiments were poor, as in the random screening, and they did not improve when the number of proteins was increased.

Table 8 shows overlapping volume between two distributions of series of compounds. These distributions correspond to the sets of histamine H1 receptor antagonists, adrenaline beta-receptor agonists/antagonists, serotonin receptor agonists/antagonists and dopamine D2 receptor agonists/antagonists listed in Tables 1–7, respectively. In a 10-dimensional PCA space, the average radius of each distribution was calculated assuming that it was a spherical distribution. The volumes and overlapping volumes among these spheres were then calculated in the 10-dimensional space. The distribution of adrenaline beta-receptor antagonists was the most localized, and the volume was set to unity. The relative volumes of the distributions were 178.9, 2.5, 1.0, 6.2, 5.8, 102.9 and 55.0 for histamine H1 receptor antagonists, adrenaline beta-receptor agonists/antagonists, serotonin receptor agonists/antagonists

Table 8
Relative volume and overlapping volume among distributions of GPCR ligands

	Volume	a	b	c	d	e	f	g
a	178.9	100	0.3	0	0.5	0.6	13.7	16.3
b	2.5	24.5	100	15.7	20.5	54.4	99.7	44.4
c	1	8.9	39.7	100	6.9	55.7	87.9	21.9
d	6.2	15.7	8.4	1.1	100	17.6	67.2	37.7
e	5.8	20	23.7	9.6	18.7	100	96	52.5
f	102.9	23.9	2.5	0.9	4	5.4	100	31.6
g	55	53.1	2	0.4	4.2	5.5	59.1	100

a: Histamine H1 receptor antagonists; b: adrenaline beta-receptor agonists; c: adrenaline beta-receptor antagonists; d: serotonin receptor agonists; e: serotonin receptor antagonists; f: dopamine D2 receptor agonists; g: dopamine D2 receptor antagonists. The overlapping volumes in rows are divided by the relative volume and scaled to %.

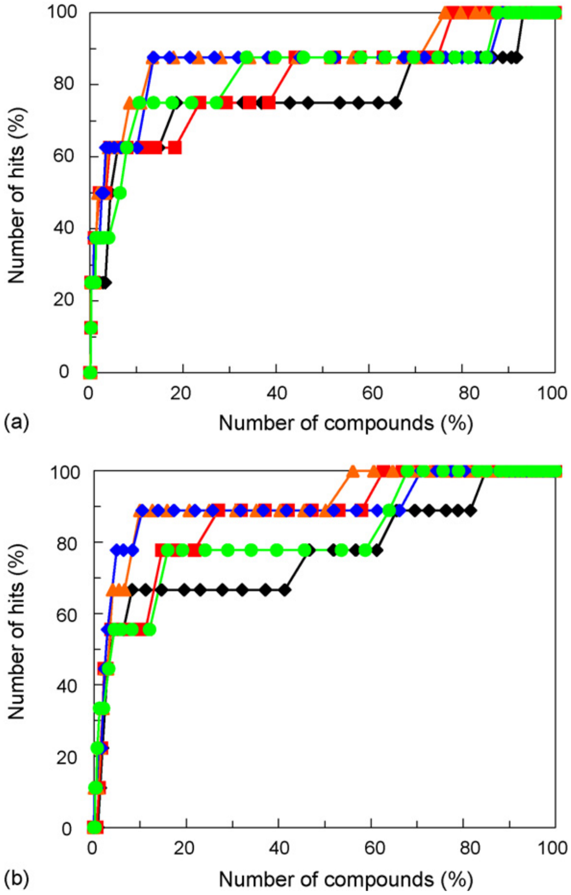

Fig. 5. Database enrichment results of serotonin receptor agonists. The black diamonds, red squares, orange triangles, blue diamonds and green diamonds represent the results obtained for the 10, 20, 30, 50 and 124 proteins, respectively. (a) The sphere center is set to the native agonist serotonin; (b) the sphere center is set to the mass center of the known agonists listed in Table 4.

and dopamine D2 receptor agonists/antagonists, respectively. The larger volumes correspond to a wider distribution and a lower screening efficiency. The overlapping volumes were scaled in terms of %, with the overlapping volume between the *i*th sphere and the *j*th sphere being divided by the volume of the *i*th sphere. The distribution of serotonin antagonists includes 54.4% and 55.7% of the distributions of adrenaline beta-receptor agonists and antagonists, respectively, and the distribution of dopamine D2 agonists includes 99.7%, 87.9%, 67.2% and 96.0% of the distributions of adrenaline beta-receptor agonists and antagonists, and serotonin receptor agonists and antagonists, respectively. Additionally, the distribution of the dopamine D2 antagonists includes most of the other distributions. The large overlap signifies lower selectivity.

Table 9 shows the database enrichments made by the DSI and TO methods. The database enrichment curves were close to each other, and the difference was less than 1%. In some cases, the TO method showed better results than the DSI method, however the difference was less than 0.3% of the 1077 compounds (only three compounds). Thus, there was no statistically significant difference between the two methods.

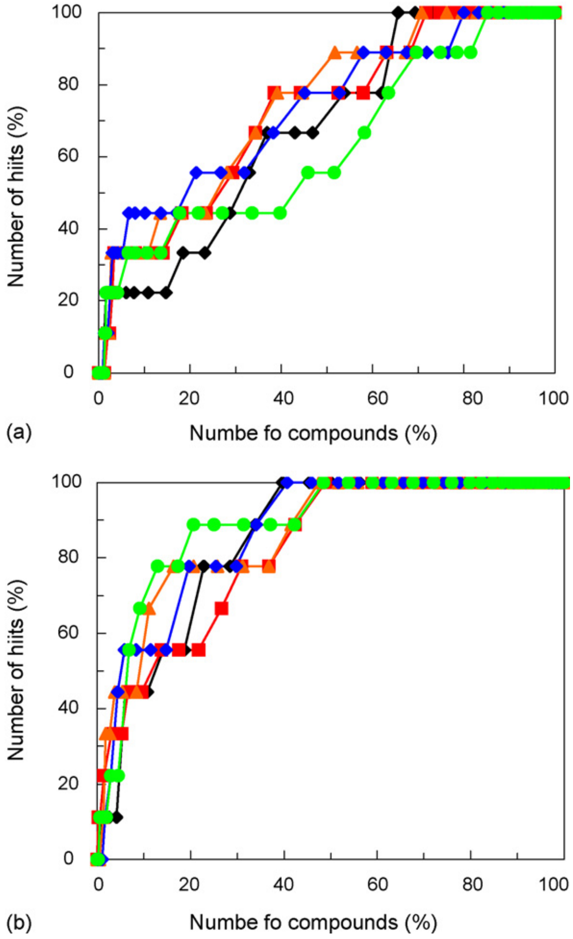

Fig. 6. Database enrichment results of serotonin receptor antagonists. The black diamonds, red squares, orange triangles, blue diamonds and green diamonds represent the results obtained for the 10, 20, 30, 50 and 124 proteins, respectively. (a) The sphere center is set to the native agonist serotonin; (b) the sphere center is set to the mass center of the known antagonists listed in Table 5.

5. Discussion

The DSI method was found to be able to produce high database enrichment in many cases, even if the 3D structure of the target protein was not available. The present method

Table 9
Database enrichment

Compound name	First active compound[a]		50% active compounds[b]	
	DSI	TO	DSI	TO
Histamine H1 antagonists	2.41	2.41	13.24	13.15
Adrenaline beta agonists	0.19	0.19	2.87	2.59
Adrenaline beta antagonists	2.04	2.13	8.98	8.98
Serotonin agonists	0.19	0.19	2.78	2.78
Serotonin antagonists	1.11	1.11	19.17	18.89
Dopamine D2 agonists	5.00	4.72	31.57	31.39
Dopamine D2 antagonists	3.33	3.33	37.87	37.78

[a] The number indicates how many compounds must be selected to identify the first hit compound.

[b] The number indicates how many compounds must be selected to identify 50% of the hit compounds. The number of compounds is scaled to %.

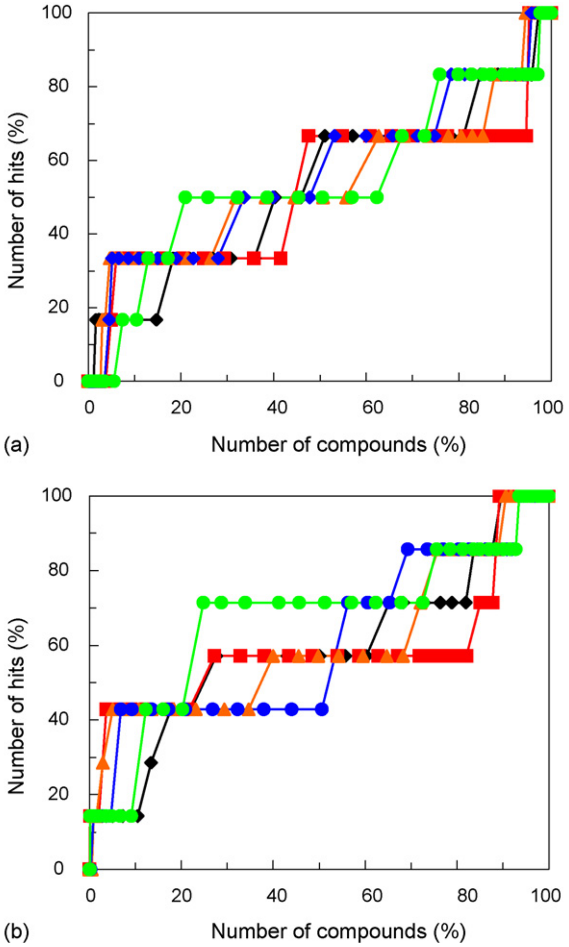

Fig. 7. Database enrichment results of dopamine D2 receptor agonists. The black diamonds, red squares, orange triangles, blue diamonds and green diamonds represent the results obtained for the 10, 20, 30, 50 and 124 proteins, respectively. (a) The sphere center is set to the native agonist dopamine; (b) the sphere center is set to the mass center of the known agonists listed in Table 6.

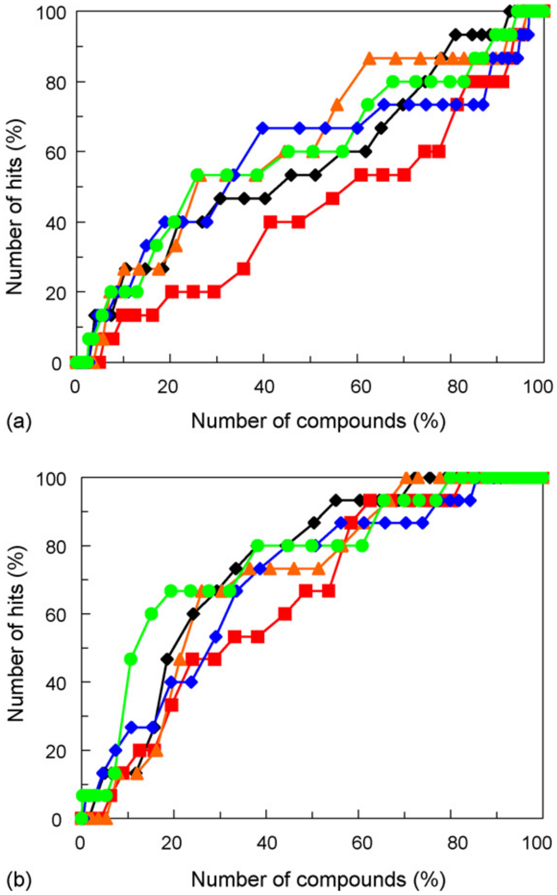

Fig. 8. Database enrichment results of dopamine D2 receptor antagonists. The black diamonds, red squares, orange triangles, blue diamonds and green diamonds represent the results obtained for the 10, 20, 30, 50 and 124 proteins, respectively. (a) The sphere center is set to the native agonist dopamine; (b) the sphere center is set to the mass center of the known antagonists listed in Table 7.

measures the distances among compounds using docking scores. The protein pockets were used as probes to examine the chemical structure, whether partial or whole, of each compound. Thus, when we adopt a sufficient number of protein pockets to distinguish all of the compounds in the compound library, the present method is able to distinguish the active compounds from the negative compounds even without the target protein.

In many cases, the enrichment carried out by using the average positions of the known ligands (shown in Figs. 2b, 3b, 4b, 5b, 6b, 7b and 8b) is better than that performed using the native agonist (shown in Figs. 2a, 3a, 4a, 5a, 6a, 7a and 8a). Clearly, the distribution of the known ligands, apart from the native agonist, is closer to the average position of these ligands than to that of the native agonist in the PCA space when the distribution of the ligands is localized.

The sphere center for the screening was set to the native agonist, which seems to be more similar to the synthesized agonist than to the synthesized antagonist; thus, the enrichment of the agonist was better than that of the antagonist, as shown in Figs. 3–6. These results do not suggest that the chemical structure of the native agonist is similar to that of the synthesized agonist, but they do suggest that the protein–ligand interaction of the native agonist is similar to that of the synthesized agonist. The number of atoms in histamine is 18, while those in the antagonists listed in Table 1 range from 40 to 65. Clearly, the synthesized antagonists are much bigger than the native ligand, and their chemical structures are thus expected to be different, but the enrichment shown in Fig. 2a is not poor. The number of atoms in adrenaline is 27, and those in its synthesized agonists and antagonists, listed in Tables 2 and 3, range from 27 to 49 and from 39 to 57, respectively. Although the synthesized agonists and antagonists are bigger than the native ligand, the enrichments shown in Figs. 3 and 4 are quite good. The number of atoms in serotonin is 26, and those in its synthesized agonists and antagonists, listed in Tables 4 and 5, range from 29 to 57 and from 38 to 67, respectively. The number of atoms in dopamine is 23, and those in its synthesized agonists and antagonists, listed in Tables 6 and 7, range from 37 to 84 and from 41 to 64, respectively. The size of the synthesized agonists and antagonists for dopamine receptor are not very different from those for other receptors, but the enrichments shown in Figs. 7 and 8 are quite poor compared to the other enrichments.

Poor enrichment was observed for the dopamine D2 receptor–ligands. Figs. 7b and 8b show poor enrichment when the sphere centers were set to the average position of the known agonists and antagonists. The obtained results show that the distribution of the dopamine receptor–ligands is not localized in the PCA space; our screening method does not work when the ligands are not localized in the PCA space. This result is consistent with the volume of the distribution shown in Table 8. The volumes of the distributions of the dopamine D2 receptor agonists and antagonists are much larger than the values of the other compounds, except the histamine H1 receptor antagonists. The large volume corresponds to wide distribution and low screening efficiency. In addition, the distributions of the dopamine D2 receptor agonists and antagonists include most of the other distributions, and the large overlap signifies low selectivity. The ligands of serotonin receptor bind the various subtypes, as shown in Tables 4 and 5. Figs. 5b and 6b show good enrichments, suggesting that these compounds are similar in the PCA space. In this case, even if the target subtypes are different from each other, the enrichment performed using the native serotonin is good, as shown in Fig. 5a.

Figs. 3–6 show that approximately 30 proteins are sufficient to attain database enrichment. In some cases, however, enrichment performed with all 124 proteins was not better than that carried out with a subset of the proteins. As shown in our previous report, the protein data set is redundant [34]. For example, 1abe, 1abf and 5abp are exactly the same sugar binding proteins; 1htf, 1hos and 1ida are the same HIV proteases; and 1tlp, 1lna and 1tmn are all thermolysins. Thus, the increase in the number of proteins from 50 to 124 does not necessarily signify an increase by a factor of two in the number of unique proteins.

There is no significant difference between the DSI and TO methods. The enrichment changed due to the number of principal components used as follows: 58.3%, 83.3%, 83.3%, 91.7%, 83.3%, 58.3%, 75.0% and 66.7% of the adrenaline agonists were found by the DSI method within the first 10% of the database with 1, 5, 10, 15, 20, 30, 50 and all of the principal components, respectively, when all 142 proteins were used. The increase in the number of principal components used does not entail an increase in the enrichment. The first principal component, the first 5 principal components, the first 10 principal components, the first 20 principal components, and the first 30 principal components include 38.38%, 65.43%, 70.57%, 76.82% and 80.95%, respectively, of the total information. The minor principal components will contain noise, which is a computational error. Not all of the information from the protein–compound affinity matrix is required in order to achieve a good enrichment. The TO method gathers information about the known active compound from all the principal components, but the minor components provide little meaningful information, and this information would therefore not improve database enrichment.

In Tables 1–7, the Euclidian distances from the native ligand to the active compounds in the 10-dimensional PCA space are shown using all of the 124 proteins. The numbers of atoms of these compounds are also indicated. The correlation coefficients between the distances from the native ligands to the compound and the numbers of atoms of the compounds are 0.68, 0.63 and 0.89 for the adrenaline receptor agonists, serotonin receptor agonists and dopamine receptor agonists, respectively. The correlation coefficient of the histamine receptor antagonists is high (0.82), but the values of the adrenaline receptor antagonists, serotonin receptor antagonists and dopamine receptor antagonists are low at 0.02, 0.02 and 0.39, respectively. These results suggest that similar antagonists do not localize in the PCA space as well as dissimilar antagonists, while the similar agonists can localize in the PCA space. Namely, two antagonists which are close to each other in the PCA space are not always similar in chemical structure.

6. Conclusion

We applied our novel DSI method to identify GPCR ligands using a protein–compound docking simulation for the set of soluble proteins. The ligands used in the present study were agonists and antagonists of histamine H1, adrenaline beta, serotonin and dopamine D2 receptors. Our method was found to be effective in finding the ligands of GPCRs based on known native ligands even if the receptor structures were unavailable. In the DSI method, the number of proteins needed to achieve database enrichment was approximately 30. Even if 124 proteins were used in the analysis, the result was not necessarily improved.

The database enrichments of agonists were better than those of antagonists, when the native agonists were set at the center of the sphere for screening under the DSI method. This is due to the fact that the native agonist is more similar to the synthesized agonist than to the synthesized antagonist, but the difference in molecular size between the native ligand and the synthesized ligand affects database enrichment only slightly.

The DSI method, though efficient, can be applied only when the ligands are localized in the PCA space. Additionally, it is difficult to distinguish agonists and antagonists with this method. Nevertheless, although this method showed poor database enrichments for serotonin antagonists and dopamine D2 agonists and antagonists, it was found to achieve good database enrichments for histamine H1 antagonists, adrenaline beta agonists and antagonists, and serotonin agonists.

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 following 142 complexes were used, listed here by PDB identifier: 1a28, 1a42, 1a4g, 1a4q, 1abe*, 1abf*, 1aco*, 1ai5, 1aoe, 1apt, 1apu, 1aqw, 1atl, 1b58, 1b9v, 1bkc, 1bma, 1bqq, 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, 1gcz, 1glp*, 1hdc, 1hfc, 1hos, 1hpv, 1hsb, 1hsl*, 1htf, 1hyt, 1ida, 1ivb, 1jap, 1kjo*, 1lah*, 1lcp, 1ldm, 1lic, 1lna, 1lst*, 1mbi, 1mdr, 1mld, 1mmq, 1mrg, 1mts, 1mup, 1nco, 1ngp, 1nis, 1okl, 1pbd, 1pdz, 1phd, 1phg, 1poc, 1ppc, 1pph, 1pso, 1qbr, 1qbu, 1qpq, 1r55*, 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 receptor pockets were prepared, since these proteins bind two ligands each. The 18 proteins with "*" were removed in the analysis, since the docking scores with the protein were not calculated frequently.

Appendix B

The 10 selected representative proteins of the first cluster set were 2ada, 1ngp, 1hfc, 1mup, 1fl3, 2ctc, 4aah, 2cmd, 1pbd and 1d3h. The 20 selected representative proteins of the second cluster set were 1cbx, 1d3h, 1epb, 1fl3, 1hfc, 1lcp, 1ldm, 1mrg, 1mup, 1ngp, 1pbd, 1pdz, 1qpq, 2ack, 2ada, 2cmd, 2ctc, 2fox, 3ert and 4aah. The 30 selected representative proteins of the third cluster set were 1a28, 1cbx, 1com, 1d3h, 1epb, 1fen, 1fl3, 1hfc, 1hos, 1lcp, 1ldm, 1mdr, 1mrg, 1mup, 1ngp, 1pbd, 1pdz, 1pso, 1qpq, 1tng, 2ack, 2ada, 2cmd, 2ctc, 2fox, 3cpa, 3ert, 3tpi, 4aah and 4lbd. The 50 selected representative proteins of the fourth cluster set were 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.

Appendix C. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2006.05.001.

References

[1] I.D. Kuntz, J.M. Blaney, S.J. Oatley, R. Langridge, T.E. Ferrin, A geometric approach to macromolecule–ligand interactions, J. Mol. Biol. 161 (1982) 269–288.

[2] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, A fast flexible docking method using an incremental construction algorithm, J. Mol. Biol. 261 (1996) 470–489.

[3] G. Jones, P. Willet, R.C. Glen, A.R. Leach, R. Taylor, Development and validation of a genetic algorithm for flexible docking, J. Mol. Biol. 267 (1997) 727–748.

[4] N. Paul, D. Rognan, ConsDock: a new program for the consensus analysis of protein–ligand interactions, Proteins Struct. Funct. Genet. 47 (2002) 521–533.

[5] C.A. Baxter, C.W. Murray, D.E. Clark, D.R. Westhead, M.D. Eldridge, Flexible docking using tabu search and an empirical estimate of binding affinity, Proteins Struct. Funct. Genet. 33 (1998) 367–382.

[6] M.R. McGann, H.R. Almond, A. Nicholls, J.A. Grant, F.K. Brown, Gaussian docking functions, Biopolymers 68 (2003) 76–90.

[7] D.S. Goodsell, A.J. Olson, Automated docking of substrates to proteins by simulated annealing, Proteins Struct. Funct. Genet. 8 (1990) 195–202.

[8] J.S. Taylor, R.M. Burnett, DARWIN: a program for docking flexible molecules, Proteins Struct. Funct. Genet. 41 (2000) 173–191.

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

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

[11] A. Krammer, P.D. Kirchhoff, X. Jiang, C.M. Venkatachalam, M. Waldman, LigScore: a novel scoring function for predicting binding affinities, J. Mol. Graphics Model. 23 (2005) 395–407.

[12] Y. Fukunishi, Y. Mikami, H. Nakamura, Similarities among receptor pockets and among compounds: analysis and application to in silico ligand screening, J. Mol. Graphics Model. 24 (2005) 34–45.

[13] 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, Coumarin and chromen-4-one analogues as tautomerase inhibitors of macrophage migration inhibitory factor: discovery and X-ray crystallography, J. Med. Chem. 44 (2001) 540–547.

[14] S. Cotesta, F. Giordanetto, J.-Y. Trosset, P. Crivori, R.T. Kroemer, P.F.W. Stouten, A. Vulpetti, Virtual screening to enrich a compound collection with CDK2 inhibitors using docking, scoring, and composite scoring models, Proteins Struct. Funct. Bioinf. 60 (2005) 629–643.

[15] I. Schellhammer, M. Rarey, FlexX-Scan: fast, structure-based virtual screening, Proteins Struct. Funct. Bioinf. 57 (2004) 504–517.

[16] A. Evers, G. Hessler, H. Matter, T. Klabunde, Virtual screening of biogenic amine-binding G-protein-coupled receptors: comparative evaluation of protein- and ligand-based virtual screening protocols, J. Med. Chem. 48 (2005) 5448–5465.

[17] M.H. Howard, T. Cenizal, S. Gutteridge, W.S. Hanna, Y. Tao, M. Totrov, V.A. Wittenbach, Y.-J. Zheng, A novel class of inhibitors of peptide deformylase discovered through high-throughput screening and virtual ligand screening, J. Med. Chem. 47 (2004) 6669–6672.

[18] J.W. Godden, F.L. Stahura, J. Bajorath, POT-DMC: a virtual screening method for the identification of potent hits, J. Med. Chem. 47 (2004) 5608–5611.

[19] L. Zhao, R.D. Brinton, Structure-based virtual screening for plant-based ER^{β}-selective ligands as potential preventative therapy against age-related neurodegenerative diseases, J. Med. Chem. 48 (2005) 3463–3466.

[20] J. Mestres, G.H. Veeneman, Identification of "latent hits" in compound screening collections, J. Med. Chem. 46 (2003) 3441–3444.

[21] G.P.A. Vigers, J.P. Rizzi, Multiple active site corrections for docking and virtual screening, J. Med. Chem. 47 (2004) 80–89.

[22] Y. Fukunishi, Y. Mikami, S. Kubota, H. Nakamura, Multiple target screening method for robust and accurate in silico ligand screening, J. Mol. Graphics Model. 25 (2005) 61–70.

[23] S. Pickett, in: H.J. Boehm, G. Schneider, R. Mannhold, H. Kubinyi, G. Folkers (Eds.), Protein–Ligand Interactions from Molecular Recognition to Drug Design—Methods and Principles in Medicinal Chemistry, WILEY-VCH, Weinheim, 2003, pp. 88–91.

[24] R.S. Pearlman, K.M. Smith, Metric validation and the receptor-relevant subspace concept, J. Chem. Inf. Comput. Sci. 39 (1999) 28–35.

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

[26] C.N. Cavasotto, A.J.W. Orry, R.A. Abagyan, Structure-based identification of binding sites, native ligands and potential inhibitors for G-protein-coupled receptors, Proteins Struct. Funct. Bioinf. 51 (2003) 423–433.

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

[28] C. Bissantz, P. Bernard, M. Hibert, D. Rognan, Protein-based virtual screening of chemical database. II. Are homology models of G-protein-coupled receptors suitable targets? Proteins Struct. Funct. Bioinf. 50 (2003) 5–25.

[29] O. Dror, A. Shulman-Peleg, R. Nussinov, H.J. Wolfson, Predicting molecular interactions in silico. I. A guide to pharmacophore identification and its applications to drug design, Curr. Med. Chem. 11 (2004) 71–90.

[30] 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, Predicting ligand binding to proteins by affinity fingerprinting, Chem. Biol. 2 (1995) 107–118.

[31] H. Briem, I.D. Kuntz, Molecular similarity based on DOCK-generated fingerprints, J. Med. Chem. 39 (1996) 3401–3408.

[32] U.F. Lessel, H. Briem, Flexsim-X: a method for the detection of molecules with similar biological activity, J. Chem. Inf. Comput. Sci. 40 (2000) 246–253.

[33] 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, Novel cyclooxygenase-1 inhibitors discovered using affinity fingerprints, J. Med. Chem. 47 (2004) 4875–4880.

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

[35] R.B. Cattel, The scree test for the number of factors, Multivariate Behav. Res. 1 (1966) 245–276.

[36] H. Abdi, in: M. Lewis-Beck, T. Futing (Eds.), Encyclopedia for Research Methods for the Social Sciences, Sage, Thousand Oaks, 2003, pp. 978–982.

[37] J.W.M. Nissink, C. Murray, M. Hartshorn, M.L. Verdonk, J.C. Cole, R. Taylor, A new test set for validating predictions of protein–ligand interaction, Proteins Struct. Funct. Genet. 49 (2002) 457–471.

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

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

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

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

[42] J. Gasteiger, M. Marsili, Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges, Tetrahedron 36 (1980) 3219–3228.

[43] J. Gasteiger, M. Marsili, A new model for calculating atomic charges in molecules, Tetrahedron Lett. (1978) 3181–3184.

[44] D.A. Case, T.A. Darden, T.E. Cheatham III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, K.M. Merz, B. Wang, D.A. Pearlman, M. Crowley, 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.