J Comput Aided Mol Des (2006) 20:237–248
DOI 10.1007/s10822-006-9047-1

ORIGINAL PAPER

A virtual active compound produced from the negative image of a ligand-binding pocket, and its application to in-silico drug screening

Yoshifumi Fukunishi · Satoru Kubota · Chisato Kanai · Haruki Nakamura

Received: 4 March 2006 / Accepted: 25 April 2006 / Published online: 21 June 2006

Abstract We developed a new structure-based in-silico screening method using a negative image of a ligand-binding pocket and a multi-protein–compound interaction matrix. Based on the structure of the ligand pocket of the target protein, we designed a negative image, which consists of virtual atoms whose radii are close to those of carbon atoms. The virtual atoms fit the pocket ideally and achieve an optimal Coulomb interaction. A protein–compound docking program calculates the protein–compound interaction matrix for many proteins and many compounds including the negative image, which can be treated as a virtual compound. With specific attention to a vector of docking scores for a single compound with many proteins, we selected a compound whose score vector was similar to that of the negative image as a candidate hit compound. This method was applied to representative target proteins and showed high database enrichment with a relatively quick procedure.

Keywords Database enrichment · Docking score · Flexible docking · Negative image · Receptor–ligand docking · Virtual atom · Virtual compound · Virtual screening

Y. Fukunishi (✉) · H. Nakamura
Biological Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6, Aomi, Koto-ku, Tokyo 135-0064, Japan
e-mail: y-fukunishi@jbirc.aist.go.jp

S. Kubota · C. Kanai
Japan Biological Information Research Center (JBIRC), Japan Biological Informatics Consortium (JBIC), 2-41-6, Aomi, Koto-ku, Tokyo 135-0064, Japan

H. Nakamura
Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

Introduction

In-silico (virtual) screening is a useful tool for finding hit compounds from a chemical compound library. This method is based on a protein–compound docking program, many of which have been proposed [1–11]. These methods have been applied to many known protein–ligand complexes, and correctly predict 40–70% of complex conformations within the 2 Å root–mean–square deviation (RMSD) [4, 11]. In contrast, the prediction of protein–compound binding free energy is still poor; the average error of the prediction of free energy is almost 2.5 kcal/mol, which is close to the binding free energy of hit compounds [5, 12–15]. Recent docking programs work very fast; usually the program can predict the protein–compound affinity within one minute. The number of available compounds is increasing and has now reached approximately 10,000,000. Even using a fast docking program, in-silico screening is highly time-consuming.

Several methods have been proposed to improve database enrichment, some of which are based on the protein–compound affinity matrix [11, 15–17]. One of these methods is the multiple active site correction (MASC) score [16]. The MASC score S'_{ik} for the i-th pocket and the k-th compound has been reported by Vigers and Rizzi [16] to be the following:

$$S'_{ik} = (S_{ik} - \mu_k)/\sigma_k, \quad (1)$$

where S_{ik} is the raw docking score, and μ_k and σ_k are the average and standard deviations of the raw docking

score across all pockets for the k-th compound, respectively. When the i-th pocket is the target, we select the k-th compounds, which show a large S'_{ik}. For a target protein, it is difficult to compare the affinities of compounds because of the large error in the docking score. Under this method, the true active compound is expected to show much stronger affinity to its target protein than to other proteins. The results obtained by the MASC scoring method depend strongly on the choice of protein pockets. Additionally, our previous study revealed that the screening results strongly depend on the choice of proteins, and that it is desirable to choose proteins similar to the target protein [11]. The receptor selection (RS) method selects protein pockets which are similar to the target protein pocket, and the MASC scores are calculated [11]. The docking score obtained under the RS method is the sum of the MASC score and the raw docking score. The modification of the score and the selection of proteins under this method have improved database enrichment allowing us to achieve high and robust enrichment [11].

The results obtained by the MASC scoring method and the RS method strongly depend on the computational parameters such as grid size, atom model, scoring system, etc. The multiple target screening (MTS) method, on the other hand, could provide stable screening results [15]. For the k-th compound, the docking scores $\{s_i^k; i = 1,...,M\}$ are sorted in descending order; here the suffix i represents the i-th pocket and the total number of pockets is M. The order n_i^k is assigned to each i-th pocket depending on its value s_i^k. For example, when $n_i^k = 1$, the i-th pocket binds the k-th compound with the strongest affinity. When $n_i^k = M$, the i-th pocket binds with the weakest affinity. This procedure is repeated until the orders $\{n_i^k; i = 1,...,M|k = 1,...,N\}$ are determined for all compounds. Next, we focus on the target k-th pocket. Compounds having the order $n_i^k = 1$ are assigned to be members in compound group-1. Among group-1 members, the compound with the lowest s_i^k is the most likely hit compound. The MTS method could achieve high database enrichment at the same level as the MASC scoring method and the RS method; it is also robust against changes in the docking parameters.

Regardless of the progress of computer hardware, in-silico screening remains time-consuming. Under the MASC, RS and MTS methods, we must prepare the protein–compound affinity matrix. Once the affinity matrix is generated, it can be reused for different target proteins, and thus the required computational time is not proportional to the number of proteins used. Nevertheless, it is necessary to perform docking simulation at least 10^6 times for 10^6 compounds in order to apply the MASC, RS and MTS methods to a new target protein.

Two compounds, which bind the same proteins, are identified as similar compounds. A compound which is similar to a known active compound of the target protein could be a candidate hit compound. Once a protein–compound affinity matrix is generated and the compound library includes at least one active compound, we can select candidate hit compounds by comparing the score vectors of each compound to those of the active compounds [17]. This is knows as the docking score index (DSI) method. Under the DSI method, the computational effort is drastically reduced [17], because once the protein–compound affinity matrix is generated, the additional computations for screening are only the docking calculations among the known active compounds and the proteins, and the simple analysis calculation for the similarity search. Furthermore, this method does not require the 3D structure of the target protein, but only the known active compound of the target protein. The DSI method utilizes principal component analysis (PCA) of the protein–compound affinity matrix. Each compound is plotted as a point in the PCA space, and the compounds which are close to the known active compound are selected as the candidates of hit compounds.

The MASC scoring, RS and MTS methods do not require the known active compound, but rather the 3D structure of the target protein. These methods are as time-consuming as the conventional in-silico screening method. In contrast, the DSI method does not require the 3D structure of the target protein, but rather the known active compound. The DSI method is a new in-silico screening method in which the protein–compound affinity matrix is prepared. In the present study, we developed an in-silico screening method which is as fast as the DSI method, but which does not require the known active compound. Instead, it requires the 3D structure of the target protein, as in the MASC scoring, RS and MTS methods. Under this new method, a negative image of the ligand-binding pocket of the target protein is generated. The negative image corresponds to an ideal virtual ligand of the target protein as well as to an active compound of the target protein. Then, the DSI method is applied using the negative image. Once the protein–compound affinity matrix has been prepared, the only further required computations under this method are docking calculations between the negative image and the proteins of the protein set. Thus, by taking advantage of the concept of the DSI method, we have developed the present structure-based screening procedure, which is not only much faster than the conventional structure-based in-silico screening method, but also shows good database enrichment.

Method

Step 1 A lattice of suitable size is generated around the protein ligand pocket. When the coordinates of the known ligand of the protein–ligand complex structure are available, the lattice is set to cover the coordinates of the ligand. The unit cell of the lattice is cubic and the cell size of the lattice is set at 2.0 Å.

Step 2 The virtual atoms are set in the lattice, with one virtual atom per lattice cell. The initial coordinates of the virtual atom are randomly set in the cell. Figure 1 shows a schematic representation of this step. The open circles represent the positions of the virtual atoms.

Step 3 The force field of the protein is calculated as described in our previous paper [11]. The van der Waals (vdW), accessible surface area and Coulomb interactions are calculated, and the potential due to the hydrogen bond is omitted. A grid potential represents these interactions, assuming that the receptor structure is rigid. The rigid receptor model cannot represent induced fitting, and therefore it is possible to overestimate the repulsion energy due to atomic contact. To reduce this structural noise, the repulsive part of the ligand–receptor potential is modified in order to underestimate the atomic contact.

Using our method, the space is divided into two regions, the inner and outer regions of the protein,

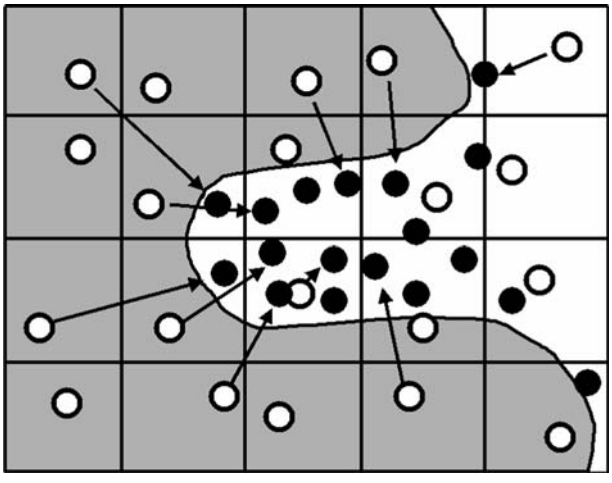

Fig. 1 Schematic representation of negative image generation. The grey region represents the protein. The grid represents the lattice generated in Step 1. The open and filled circles represent the initial and optimized positions of the virtual atoms, respectively

according to the accessible surface of the receptor. In the inner region of the receptor, we apply an artificial smoothing function, and the force field of the outer and the inner regions are smoothly connected at the interface, which is the accessible surface of the receptor.

The potential function for van der Waals interactions in the outer region is

$$E_{\mathrm{vdW}} = w(r)\sum_{a,b} 4\varepsilon_{a,b}\left(\left(\frac{\sigma_{ab}}{R_{ab}}\right)^{12} - \left(\frac{\sigma_{ab}}{R_{ab}}\right)^{6}\right), \tag{2}$$

where R_{ab} is the distance between the a-th atom of the ligand and the b-th atom of the receptor, ϵ and σ are the well-depth of the vdW radius of the atom, and the AMBER parameters are applied to these values [18]. Here, w is a weight function,

$$w(r) = 1 - e^{-C_1 r}, \tag{3}$$

where C_1 is a constant and r is the minimum distance from the a-th atom to the interface which separates the inner and outer regions. In the present study, C_1 was set at 2.0 $\mathrm{Å}^{-1}$. In the inner regions, the potential function for vdW interactions is

$$E_{\mathrm{vdW}} = Dr, \tag{4}$$

where D is a constant and r is the minimum distance from the interface. In the present study, D was set at 5.0 kcal/mol/Å. At the interface, the values of both Eq. 3 and Eq. 4 become zero, such that the vdW potential surface is continuous.

The potential function for Coulomb interactions is given by Eq. 5 for the protein's outer region, and by Eq. 6 for the inner region of the protein:

$$E_{\mathrm{elec}} = \sum_{a,b} \frac{332 q_a q_b}{\varepsilon \cdot r_{ab}^2}, \tag{5}$$

where q_a, q_b, and r_{ab} are the atomic charges of the a-th atom of the ligand and the b-th atom of the receptor, and the distance between the a-th and b-th atoms, respectively. ϵ is a dielectric constant, and we use a uniform value, $\epsilon = 4.0$, in this case:

$$E_{\mathrm{elec}} = w(r_{\mathrm{ASA}})\sum_{a,b} \frac{332 q_a q_b}{\varepsilon \cdot r_{\mathrm{ASA}}^2}, \tag{6}$$

where r_{ASA} is the minimum distance between the a-th atom of the ligand and the receptor-accessible

surface. In addition, w is the following weight function:

$$w(R) = \frac{1}{1+R}. \tag{7}$$

At the interface, the value of Eq. 5 becomes the same as that given by Eq. 6.

The potential function for the hydrophobic interaction is

$$E_{\mathrm{ASA}} = \begin{cases} \sum_{a,b} f(A + C_3 R_{ab}) : R_{ab} \le \sigma_a + \sigma_b + 2r_{\mathrm{prob}} \\ 0 : R_{ab} > \sigma_a + \sigma_b + 2r_{\mathrm{prob}} \end{cases}, \tag{8}$$

where $A = -4\pi\{(\sigma_a + r_{\mathrm{prob}})^2 + (\sigma_b + r_{\mathrm{prob}})^2\}$ and $C_3 = -A/(\sigma_a + \sigma_b + 2r_{\mathrm{prob}})$. Here, σ_a and r_{prob} are the vdW radius of the a-th atom and the probe radius. In the present study, f is an atomic solvation parameter set at 10.0 cal/mol/Å^2 for all atoms; this parameter is close to the previously reported value [19, 20]. Although the pairwise potential in Eq. 8 is only a rough approximation for the hydrophobic interaction, pairwise potentials have frequently been used in docking programs and have provided good estimates [7, 21].

We can numerically smooth the grid potential in order to reduce the structural noise.

$$v_{\alpha,\beta,\gamma} = \frac{n \cdot v_{\alpha,\beta,\gamma} + \sum_{\alpha'=\alpha-1,\alpha+1,\beta'=\beta-1,\beta+1,\gamma'=\gamma-1,\gamma+1} v_{\alpha',\beta',\gamma'}}{n+6}, \tag{9}$$

where $v_{\alpha,\beta,\gamma}$ is a score value at grid point (α, β, γ), and n is set at 2. This smoothing process is then applied three times.

Finally, for actual potential score calculations, we can apply the first-order Lagrange interpolation to the grid potential. The dimensionless raw docking score that is used for receptor–ligand docking is

$$S_{\mathrm{raw}} = g \cdot (E_{\mathrm{vdW}} + E_{\mathrm{elec}} + E_{\mathrm{H-bond}} + E_{\mathrm{ASA}}), \tag{10}$$

where g is a parameter set at 0.01 mol/kcal.

Step 4 The coordinates of the virtual atoms are optimized with an intramolecule interaction and the protein–ligand interaction described in the previous step. As shown in Fig. 1, the open circles, which represent the initial positions of the virtual atoms, move to the filled circles, which represent the optimized positions of the virtual atoms. The intramolecule interaction, which is an interatomic interaction between two virtual atoms, with the interatomic distance R is

$$E_{\mathrm{V}} = 4\varepsilon\left\{\left(\frac{\sigma}{R}\right)^a - \left(\frac{\sigma}{R}\right)^b\right\}, \tag{11}$$

where ϵ, a, b and σ are set at 10.0 kcal/mol, 9, 6 and 1.5 Å, respectively.

Step 5 The interatomic bonding network is generated for the energy-optimized virtual atoms. The bond of the network is not a chemical bond, but a virtual bond, and is generated for the atom pair whose interatomic distance is less than 3 Å. In many cases, the bonding network is not complete; specifically, some atoms or some clusters of atoms are not connected. Bonds of less than 5 Å are allowed to connect the non-connected atoms and non-connected clusters. The major graph (cluster), that is, the one which consists of the maximum number of atoms, is chosen as the negative image and the other minor graphs (clusters and atoms) are removed.

Step 6 The charges of the atoms of the negative image are calculated iteratively. Let q_i be the atomic charge of the i-th atom, and the initial charges of all atoms are set at zero. Suppose the i-th atom is directly connected to the j-th atom by a virtual bond, and that the i-th atom has n bonds. The new atomic charges q_i and q_j are then calculated based on their old values, q_i^{old} and q_j^{old}, respectively.

$$\begin{aligned} q_i &= q_i^{\mathrm{old}} + \Delta q_i \\ q_j &= q_j^{\mathrm{old}} - \Delta q_i/n. \end{aligned} \tag{12}$$

Here, (q_i is selected from –0.001, 0, to +0.001. The protein–ligand Coulomb interaction is calculated as shown by Eqs. 5 and 6. If the total energy decreases with the new charges, the new charges are adopted; otherwise, the other (q_i value is tested.

Step 7 The negative image of the pocket is prepared in Steps 1–6, and we then initiate the docking study. A set of proteins and a chemical compound library are prepared. All compounds are docked to proteins individually. We prepare a set of pockets $P = \{p_1, p_2, p_3, \ldots, p_{N\mathrm{r}}\}$, where $\mathrm{p_i}$ represents the i-th pocket and N_{r} is the total number of pockets, and a set of compounds $X = \{x^1, x^2, \ldots, x^{N\mathrm{c}}\}$, 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 score s_i^k between the i-th pocket and the k-th compound. Here, s_i^k corresponds to the binding free energy. For usual compounds, flexible docking is applied, that is, the

ligand is flexible with the rigid protein. For the negative image (virtual ligand), rigid docking is applied since the bonds of the negative image are virtual and the shape of the negative image must be maintained.

Step 8 Next, the docking score of the negative image is scaled. The density of the atoms of the negative image is higher than that of a usual compound, and higher density of atoms causes stronger affinity (lower score). The RMSD of the scores of usual compounds (R_1) is

$$R_1 = \frac{1}{N_c}\sum_{i}^{Nc}\sqrt{\frac{\sum_{k}^{Nr}\left(s_i^{0k} - <s_i^0>\right)^2}{N_r}} \tag{13}$$

where i, k, s_i^{0k} and $< s_i^0>$ represent the i-th compound, the k-th protein, the docking score between the i-th compound and the k-th protein, and the average value of s_i^{0k} over k.

The RMSD of the scores of the negative image (R_2) is

$$R_2 = \sqrt{\frac{\sum_{k}^{Nr}\left(s_v^{0k} - <s_v^0>\right)^2}{N_r}}. \tag{14}$$

where v represents the negative image.

The new docking score s_v^j is given by scaling the original docking score s_v^{0j} as

$$s_v^j = f_c\frac{R_1}{R_2}s_v^{0j} \tag{15}$$

where f_c is a scaling factor.

Step 9 Finally, the DSI method is applied to the protein–compound interaction matrix [17]. The covariance matrix M^P of the proteins is defined as

$$M_{ij}^P = \frac{1}{N_c}\sum_{k=1}^{N_c}\left(s_i^k - \overline{s_i}\right)\left(s_j^k - \overline{s_j}\right), \tag{16}$$

and (16)

$$\overline{s_i} = \frac{1}{N_c}\sum_{k}^{N_c}s_i^k, \tag{17}$$

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 $X_k = (s_1^k, s_2^k,...,s_{Nr}^k)$ is represented by the linear combination of ϕ_j as

$$X_k = \sum_{j=1}^{Nr}c_j^k\phi_j. \tag{18}$$

The coefficient $\{c_j^k\}$ represents the j-th coordinate of the PCA space (principal component) of the k-th compound.

In the PCA space, the compounds in a sphere whose center was set to the negative image were selected as the candidate hit compounds. The coefficients $\{c_j^k\}$ were scaled to set the standard deviation of the distribution of compounds of each axis to 1.

We call the method described here in Steps 1–9 above the negative-image screening (NIS) method. The parameters in Step 3 are the same parameters of our previous study [11]. The parameters in Steps 1, 4 and 5 are determined to get the suitable size of the negative image, namely, the size of the negative image is almost equivalent to the average size of the compounds in the compound library. The first principal component of the DSI method corresponds to the size of the compound [17]. Thus if the negative image is too large or too small comparing to the average size of the compounds in the library, the large or small compound is likely selected by the NIS method. To avoid such bias, the size of the negative image is set to the average size of the compounds in the library.

Preparation of materials

In order to evaluate our proposed screening method, we performed screening studies for macrophage migration inhibitory factor (MIF), cyclooxygenase 2 (COX–2) and thermolysin. The protein structures we used were selected from known complex structures registered in the Protein Data Bank (PDB). Here, the most of the complexes were selected from the database, which was also used by Nissink et al. in the evaluation of GOLD and FlexX [22]. 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. The lack of atom coordinates is almost zero and the all-atomic structures around the ligand pockets are quite reliable. Thus, this data set was used in the present in-silico screening. We removed from the original data set those complexes containing a covalent bond between the protein and ligand, since our docking software cannot treat covalent bonds. All water

molecules and cofactors were removed from the proteins and all missing hydrogen atoms were added to form atom models of proteins. In addition, some structures of AMP/ADP/ATP-binding proteins, sugar-binding proteins, COX-1 and COX-2 were added to the database. There are total 146 protein structures whose PDB identifiers are summarized in Appendix 1.

The compound set consisted of 12 active compounds of MIF, 10 active compounds of COX-2, 25 active compounds of thermolysin and 1000 inactive compounds extracted from the Coelacanth chemical compound library (Coelacanth Corporation, East Windsor, NJ, USA). The 12 active compounds of MIF are listed in Fig. 2. Compounds **7** and **12** were selected from the PDB, and compounds **2**, **5**, **9**, **10** and **11** were reported in our previous study [17]; the others (compounds **1**, **3**, **4**, **6**, **8**) were prepared in another previous work [23]. The active compounds of COX-2 and thermolysin are listed in Appendices 2 and 3, respectively. The 11047 compounds of the original Coelacanth chemical compound library, which is a random library, were put in alphabetical order and the top 1000 inactive compounds were selected. The correlation coefficient between the order and the number of atoms of the compounds is only 0.12; thus, these 1000 compounds form a random library. The average van der Waals volume of the used compounds is 417.2 $Å^3$.

Table 1 shows the size distribution of the compounds. The average compound size of the library of the 1000 inactive compounds was 64.3 atoms. The average size of the active compounds was slightly smaller than that of inactive compounds. The average compound sizes of the 12 MIF active compounds, 10 COX-2 active compounds and 25 thermolysin active compounds were 33.7, 34.5 and 43.6 atoms, respectively. The distribution overlap between the library and the active compounds of MIF and COX-2 is poor, but the distributions of the library and the active compounds of thermolysin overlapped well.

The affinities of the MIF active compounds are not so strong, namely, the IC_{50} values of compound 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 and 11 are 0.038, 0.4, 0.47, 0.55, 3.4, 4.3, 7, 7.4, 8, 8.1 and 30 μM, respectively [17, 23]. The IC_{50} value of compound 12 was unknown. The affinities of the COX-2 active compounds should be strong, since the most of them are commercially available medicines. The affinities of the thermolysin active compounds distribute widely as shown in Appendix 3, namely, the *Ki* values are 3.8 nM–190 μM.

The 3D coordinates of the chemical compounds were generated by the Concord program (Tripos, St. Louis, MO, USA) from the 2D Sybyl SD files provided by the Coelacanth Corporation. The atomic charges of each ligand were determined by the Gasteiger method [24, 25]. The atomic charges of proteins were the same as the atomic charges of AMBER parm99 [26].

Results and discussion

Application to MIF

The proposed NIS method was applied to the MIF. The PDB code of the structure used was 1gc7. The

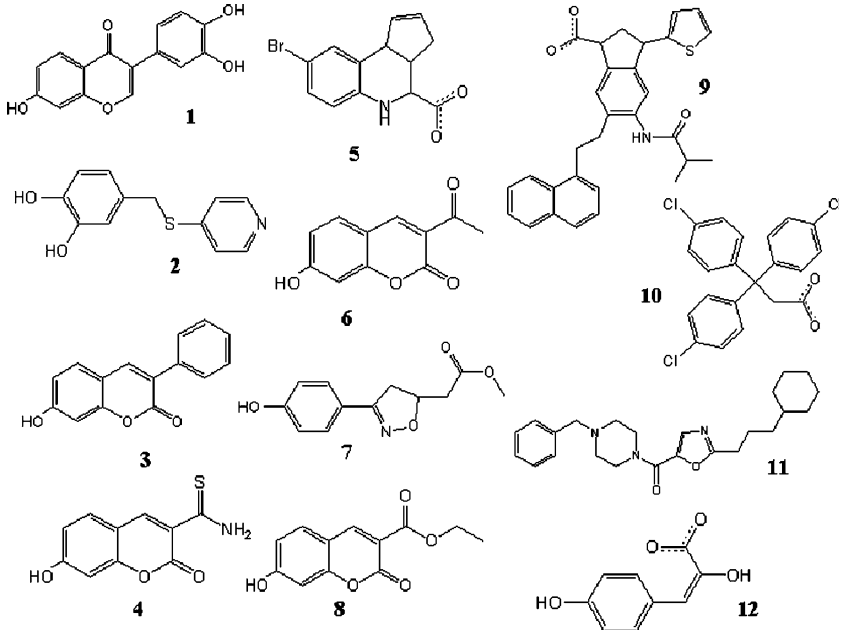

Fig. 2 Migration inhibitory factor (MIF) active compounds

Table 1 Size distribution of the compound library and the active compounds for MIF, COX-2 and thermolysin

Size	Library	MIF	COX-2	Thermolysin
0–19 atoms	0.0	0.0	0.0	0.0
20–29 atoms	0.5	58.3	10.0	16.0
30–39 atoms	0.5	16.7	70.0	36.0
40–49 atoms	6.5	8.3	20.0	4.0
50–59 atoms	22.5	0.0	0.0	12.0
60–69 atoms	40.4	16.7	0.0	28.0
70–79 atoms	22.1	0.0	0.0	4.0
80- atoms	7.4	0.0	0.0	0.0

The size represents the number of atoms of compound. ''Library'' represents the random library of 1000 compounds, and ''MIF'', ''COX-2'' and ''Thermolysin'' represent the active compounds for MIF, COX-2 and thermolysin, respectively. The numbers in this table is %

coordinates of the ligand were used to describe the pocket for designing the negative image.

Figure 3 shows the original ligand of the MIF and the negative image of its ligand-binding pocket. The original ligand was 7-hydroxy-2-*oxo*-chromene-3-carboxylic acid ethyl ester (compound **8**), which is the ligand of complex 1gc7. The number of atoms of the original ligand is 27 (17 heavy atoms and 10 hydrogen atoms) and the number of virtual atoms of the negative image is 63. The vdW volume of the negative image is 506.4 $Å^3$, which is close to the average vdW volume (417.2 $Å^3$) of the compounds in the compound library.

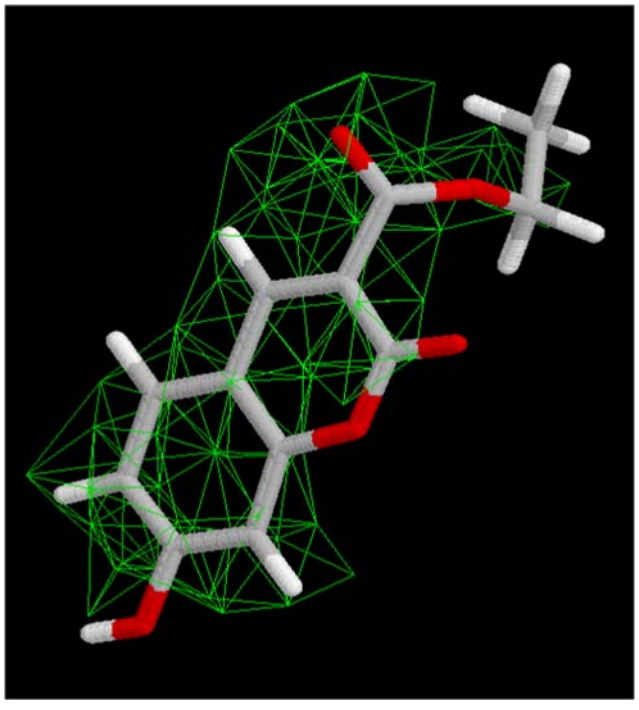

Fig. 3 Ligand of the migration inhibitory factor (MIF) and the negative image of the ligand-binding pocket. The stick model represents the original ligand and the nodes of the green structure represent the optimized coordinates of the virtual atoms

The structure of the negative image looks similar to that of the original ligand, although the negative image is slightly bigger, since the virtual atoms of the negative image achieve closed packing in the pocket.

Figure 4 shows the PCA result of the negative image, and the MIF active and inactive compounds. The 12 active compounds are localized in the PCA space, which is consistent with our previous results [17]. The negative image of the MIF is very close to the distribution of the active compounds. Thus, good database enrichment is expected by using the NIS method.

We calculated the database enrichment by changing the scaling factor *fc* from 0.0 to 1.0. Figure 5 shows database enrichment by the NIS method for the MIF active compound. The atom type of the virtual atom

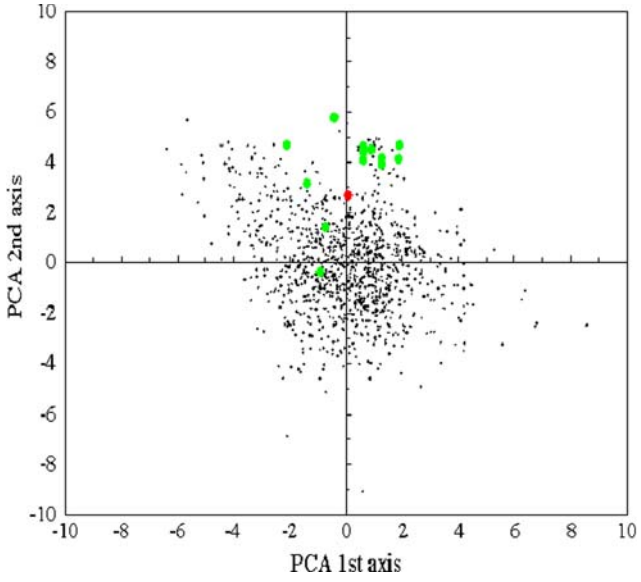

Fig. 4 Principle component analysis (PCA) plot of migration inhibitory factor (MIF) active and inactive compounds and the negative image with the first and second major coordinates. All 146 proteins were used for the analysis. The red circle represents the negative image, and the green circles represent the 12 active compounds. The black dots represent the inactive compounds

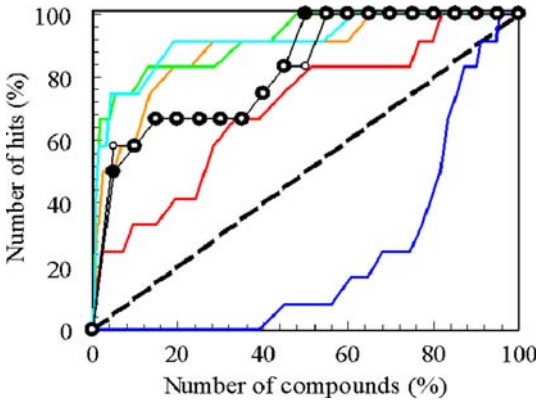

Fig. 5 Database enrichment results of migration inhibitory factor (MIF) by various methods. The dashed line, open circles, and filled circles represent the results obtained with a uniform sampling, the multiple active site correction (MASC) scoring method, and the multiple target screening (MTS) method, respectively. The red, orange, green, light blue and dark blue lines represent the results obtained by the negative-image screening (NIS) method with fc = 1.0, 0.75, 0.5, 0.25 and 0.0, respectively. The numbers of compounds and hits were scaled to %

was set to carbon, whose vdW radius is 1.91 Å. The database enrichment increases when the scaling factor *fc* decreases, and an *fc* value of less than 0.5 gives the optimal result, which is quite good enrichment. This result shows that even if the RMSD of the docking score of the virtual ligand is scaled to the standard value, the absolute value of the docking score of the virtual ligand is still overestimated compared to the optimal value.

The database enrichment obtained by the NIS method was compared with the results obtained by the MTS method. In Fig. 5, the database enrichment by the NIS method is almost equivalent to or better than the results given by the MTS method. The NIS method with $fc = 0.5$ yielded a 7.5–15-fold enrichment, with 75.0 and 75.0% of the ligands found among the first 5 and 10% of the database, respectively; the MASC scoring method yielded a 5.8–11.7-fold enrichment, with 58.3 and 58.3% of the ligands found among the first 5 and 10% of the database, respectively; and the MTS method yielded a 5.8–10-fold enrichment, with 50.0 and 58.3% of the ligands found among the first 5 and 10% of the database, respectively.

Ten negative images were generated with different series of random numbers in Step 2. Using these 12 negative images, we calculated 12 database enrichment results. The scale factor *fc* was set at 0.5 for all cases. The computational conditions were the same as those used in the experiment whose results are shown in Fig. 4. We introduced the following measure to evaluate the database enrichment:

$$S = 100 \int_0^{100} f(x)\mathrm{d}x \Big/ \int_0^{100} f_{\mathrm{ideal}}(x)\mathrm{d}x \qquad (19)$$

where x is the number of compounds in %, $f(x)$ is the database enrichment curve and $f_{\mathrm{ideal}}(x)$ is the ideal database enrichment curve, in which all active compounds are found before any of the inactive compounds. Obviously, $0 < S \leq 100$. Higher database enrichment corresponds to a higher S value, and the value of S obtained by random screening is almost 50. The average S value of these 12 database enrichment curves was 94.99, which is very good, and their RMSD was only 0.27. Thus, the dependence of the negative image on its initial coordinates can safely be ignored.

We also examined the atom-type dependence of the database enrichment. Three atom types were examined: oxygen (vdW radius = 1.66 Å), nitrogen (vdW radius = 1.82 Å) and carbon (vdW radius = 1.91 Å). The negative images consisted of 64, 68 and 63 atoms for the oxygen, nitrogen and carbon atom types, respectively. The results given by these three atom types were very similar to each other, with all three atom types yielding a 15–23-fold enrichment, with 46.2 and 75.0% of the ligands found among the first 2 and 5% of the database, respectively. Nevertheless, there were slight differences. The first active compound was found among the first 0.7, 0.6 and 0.6% of the database by the negative images of oxygen, nitrogen and carbon atoms, respectively, and the last active compound was found among the first 88.1, 89.3 and 90.0% of the database.

Application to COX-2

We applied the proposed NIS method to COX-2 using negative images designed for 1cx2, 1pxx, 4cox, 3pgh and 6cox. The virtual atoms were set to carbons, and the scaling factor *fc* was set at 0.5. For COX-2, the negative image was designed slightly differently from the case of MIF. The negative images, which were designed based on the COX-2 ligands, were so similar to the shapes of the original ligands. If the shape of the negative image was similar to that of the original ligand, the screening result is trivial. Thus, we designed the negative image around the center of the ligand-binding pocket. The center of the ligand-binding pocket was the average position of the ligand atoms in each complex structure. The mesh was generated on the accessible surface of the protein, with a mesh size of 1.5 Å and a probe radius of the accessible surface of 1.4 Å. We then selected the mesh points within a sphere whose center was the center of the pocket and

whose radius was 6.5 Å. These mesh points were used as the coordinates of the known ligand in Step 1. The negative images consisted of 76, 62, 72, 116 and 68 virtual atoms for 1cx2, 1pxx, 3pgh, 4cox and 6cox, respectively. The vdW volumes were 706.5, 501.1, 1091.5, 702.1 and 571.2 Å^3 for 1cx2, 1pxx, 3pgh, 4cox and 6cox, respectively. The negative image realizes closed packing, thus the vdW volume of the negative image becomes larger than the average size of the compounds in the compound library. Figure 6 shows the PCA results of the negative images, and of COX-2 active and inactive compounds. The 10 active compounds are localized in the PCA space, which is consistent with the results obtained in our previous work [17]. The distribution of the negative images is far from the distribution of the active compounds. These images were selected as the four best compounds by the MTS method. The negative images were designed as the optimal virtual ligands, but an optimal virtual ligand is different from a real chemical compound.

Figure 7a, b show the database enrichments given by the negative images and by the MTS method, respectively. The results obtained with the negative images are better than those by the MTS method. The distribution of the negative images was different from that of the active compounds in the 2D space, and the distribution of the active compounds was closer to that of the negative images than to that of the inactive compounds.

The numbers of virtual atoms of these 5 negative images are different from each other, but the 5 database enrichment results obtained by the proposed NIS method are similar to each other. Furthermore, the distribution of these 5 negative images is well localized. On the contrary, the results obtained by the MTS method are different from each other. These results suggest that the NIS method is robust against changes in the shape of the protein pocket due to the induced fitting.

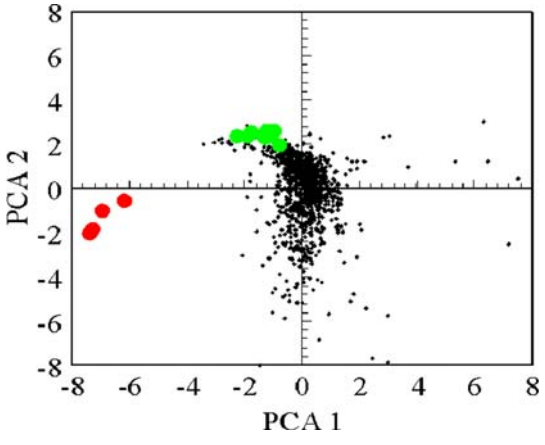

Fig. 6 Principle component analysis (PCA) plot of COX-2 active and inactive compounds and of the negative image with the first and second major coordinates. The red circles represent the negative images, and the green circles represent the 10 active compounds. The black dots represent the inactive compounds

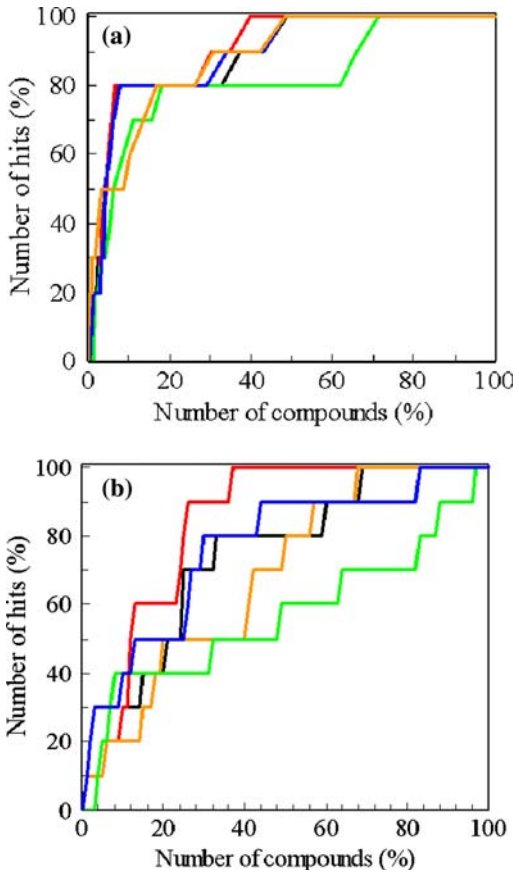

Fig. 7 Database enrichment results of COX-2. The black, red, orange, green and blue lines represent the results for 1cx2, 1pxx, 3pgh, 4cox and 6cox, respectively. The numbers of compounds and hits were scaled to %. (**a**) Database enrichment results by the negative-image screening (NIS) method, (**b**) Database enrichment results by the multiple target screening (MTS) method

To evaluate the robustness of the NIS method against the induced fit, the current NIS method was applied to the apo structure of COX-2 (PDB code: 5cox) that is a structure of COX-2 without ligand. The ligand-binding pocket was determined by comparing to the structure of 6cox. We compared the structure of the apo structure of the COX-2 and that of the protein–ligand complex structure, and the root mean square deviation (RMSD) of the heavy atoms was 1.06 Å for 6cox. The negative images consisted of 64 virtual atoms and the vdW volume was 548.5 Å^3. The NIS method found 40–60% and 60–80% of the ligands among the

first 5 and 10% of the database, respectively for 1cx2, 1pxx, 4cox, 6cox and 3pgh. On the contrary, the NIS method showed slight enrichment, namely, it found 30 and 40% of the ligands among the first 5 and 10% of the database, respectively for 5cox, where the virtual compound was generated from the apo protein structure. Thus, the negative image generated by our automatic method based on the apo protein structure is worse than that based on the protein–ligand complex structure, however, careful modeling of the target protein structure would solve this problem.

Application to thermolysin

The proposed NIS method was applied to thermolysin using negative images designed for 1lna, 1tlp, 1tmn and 2tmn. The shape of each ligand-binding pocket was indicated by the ligand of each complex structure as in the case of MIF, that is, the known ligand coordinates were used in Step 1. The numbers of atoms of the ligand were 41, 69, 67 and 26 for 1lna, 1tlp, 1tmn and 2tmn, respectively, and the numbers of virtual atoms of the negative images were 87, 154, 168 and 66. The vdW volumes were 593.4, 948.9, 998.8 and 470.6 $Å^3$ for 1lna, 1tlp, 1tmn and 2tmn, respectively. The virtual atoms were set to carbons, and the scaling factor *fc* was set at 0.5. Figure 8 shows the PCA results of the negative images, and of the thermolysin active and inactive compounds. The 25 active compounds are localized in the PCA space, which is consistent with our previous results [17]. The distribution of the negative images is close to the distribution of the active compounds. The negative images were selected as the 4 best compounds by the MTS method. The negative images were designed as optimal virtual ligands.

Figure 9a, b show the database enrichments by the negative images and by the MTS method, respectively. The results given by the negative images are better than those obtained by the MTS method. The distribution of the negative images was different from that of the active compounds in the 2D space, and the distribution of the active compounds was closer to that of the negative images than to that of the inactive compounds.

As in the case of COX-2, the numbers of the atoms of the negative images of these 4 proteins are different from each other, but the distribution of these 4 negative images is well localized and the database enrichment results are also similar to each other. In contrast, the 4 database enrichment results given by the MTS method are different from each other. This trend is the same as that found in the case of COX-2. Again these

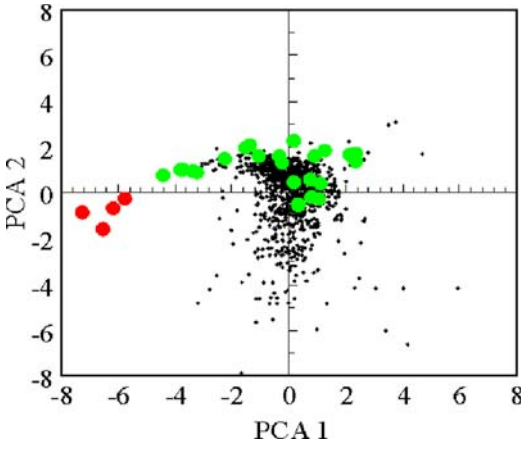

Fig. 8 Principle component analysis (PCA) plot of thermolysin active and inactive compounds and of the negative image with the first and second major coordinates. The red circles represent the negative images, and the green circles represent the 25 active compounds. The black dots represent the inactive compounds

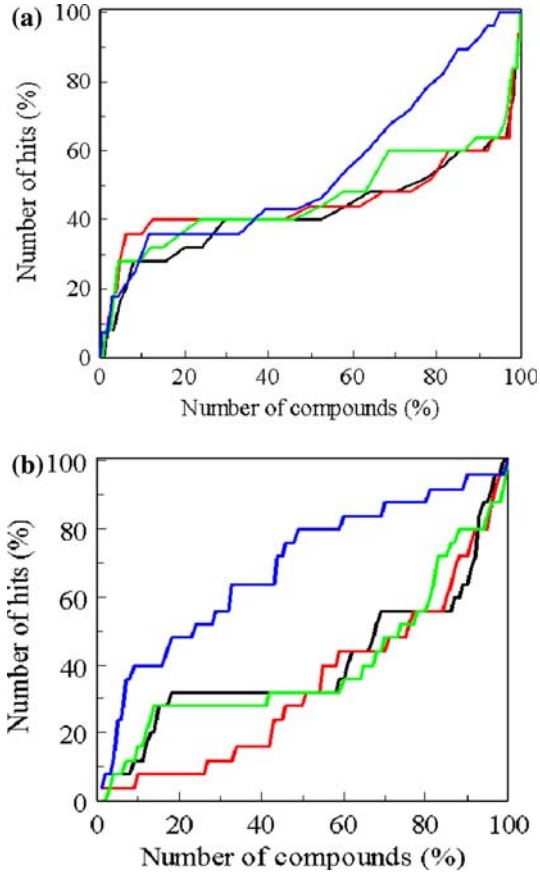

Fig. 9 Database enrichment results of thermolysin. The black, red, green and blue lines represent the results for 1tlp, 1lna, 1tmn and 2tmn, respectively. The numbers of compounds and hits were scaled to %. (**a**) Database enrichment results by the negative-image screening (NIS) method, (**b**) Database enrichment results by the multiple target screening (MTS) method

results suggest that the NIS method is robust against changes in the shape of the protein pocket due to the induced fitting.

To evaluate the robustness of the NIS method against the induced fit, the current NIS method was applied to the apo structure of thermolysin (PDB code: 1l3f). The center of the ligand-binding pocket was set as the zinc ion of the reaction center and the same procedure as the COX-2 case was applied. We compared the structure of the apo structure of the thermolysin and those of the protein–ligand complex structures, and the RMSDs of the heavy atoms were 1.19, 1.69, 1.72 and 1.41 Å for 1lna, 1tlp, 1tmn and 2tmn, respectively. The negative images consisted of 49 virtual atoms and the vdW volume was 456.9 $Å^3$. The NIS method found 16–28% and 28–36% of the ligands among the first 5 and 10% of the database, respectively for 1tlp, 1lna, 1tmn and 2tmn. On the contrary, the NIS method showed slight enrichment, namely, it found 16 and 20% of the ligands among the first 5 and 10% of the database, respectively for 1l3f, where the virtual compound was generated from the apo protein structure. As well as in the COX-2 case, the negative image generated by our automatic method based on the apo protein structure is worse than that based on the protein–ligand complex structure.

Conclusions

We developed a new in-silico screening method, the NIS method. Under this method, the negative image of the ligand-binding pocket of the target protein is generated using virtual atoms, and with the expectation that it could represent an optimal ligand of the target protein. Then, using the negative image as the active compound, we applied the DSI method to select candidate hit compounds. This method was applied to MIF, COX-2 and thermolysin, and was found to achieve high database enrichment in all cases.

The negative image generated by the NIS method is one of the optimal virtual ligands. These images were selected as the best ligands by the MTS method. The negative image is different from the actual active compound. In the PCA space, the distribution of the negative image is different from that of the active compounds. In the present study, the negative image was designed based on the ligand structure of the protein–ligand complex or the accessible surface of the ligand-binding pocket. Instead of using the negative image produced by the virtual atoms, we can use a de novo designed compound, even if it cannot be synthesized.

The screening results obtained by the NIS method are close to or better than those given by the MTS method, which is a structure-based in-silico screening method. Additionally, the NIS method is much faster than the conventional in-silico screening method. The number of dockings required for a new target is equal to the number of previously selected proteins when the docking calculations have already been performed for all compounds of the library against the selected proteins, which are different from the target protein. In contrast, under the conventional in-silico screening, the number of dockings required for the new target protein is equal to the number of compounds of the library. The proposed NIS method thus gives good results and is more efficient than conventional methods.

The NIS method did not require the precise coordinates of ligand of the protein–ligand complex and this method can show the good database enrichment for the protein structure extracted from the protein–ligand complex structure. However, for the protein structure without ligand, the database enrichment by the NIS method is worse than that calculated based on the protein–ligand complex structure.

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 1

The following PDB identifier list of complexes was used: 12asy, 1a28, 1a42, 1a4g, 1a4q, 1ady, 1aer, 1ai5, 1aoe, 1apt, 1apu, 1aqw, 1aszy, 1atl, 1aux, 1b58, 1b76, 1b9v, 1bdg, 1bma, 1byb, 1byg, 1c1e, 1c5c, 1c83, 1cbs, 1cbx, 1cdg, 1ckp, 1com, 1coy, 1cps, 1cqez, 1csny, 1cvu, 1cx2z, 1d0l, 1d3h, 1dd7, 1dg5, 1dhf, 1dog, 1dr1, 1eed, 1efv, 1ejn, 1epb, 1epo, 1eqg, 1eqh, 1ets, 1f0r, 1f0s, 1fen, 1fkg, 1fki, 1fl3, 1gol, 1gtr, 1hck, 1hdc, 1hfc, 1hos, 1hpv, 1hsb, 1htf, 1hyt, 1ida, 1ivb, 1jap, 1lcp, 1lic, 1lna, 1mbi, 1mdr, 1gcz, 1mld, 1mmq, 1mmu, 1mrg, 1mts, 1nco, 1ngp, 1nks, 1okl, 1phd, 1phg, 1poc, 1ppc, 1pph, 1pso, 1pxx, 1pyg, 1qbr, 1qbu, 1qh7, 1qpq, 1rds, 1rne, 1rnt, 1rob, 1s2a, 1s2c, 1ses, 1snc, 1so0, 1tlp, 1tmn, 1tng, 1tnh, 1tni, 1tnl, 1tyl, 1xid, 1xie, 1yee, 2aac, 2aad, 2ack, 2ada, 2cmd, 2cpp, 2fox, 2ifb, 2pk4, 2qwk, 2tmd, 2tmn, 3cla, 3cpa, 3erd, 3ert, 3pgh, 3r1r, 3tpi, 4cox, 4est, 4lbd, 4phv, 5cpp, 5er1, 6cox, 6rnt, and 7tim. For 1htf and 1s2c, two receptor pockets were prepared, since these proteins both bind two ligands each.

Appendix 2

The names of the COX-2 inhibitors used in the present study are suprofen, flubiprofen, indomethacin, ketoprofen, naproxen, etodolac, nimesulide, rofecoxib, diclofenac and Sc-558 (1-phenylsulfonamide-3-trifluoromethyl-5-parabromophenylpyrazole).

Appendix 3

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, also the *Ki* values are supplied when the value is available [27]: L-benzylsuccinate (1hyt: *Ki* = 3.8 nM), 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:*Ki* = 42 nM), [2(R,S)-2-sulfanylheptanoyl]-Phe-Ala (1qf1:*Ki* = 48 nM), [(2S)-2-sulfanyl-3-phenylpropanoyl]-Gly-(5-phenylproline) (1qf2:*Ki* = 1200 nM), *n*-(1-(2(R, S)-carboxy-4-phenylbutyl) cyclopentylcarbonyl)-(S)-tryptophan (1thl), (R)-retrothiorphan (1z9g), (S)-thiorphan (1zdp), hydroxamic acid (4tln:*Ki* = 190 μM), phenylalanine phosphinic acid (4tmn:*Ki* = 68 pM), Honh-benzylmalonyl-L-alanylglycine-P-nitroanilide (5tln), Cbz-Gly^{P}-Leu-Leu ($Zg^{P}Ll$) (5tmn:*Ki* = 9.1 nM), Cbz-Gly^{P}-(*O*)-Leu-Leu ($Zg^{P}(O)Ll$) (6tmn:*Ki* = 9 μM), 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).

References

1. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) J Mol Biol 161:269
2. Rarey M, Kramer B, Lengauer T, Klebe G (1996) J Mol Biol 261:470
3. Jones G, Willet P, Glen RC, Leach AR, Taylor R (1997) J Mol Biol 267:727
4. Paul N, Rognan D (2002) Proteins: Structure, Function, and Genetics 47:521
5. Baxter CA, Murray CW, Clark DE, Westhead DR, Eldridge MD (1998) Proteins: Structure, Function, and Genetics 33:367
6. McGann MR, Almond HR, Nicholls A, Grant JA, Brown FK (2003) Biopolymers 68:76
7. Goodsell DS, Olson AJ (1990) Proteins: Structure, Function and Genetics 8:195
8. Taylor JS, Burnett RM (2000) Proteins: Structure, Function, and Genetics 41:173
9. Abagyan R, Totrov M, Kuznetsov D (1994) J Compt Chem 15:488
10. Colman PM (1994) Curr Opin Struct Biol 4:868
11. Fukunishi Y, Mikami Y, Nakamura H (2005) J Mol Graph Model 24:34
12. Kramer A, Kirchhoff PD, Jiang X, Venkatachalam CM, Waldman M (2005) J Mol Graph Model 23:395
13. Zhang C, Liu S, Zhu Q, Zhou Y (2005) J Med Chem 48:2325
14. Muegge I, Martin YC (1999) J Med Chem 42:791
15. Fukunishi Y, Mikami Y, Kubota S, Nakamura H (2005) J Mol Graph Model 25:61
16. Vigers GPA, Rizzi JP (2004) J Med Chem 47:80
17. Fukunishi Y, Mikami Y, Takedomi K, Yamanouchi M, Shima H, Nakamura H (2006) J Med Chem 49:523
18. Case DA, Darden TA, Cheatham TE III, Simmerling CL, Wang J, Duke RE, Luo R, Merz KM, Wang B, Pearlman DA, Crowley M, Brozell S, Tsui V, Gohlke H, Mongan J, Hornak V, Cui G, Beroza P, Schafmeister C, Caldwell JW, Ross WS, Kollman PA (2004) AMBER 8. University of California, San Francisco
19. Hawkins DG, Cramer JC, Truhlar GD (1996) J Phys Chem 100:19,824
20. Ooi T, Oobatake M, Nemethy G, Scheraga HA (1987) Proc Natl Acad Sci USA 84:3086
21. Stouten PFW, Frommel C, Nakamura H, Sander C (1993) Mol Simul 10:97
22. Nissink JWM, Murray C, Hartshorn M, Verdonk ML, Cole JC, Taylor R (2002) Proteins: Structure, Function, and Genetics 49:457
23. Orita M, Yamamoto S, Katayama N, Aoki M, Takayama K, Yamagiwa Y, Seki N, Suzuki H, Kurihara H, Sakashita H, Takeuchi M, Fujita S, Yamada T, Tanaka A (2001) J Med Chem 44:540
24. Gasteiger J, Marsili M (1980) Tetrahedron 36:3219
25. Gasteiger J, Marsili M (1978) Tetrahedron Lett 3181
26. Wang J, Cieplak P, Kollman PA (2000) J Comput Chem 21:1049
27. Block P, Sotriffer CA, Dramburg I, Klebe G (2006) Nucleic Acids Res 34:D522