Journal of Computer-Aided Molecular Design, **15:** 447–463, 2001.
KLUWER/ESCOM

Design of new selective inhibitors of cyclooxygenase-2 by dynamic assembly of molecular building blocks

Jiang Zhu, Haibo Yu, Hao Fan, Haiyan Liu & Yunyu Shi*
Laboratory of Structural Biology, School of Life Science, University of Science and Technology of China (USTC), Hefei, Anhui, 230026, P.R. China

Received 11 September 2000; accepted 29 January 2001

Key words: clustering algorithm, cyclooxygenase, *De novo* structure-based drug design, molecular similarity, multiple copy, stochastic dynamics

Summary

A method of dynamically assembling molecular building blocks – DycoBlock – has been proposed and tested by Liu et al. [1]. This method is based on multiple-copy stochastic dynamics simulation in the presence of a receptor molecule. In this method, a novel algorithm was used to dynamically assemble the molecular building blocks to form candidate compounds. Currently, some new improvements have been incorporated into DycoBlock to make it more efficient. In the new version of DycoBlock, the binding energy and solvent accessible surface area (SASA) can be used to screen the resulting compounds. A simple clustering algorithm based on molecular similarity was developed and used to classify the remaining compounds. The revised DycoBlock was tested by breaking SC-558 – a selective inhibitor of cyclooxygenase-2 (COX-2) – into building blocks and reassembling them in the active site of the enzyme. The accuracy of recovery grew to 58.8% while it was only 16.7% in the previous version. Then, thirty-three kinds of molecular building blocks were used in the design of novel inhibitors and the investigation of diversity. As a result, a total of 1441 compounds was generated with high diversity. After the first screening procedure, there remained 864 reasonable compounds. The results from clustering indicate that the structural motifs in the diarylheterocycle class of COX-2-selective inhibitors [2] have been generated using the revised DycoBlock, and their binding modes were investigated.

Introduction

The goal of drug design is to discover new lead compounds, which have a high affinity with the receptor so as to alter the biological activity of the receptor, often an enzyme. Among all the methods proposed, *de novo* structure-based drug design [3] and combinatorial chemistry are two major approaches for the discovery of new drugs.

In recent years, several cases of successful application of *de novo* structure-based drug design have been reported. The use of GrowMol [4] and LUDI [5] has led to the synthesis of novel ligands which have been proved to inhibit their target enzymes (GrowMol was used to design aspartyl protease inhibitors which were proved by experiments, and LUDI has been reported to generate a novel thrombin inhibitor). Especially after the idea of multiple-copy simulation [6] was introduced into ligand design (to identify binding sites of functional groups within the active sites of a receptor molecule), a series of new methods such as HOOK [7] and CONCERTS [8] were proposed to produce candidate compounds based on the located binding sites.

A new method – DycoBlock – was proposed for *de novo* structure-based drug design [1]. The essential idea of combinatorial chemistry was incorporated into structure-based ligand design by using the multiple-copy method. Multiple-copy stochastic molecular dynamics (MCSMD) and a novel algorithm were developed to dynamically assemble molecular

*To whom correspondence should be addressed:
E-mail: yyshi@ustc.edu.cn

building blocks at binding sites. It naturally matches the requirement that generated compounds are diverse. DycoBlock has proved quite effective at reproducing ligands by breaking two known inhibitors of HIV-1 protease (L700417 and pepstatin) into fragments and reconstructing them [1].

In spite of the success achieved by DycoBlock, the problem of how to screen the huge number of candidate compounds still remains to be solved, just as in other methods of combinatorial chemistry. In the current work we mainly focused on developing methods to evaluate, rank and classify the huge number of resulting compounds. Computed binding energy between the receptor and compounds was used for this purpose, as well as solvent accessible surface areas (SASAs) of compounds. A simple clustering algorithm was used to classify the constructed lead compounds, whose reasonableness was confirmed by further analysis of results. At the same time, the efficiency of DycoBlock was dramatically enhanced by a series of improvements to the design strategy.

Cyclooxygenase-2 (COX-2) was chosen as the target receptor for testing DycoBlock. Cyclooxygenase (COX), involved in the obligatory step in prostaglandin (PG) and thromboxane biosynthesis (also known as PGH synthase), catalyses the first committed step in arachidonic-acid metabolism [9,10], which is the precursor of prostaglandin (PG). Two isoforms of COX are known as COX-1 and COX-2. The sequence identity between them is ∼60%. COX-1, constitutively expressed in most tissues, is responsible for the physiological production of PGs [11]. COX-2, induced by cytokines, mitogens and endotoxins in inflammatory cells, is responsible for the elevated production of PGs during inflammation [12, 13]. The side effects of many non-steroidal anti-inflammatory drugs (NSAIDs), such as antithrombotic and ulcerogenic effects, arise from their inhibition of COX-1 besides COX-2. Design of selective inhibitors of COX-2 is crucial for discovering anti-inflammatory drugs [2]. SC-558, as shown in Figure 1a, is a diarylheterocyclic inhibitor with a 1,900-fold selectivity for COX-2 over COX-1, having a central pyrazole ring and a sulphonamide substituent attached to one of the aryl rings. The sulphonamide of SC-558 extends into a pocket in the binding site of COX-2, and interacts with His^{90}, Gln^{192} and Arg^{513} of COX-2. A similar pocket exists in COX-1, but is inaccessible because of the substitution of Val^{523} to bulkier Ile^{523} (shown in Figure 2). The pocket is unoccupied in the complex of COX-2 and nonselective inhibitors [14].

In the current study, two tests were carried out for the new version of DycoBlock. The first test was to rebind the known ligand in the active site of the target protein. The selective inhibitor SC-558 was divided into five building blocks by cutting all rotatable single bonds between chemical groups. These building blocks were randomly put into a predefined simulation box containing the active site of COX-2. A series of simulations was performed to determine the best combination of parameters, with which SC-558 could be exactly recovered. The second test is to examine the diversity generated by DycoBlock and to discover new selective inhibitors. In this case, thirty-three building blocks were used with 200 copies for each type. The constructed compounds made up of four, five and six building blocks were recorded and ordered with respect to the binding energy, with the RMS deviation and SASA presented. At this stage, compounds with high binding energy and large SASA in the complex structure would be rejected. A further clustering analysis was used to classify the rest of compounds into several species according to their structural similarity. We found that the structural motifs that appeared in the diarylheterocycle class of COX-2-selective inhibitors (mentioned in the review by Marnett [2]) have been generated correctly by the revised DycoBlock. They were represented by several clusters containing no fewer than five members.

Materials and methods

Multiple-copy stochastic molecular dynamics (MCSMD) simulation was performed during the whole design process [1]. In step one, the MCSMD simulation is carried out to identify the binding sites of functional groups within the active sites of a receptor molecule. In step two, given the distribution and conformations of all the building blocks, candidate compounds are built by exploring different combinations of building blocks through a specially designed algorithm. In the last step, the resulting compounds are optimized during an extended MCSMD simulation combined with simulated annealing (SA) [15].

In this section, some improvements to the DycoBlock method will be described and discussed in detail. Finally, the general strategy in our study will be outlined.

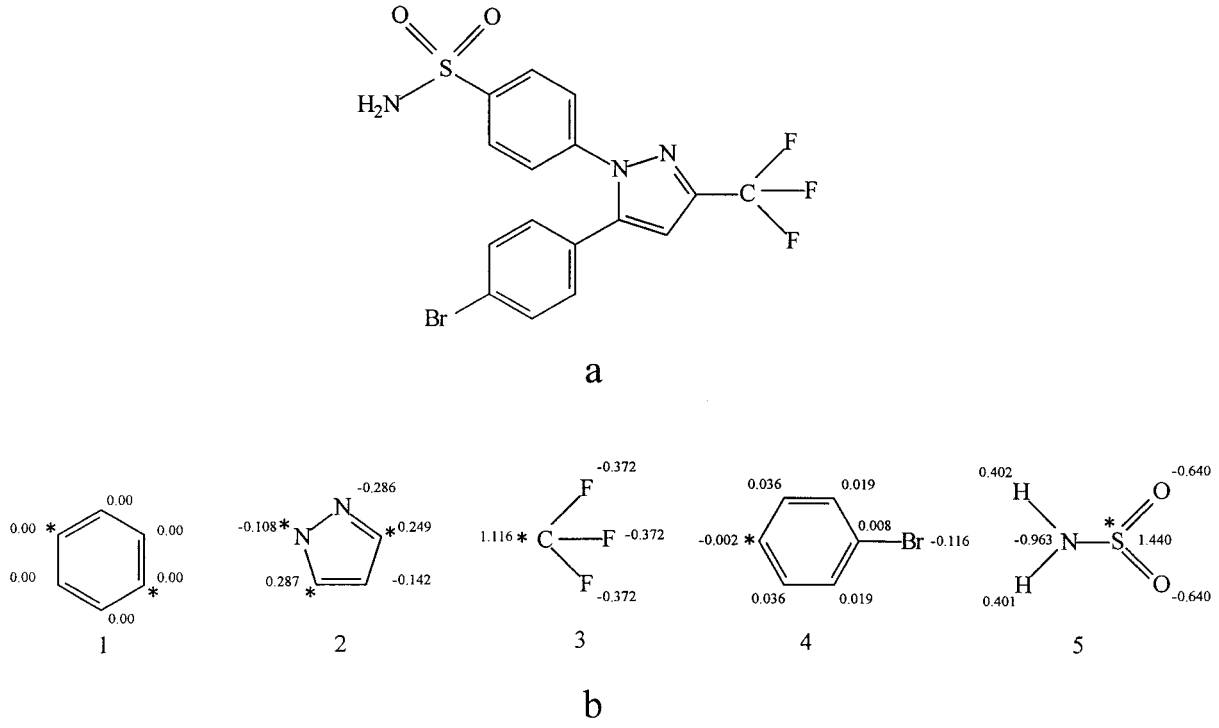


Figure 1. (A) Structure of the selective inhibitor of cyclooxygenase2 – SC-558. (B)The building blocks from SC-558 (1–5), the link site atoms are marked with asterisks (*). Only those linkage appeared in SC-558 are allowed in connecting procedure.

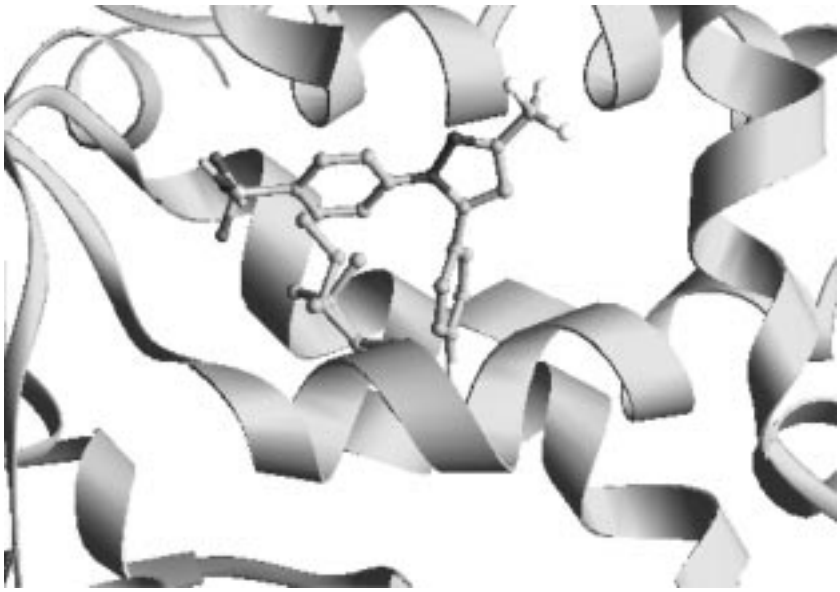


Figure 2. SC-558 from crystal structure complexed within the binding region of COX-2. The Ile^{523} in COX-1 and Val^{523} in COX-2 are explicitly shown.

Improvements on the design strategy and evaluation

Calculation of grid potential using multi-type atomic probe

In the previous version of DycoBlock, the effect of protein atoms beyond the simulation box was calculated using a three-dimensional grid covering the active site. At each grid point, electrostatic potential caused by the receptor atoms was calculated and stored. The Van der Waals interaction of a nonhydrogen ligand atom with these atoms was calculated and assigned to this grid point using a peptide nitrogen atom as a probe. The force and energy of the interaction between atoms of building block copies and receptor atoms beyond the simulation box was simply evaluated by interpolation. This technique speeds up the simulation, however, it seems problematic when the binding energy is used to evaluate the resulting compounds, due to the approximation that all atom types are treated as one type.

The bond type of the linkage between two building blocks uniquely depends on the atom types of the linking atoms, therefore, 43 atom types in the GROMOS96 force field have been extended to 59 to avoid ambiguity. The extended atom types corresponding to those in the GROMOS96 force field are listed in Table 1. The Van der Waals interactions of all 59 types of atoms with the protein atoms beyond the simulation box have been mapped on grid points of the box in the revised DycoBlock. Compared to a single-type atomic probe, using a multi-type atomic probe consumes more time at the beginning of simulation, but it leads to more reasonable results. Two runs with a peptide nitrogen atom as probe and with the 59 atomic types probe in the calculation of potential at grid points have been performed (data not shown). The results indicate that the RMS deviation between the recovered ligands and SC-558 from crystal structure will be significantly reduced by this improvement. Consequently, the resulting compounds can be ordered and evaluated according to their binding energy.

Table 1. The 43 atom types of gromos96 and the extended 59 types in simulation

Atom type number	Force field	
	GROMOS96	Design strategy
1	O	O
2	OM	OM
3	OA	OA
4	OW	OW
5	N	N, NP
6	NT	NT
7	NL	NL
8	NR	NR5, NR6, NR5H, NR6H
9	NZ	NZ
10	NE	NE
11	C	C, C5, C6
12	CH1	CH1, CH1S
13	CH2	CH2, CH25, CH26, CH2S
14	CH3	CH3, CH3S
15	CH4	CH4
16	CR1	CR5, CR6
17	HC	HC
18	H	H
19	DUM	DUM
20	S	S
21	CU1+	CU1+
22	CU2+	CU2+
23	FE	FE
24	ZN2+	ZN2+
25	MG2+	MG2+
26	CA2+	CA2+
27	P/SI	P/SI
28	AR	AR
29	F	F
30	CL	CL
31	BR	BR
32	CMet	CMet
33	OMet	OMet
34	NA+	NA+
35	CL−	CL−
36	CChl	CChl
37	CLChl	CLChl
38	HChl	HChl
39	SDmso	SDmso
40	CDmso	CDmso
41	ODmso	ODmso
42	CCl4	CCl4
43	CLCl4	CLCl4

Automatically replace high-energy building blocks

According to the mean-field theory (MFT), in the multiple-copy simulation [6, 16], the probability density ρ of the whole system can be replaced by a product of two independent probability densities of different subsystems, namely, of the molecular building blocks

(or constructed compounds) and of the receptor. This can be expressed as:

$$\rho(P, Q, t) \approx \rho_s(P_s, Q_s, t)\rho_{N-s}(P_{N-s}, Q_{N-s}, t) \quad (1)$$

where s designates the subsystem of our interest, and $N - s$ denotes the rest of system, the receptor in our case. Since the protein was treated as a rigid structure, no energy transfer took place. Furthermore, ρ_s can be expanded in terms of δ-functions, as

$$\rho_s(P_s, Q_s, t) = \sum_w w\delta_w[P_s - P^o_{s,w}(t), Q_s - Q^o_{s,w}(t)], \quad (2)$$

where w represents the weight factor determined by the initial form of the distribution. It is clear that the more copies of building blocks are put into the active region of receptor, the more efficient the sampling will be. However, many more copies will result in a huge computational burden for both the computer memory and computing time. Automatically replacing high-energy building blocks, in this sense, is identical to increasing the copies of building blocks, but without any additional computational expense. Based on this consideration, in the current work, if the interaction energy of building blocks with receptor atoms is higher than the threshold, they will be automatically replaced during the search for binding sites. The sampling efficiency was greatly enhanced in this way.

Shake for bonds in building blocks and in candidate compounds

At the beginning of step one, many copies of the molecular building blocks are put at randomly chosen positions and with random orientations in the simulation box. A great proportion of them will suffer from bad interactions with the receptor atoms in the box. A bad interaction leads to two cases. In the first case, the building blocks with higher energy than the given threshold will be replaced by newly generated ones; in the second case, building blocks whose energies are lower than the threshold will be optimized continuously during the simulation and search for their binding sites. For the second case, the structures of building blocks are most likely to be distorted by the unfavorable interactions at the beginning of simulation. Without any bond constraints, it will take many MCSMD steps for building blocks to be optimized and to adjust the conformations. On the other hand, it is still problematic whether the binding sites found by the distorted building blocks are reliable. It seems true that the linkage between two possible linking atoms is sensitive to their relative distance, therefore, it is also uncertain whether the distance between two atoms of distorted building blocks can be used as the basis for connecting.

For all the reasons above, the standard constraint method (SHAKE) [17] has been incorporated into the revised DycoBlock. It is used to impose bond length constraints onto not only the building blocks but the candidate compounds as well. The bond length constraints are applied to all bonds except those between two building blocks, which will be optimized little by little during the simulation.

Simulated annealing (SA)

Simulated annealing is an approach for global optimization [15]. In order to enhance the efficiency of sampling, a simple linear simulated annealing method is used in both the optimization of constructed compounds and the search of the binding sites of building blocks.

Gromos96 force field

In the revised DycoBlock, a standard molecular mechanics force field (GROMOS96 [18]) was used to replace the previous one (GROMOS87 [19]).

Calculation of solvent accessible surface area (SASA)

An analytical algorithm for the calculation of solvent accessible surface area [20] has been incorporated into the revised DycoBlock. The atomic radii for all atom types are determined according to the GROMOS96 force field. The radius of a water molecule is set to 1.4 Å. When the constructed compounds are ordered according to the binding energy, their SASAs are also presented, not only in the binding but also in the free state. There is no doubt that a tight binding mode of the ligand will lead to less solvent exposure characterized by a smaller SASA value in the complex.

Clustering algorithm based on molecular similarity

In order to deal with a large number of resulting compounds, a clustering algorithm based on molecular similarity has been developed to classify them into groups according to their structural similarity.

(1) *Molecular similarity matrix*

Constructing the molecular similarity matrix, which contains the similarities between each pair of compounds, is a necessary first step in the clustering procedure. In this study, we adopted a simple definition of similarity, which can be calculated rapidly. Otherwise, as the number of compounds grows to some degree, the computing time will become too expensive to be handled. The similarity between two compounds A and B is defined as

$$S_{A\leftrightarrow B} = \frac{2 \times \sum_{i}^{Ns} Ii}{N_A + N_B} \tag{3}$$

where N_A is the total number of atoms in compounds A, N_B is the total number of atoms in compounds B. N_s is the total number of atoms with the same cell number (see below). I_i is the index of similarity for atom *i*. It will be calculated by the following strategy.

First, the smallest box containing all compounds will be determined. Then, it will be divided into $0.7 \times 0.7 \times 0.7$ Å^3 cubic cells to make sure that the longest diagonal line of each cell is at most equal to the shortest bond length in the force field. Each atom of a specific compound will be marked with a unique cell number in which the atom is located.

Atoms are classified into three types according to their polarity. (1) Polar atoms: fluorine (F), chlorine (Cl), bromine (Br) and oxygen (O) are marked with type number -2, and the nitrogen (N) is marked with type number -1. (2) Non-polar atoms: in this study, all carbon types (C) in force field are treated as the same type and marked with type number 2, sulfide (S) is marked with type number 1 because of its positive charge. (3) Hydrogen atoms: all hydrogen atoms except those in united atoms will be marked with type number 0.

Second, atoms with the same cell number in two compounds are compared. If their type numbers are the same, I_i is set to 1; if their product is more than zero, I_i is set to 0.50.

Thirdly, the sum of I_i for all atoms with the same cell numbers in two compounds will be used to calculate the similarity between them.

(2) *Clustering algorithm*

At first, the compounds for clustering are ordered according to their binding energy and put into a large pool. The first compound will be picked out from the pool and put into an empty cluster, then the compound having the largest similarity value with the former one will be sought out from the pool. If the value is more than the user-specified threshold, this compound will be added to the cluster and removed from the pool. A new cluster will not be constructed until no compound can be found matching the request. This process will be iterated until all compounds have been removed from the pool.

The algorithm used here is so simple that it suffers two major deficiencies. One is that the result is dependent on the initial sequence of the compounds. The other is that although the threshold merely concerns the compounds most recently selected from the pool, it determines the size of whole cluster. Therefore, as the cluster grows larger, the center of the cluster is drifting away from the initially selected compound. It is not surprising that the similarity value between two randomly chosen compounds in one cluster is less than the threshold. Fortunately, the algorithm seems effective enough in our case to discriminate the different structural motifs.

General designing strategy

In our previous work, the molecular building blocks are constructed as initial structures for the step one. However, in the revised DycoBlock, we adopted a new strategy, namely, a preliminary step will be first performed before step one, called step zero.

In a preliminary step before the simulation, single point (SP) energy will be calculated for the constructed molecular building blocks with 6–31 g(d) basis set by using Gaussian98 package [21]. The atomic charges from the calculation will be assigned to every atom except hydrogen atoms. Polar hydrogen atoms will be explicitly treated with their atomic charges assigned, whereas, non-polar hydrogen atoms will be merged into heavy atoms. The charge of the united atom comes from the sum of atomic charges. In the present application, other parameters such as bond lengths, bond angles and (improper) dihedral angles are chosen from the GROMOS96 force field, which is augmented when necessary by information from *ab initio* calculation, and are specified in the molecular topology file. Then, an energy minimization (EM) will be performed for the molecular building blocks. The steepest-descents (SD) and the conjugated-gradients (CG) minimization methods can be specified to do EM. In the current work, 1000 CG optimization were performed. During the optimization, the structures of molecular building blocks will be relaxed. The optimized coordinates and intramolecular potential

energies will not be recorded until the energy difference between two succeeding steps is less than the given threshold. The optimized coordinates are stored as standard structures, which will be used as initial structures or to replace high-energy ones in step one.

At step one, a multiple copy [6] stochastic molecular dynamics [22] (MCSMD) simulation combined with simulated annealing [15] was carried out for 10 000 steps from 1800 K down to 10 K to search for the binding sites of the building blocks. The number of copies of each building block was 200. The time step used was 0.5 fs and the SHAKE [17] algorithm was used to constrain the bond lengths with a tolerance of 0.0001. The friction coefficient used was 80 ps^{-1} for all atoms. The simulation box containing the active sites $(20 \times 20 \times 20$ Å$)^3$ centered at the geometric center of the C_α atoms of the residues Phe^{381}, Leu^{384}, Tyr^{385}, Trp^{387}, Met^{113}, Val^{116}, Val^{349}, Tyr^{355}, His^{90}, Gln^{192}, Arg^{513}, Val^{530}. It was divided into $1 \times 1 \times 1$ $Å^3$ cubic cells. Every 100 MCSMD steps, the conformations of the building block copies with low interaction energy were clustered and recorded as seeds, from which the separate building block copies would be linked together to produce a new compound in the next step.

At step two, MCSMD simulation was continued for 2500 steps at a constant temperature of 300 K, with a time step of 0.5 fs. This step was divided into 25 constructing and optimizing units (CO unit, 100 steps per unit). Given the distribution and conformations of all the simulated building blocks at the time of connection, different possible combinations of building blocks are explored and building blocks are dynamically assembled into candidate compounds using the special algorithm incorporated in DycoBlock [1]. The maximal number of building block copies in a single compound was set at 5 for the recovery study and 4, 5 and 6 for producing structures and binding modes different from SC-558.

At step three, 5 ps stochastic molecular dynamics simulation combined with linear simulated annealing (LSA) (from 300 K to 10 K) was carried out to optimize the structures of the constructed compounds. In the end, the resulting compounds were ordered according to their binding energies, and at the same time free and binding SASAs of each compound were also calculated. Compared to the previous steps, this procedure usually takes relatively less computing time from minutes to several hours, determined by the number of compounds.

Results and discussion

Reassemble the known ligand

The structure of COX-2 was taken from the Protein Data Bank (PDB) (id 1cx2) and all the hydrogen atoms were first deleted. Then, the hydrogen atoms that are explicitly treated were regenerated by PROGCH in the GROMOS96 package [18]. A steepest-descent (SD) energy minimization was used to relax the receptor. SC-558 was broken into five moderate building blocks as shown in Figure 1b. Two testing runs were performed with the old version of DycoBlock (first run) and the improved version (second run). In this study, only the linkages that exist in SC-558 were allowed to form during the connecting procedure, and only compounds with five building blocks were recorded for comparison.

The results from the first run are listed in Table 2. There are 24 resulting compounds in total, in which the compounds 4, 11, 14 and 19 are recovered SC-558. However, if the 24 compounds were ranked by binding energy, no compounds corresponding to SC-558 could be found within the first three compounds having the lowest binding energies. It seems that the ligands in the crystal structure can't be recognized by the binding energy, if a peptide nitrogen was used as a probe to calculate the grid potential. It also indicates that the ligand in the crystal structure can't be regenerated as a dominant distribution by the old version of DycoBlock. The ratio of recovery is a mere 16.7%. In addition, the RMS deviations of the regenerated ligands with respect to the crystal structure are still large, 0.886 Å, 0.920 Å, 1.043 Å and 1.018 Å, respectively. Least squares fit was not performed on the compound when calculating RMS deviation, since we cared about whether the ligand structure and binding mode can be regenerated simultaneously.

The results from the new revised DycoBlock are rather encouraging (shown in Table 3). The total number of resulting compounds decreased to 17, on the contrary, the number of recovered ligands grew to 10, so the ratio of recovery reached 58.8%. The ten recovered ligands together with SC-558 are shown in Figure 3. Regenerated ligands are almost the same as SC-558 in the crystal structure, not only their molecular skeletons, but also the orientations of the branching groups such as the trifluoromethyl and sulphonamide. In contrast with the results from the old version, the recovered ligands occupy the top four places in the compound list with lowest binding energy. The RMS

Table 2. Candidate compounds ordered with respect to binding energies

	E_{tot}	E_{bind}	Explicit+grid (kT/mol)		SASA (Å^2)		RMSD (Å)
			E_{ele}	E_{lj}	Bind	Free	
1	318.037	−263.495	−90.769	−172.726	3.949	698.534	3.747
2	310.688	−260.626	−87.561	−173.065	4.493	699.341	3.756
3	233.796	−253.694	−107.805	−145.888	2.865	702.854	2.784
4*	320.468	−252.709	−55.689	−197.021	2.024	692.910	0.886
5	251.453	−251.515	−107.933	−143.582	5.000	700.106	2.772
6	321.932	−250.942	−109.426	−141.516	3.041	698.418	2.844
7	318.130	−250.915	−70.175	−180.740	3.692	695.485	3.710
8	305.294	−248.347	−65.457	−182.891	4.615	697.962	3.727
9	368.202	−247.773	−100.558	−147.216	2.510	697.737	2.791
10	276.440	−246.527	−114.721	−131.806	4.607	700.764	2.825
11*	263.647	−246.142	−44.300	−201.842	1.946	698.103	0.920
12	201.021	−245.466	−85.472	−159.994	3.109	698.055	2.774
13	279.164	−245.442	−120.901	−124.542	6.584	696.724	2.822
14*	241.983	−241.693	−58.475	−183.218	2.044	690.348	1.043
15	186.322	−240.840	−83.877	−156.964	3.506	702.646	2.798
16	332.812	−240.768	−63.618	−177.150	4.169	698.892	3.743
17	363.254	−239.045	−96.155	−142.890	2.446	697.765	2.790
18	245.890	−238.924	−68.230	−170.694	4.527	698.180	3.743
19*	253.359	−237.371	−46.927	−190.444	3.059	698.920	1.018
20	214.097	−233.651	−79.325	−154.327	4.099	695.269	2.786
21	202.658	−232.362	−77.376	−154.986	2.663	697.982	2.791
22	234.756	−230.877	−86.132	−144.745	3.290	701.994	2.810
23	368.751	−227.881	−97.957	−129.924	3.290	695.426	2.782
24	369.149	−225.087	−89.401	−135.687	5.162	697.749	2.948

[a]The sequence numbers of recovered compounds are marked with asterisk (*).
[b]The compounds listed in this table and their energy terms were generated by the old version of DycoBlock.

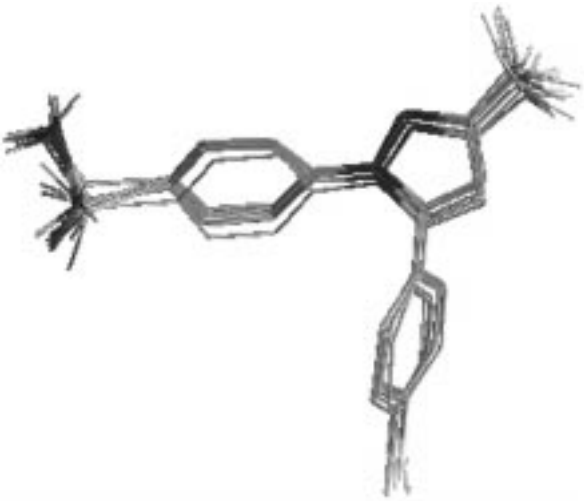


Figure 3. Superimposition of the recovered ligands by revised DycoBlock and SC-558 from crystal structure. The position root-mean-square deviation (RMSD) between them and the native one ranges from 0.616 Å to 1.006 Å.

deviations of the recovered inhibitors with respect to the crystal structure are remarkably less than in the first run. The RMS deviation of the thirteenth compound is 1.006 Å while all other RMS deviation values are less than 1 Å (without fitting to crystal structure). The third and the fourth compounds are most similar to SC-558 in the crystal structure, with the same RMS deviation (0.616 Å). It is not surprising that the SASAs of all recovered inhibitors range from 1.990 Å^2 to 2.511 Å^2, obviously less than those of other binding modes. It should be attributed to the steric complementarity between ligand and receptor. This finding indicates that the correct binding modes are often associated with less SASA exposed than the incorrect ones, therefore, SASA can be used as a criterion to judge whether the compound fits the active site of the receptor.

The clustering was performed for the second set to test the reasonableness of the algorithm. The similarity

Table 3. Candidate compounds ordered with respect to binding energies

	E_{tot}	E_{bind}	Explicit+grid (kT/mol)		SASA (Å^2)		RMSD (Å)
			E_{ele}	E_{lj}	Bind	Free	
1*	414.371	−259.434	−61.330	−198.104	2.251	693.400	0.927
2*	370.355	−252.419	−52.731	−199.688	2.511	695.204	0.729
3*	430.078	−249.652	−49.474	−200.178	2.407	693.442	0.616
4*	431.521	−245.847	−48.886	−196.961	2.404	693.942	0.616
5	345.731	−244.229	−90.802	−153.428	3.642	695.460	2.785
6	402.483	−234.094	−102.651	−131.443	4.008	697.979	2.786
7*	426.346	−232.333	−36.414	−195.920	2.379	694.150	0.752
8	352.993	−226.906	−74.326	−152.580	3.952	695.566	2.930
9*	342.454	−222.623	−24.257	−198.367	2.264	693.323	0.828
10	465.764	−222.418	−96.286	−126.132	6.503	691.022	2.771
11	376.371	−222.264	−73.907	−148.357	3.484	698.932	2.847
12*	428.968	−220.617	−27.296	−193.321	2.260	692.688	0.863
13*	488.682	−215.647	−28.746	−186.901	1.990	694.124	1.006
14*	396.009	−214.662	−13.759	−200.903	2.122	693.519	0.899
15*	444.535	−214.411	−16.249	−198.162	2.114	694.220	0.830
16	498.560	−211.961	−23.233	−188.729	3.774	683.702	3.358
17	550.566	−210.636	−29.017	−181.619	3.731	684.054	3.375

[a]The sequence numbers of recovered compounds are marked with asterisk (*).
[b]The results listed in this table are generated by the new version of DycoBlock with all improvements done.

threshold is arbitrarily set at 0.70. As a result, five clusters are generated as shown in Table 4, which is further confirmed by the visualization of the structures in each cluster. The first cluster includes ten recovered ligands. It indicates that the clustering algorithm actually discriminates different binding modes correctly.

Diversity of lead compounds

Diversity is the ultimate test for a theoretical prediction of novel ligands. Before the structures of COX-2 were solved, many lead compounds used in the development of current clinical candidates had been discovered. To date new classes of selective ligands for COX-2 were still being reported [23,24], the diversity of possible ligands for COX-2 has been expressed by different classes of known ligands. In addition, the structure of the selective inhibitor SC-558 complexed with COX-2 has been determined [14]. These considerations make COX-2 a good test system not only for the study of recovering a known ligand, but for the diversity of compounds produced by present method as well.

Lead compounds from several different structural classes have been identified and shown to be slow, tight-binding and selective inhibitors for COX-2 [2]. Among the species of inhibitors the diarylheterocycle class of COX-2-selective inhibitors is of great interest because the majority of its component chemical groups can be found from the molecular building blocks in this study. The common structural motifs are a *cis*-stilbene moiety substituted in different pendant phenyl rings with a 4-methylsulfone or sulfonamide substituent. The ring system that is fused to the stilbene framework has been extensively manipulated to include every imaginable heterocyclic and carbocyclic skeleton of varying ring sizes. The oxidation state of the sulfur is a very important determinant of selectivity. Sulfone and sulfonamide are selective for COX-2, whereas sulfoxide and sulfide are not. An apparent explanation for the COX-2 selectivity of this class of compounds is that a 4-methylsulfone or sulfonamide substituent is inserted into the side pocket and forms hydrogen bonds with side chain and main chain residues.

The thirty-three molecular building blocks used in this study are presented in Figure 4. Most of them appear often in known organic compounds. As stated above, the rest of the building blocks are from some known selective inhibitors of COX-2. The resulting compounds with four, five and six molecular building blocks were recorded for comparison with chosen known ligands.

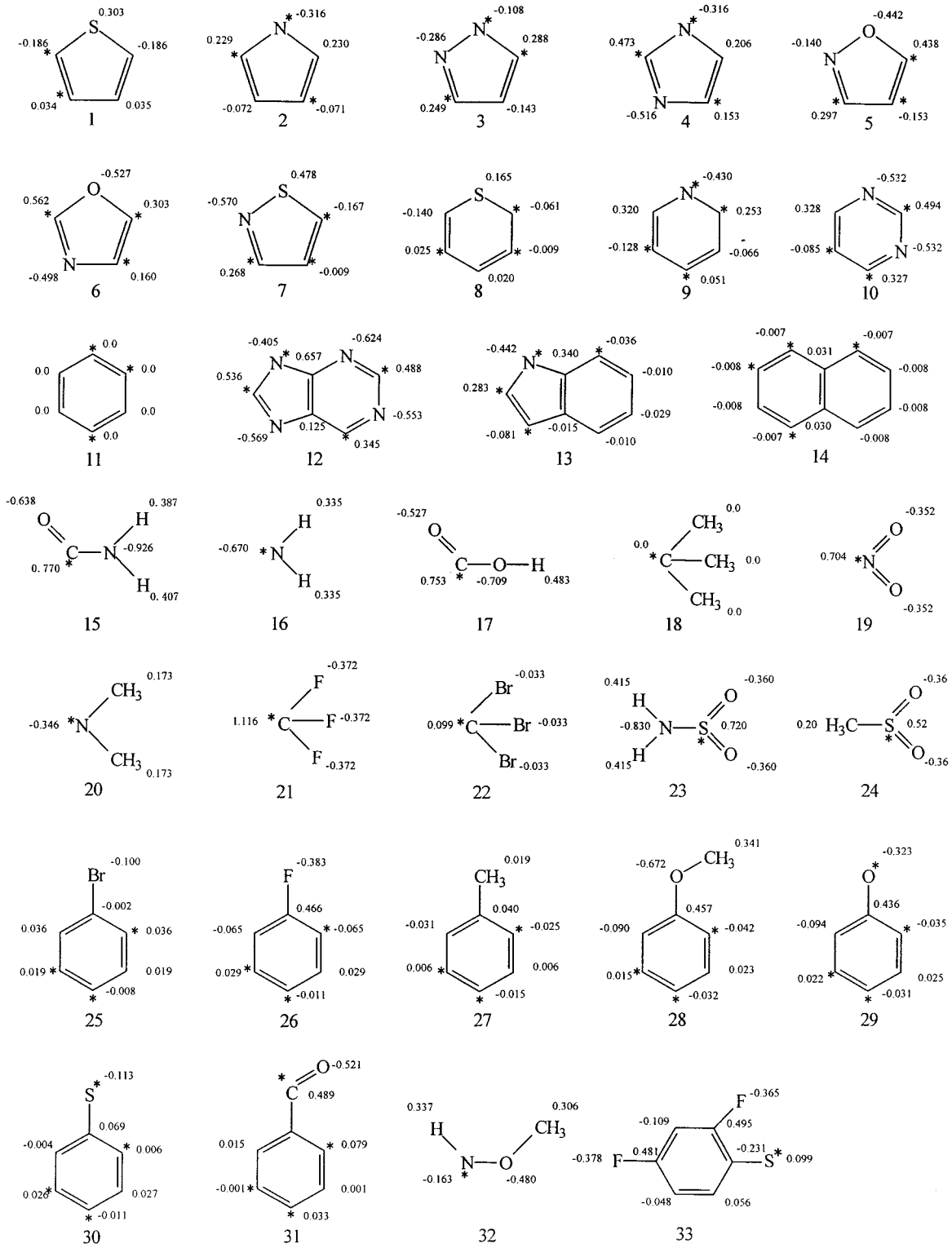


Figure 4. Thirty-three building blocks used in further design, the link site atoms are marked with asterisks (*). Only chemically reasonable linkages are allowed in connecting procedure.

Table 4. Clustering results of constructed compounds in recovery study

Sequence number of clusters	Sequence number of compounds in each cluster
1[a]	1, 2, 3, 4, 6, 9, 12, 13, 14, 15
2	5, 11
3	7, 10
4	8
5	16, 17

[a]The first cluster represents the set of recovered ligands.

Table 5. Clustering results of constructed compounds in recovery study

Cluster	Number of compounds in one cluster		
	4mbb[a]	5mbb	6mbb
1	14	9	5
2	5	6	6
3	8	7	
4	8		
5	5		
6	5		
7	6		

[a]mbb means molecular building blocks.

In our study, 476 compounds with four molecular building blocks, 479 compounds with five building blocks and 486 compounds with six building blocks are generated respectively. In order to exclude unreasonable compounds with bad binding modes or distorted structures, three criteria were chosen for the first screening process. First, the compounds with positive binding energy will be discarded. Second, those compounds characterized by intramolecular potential energy higher than 1000 kJ/mol, which indicated the unfavorable intramolecular interactions, will be also discarded. Third, a large solvent accessible surface area usually suggests that the compound can't fit the active site of the receptor well, even when optimized after a long simulation. Thus, the compound will be excluded if the SASA in the binding state is larger than 5.0 $Å^2$. As a result, there remained 338, 297 and 229 compounds, for four, five and six molecular building blocks respectively.

The similarity threshold used in clustering was set at 0.35, 0.30 and 0.25, respectively, for compounds with four, five and six building blocks. In this study, we focused on the major structural motifs that exist in the large amount of resulting compounds. We found that the lower similarity threshold used here allowed the appearance of the clusters in which the compounds had different number of atoms, atom types, but shared the similar structure. Certainly, some variations in the structure were allowed. The clusters with more than five members were recorded for further analysis.

For the compounds with four building blocks, there are a total of 259 clusters and 7 clusters with more than 5 members. There are 243 clusters for the compounds of five building blocks among which 3 clusters have more than 5 members. It is not surprising that as the number of constituting building blocks grows larger, the similarity between two compounds decreases. The compounds with six building blocks fell into 201 clusters and only 2 clusters have more than 5 members. We found that some clusters shared the same or similar structural motifs with the known inhibitors, while others seemed to be novel. The results are presented in Table 5.

For the compounds with four building blocks, the largest cluster (with fourteen compounds) has a similar structural motif to SC-558. However, their structures are not identical to that of SC-558 but vary a lot at the position corresponding to the bromophenyl ring of SC-558. The phenyl ring of SC-558, to which the sulphonamide attaches, is replaced by five-membered heterocyclic rings. With the number of building blocks limited to 4, almost no polar groups can be found corresponding to the trifluoromethyl group of SC-558. After screening, compounds ranking in the top one hundred are shown in Table 6. The average properties of recovered SC-558 are also listed for comparison. There are ten compounds with lower binding energies than the average value. The superposition of the compounds is illustrated in Figure 5.

There are three compounds with fused heterocyclic rings in a cluster with high similarity, which share this characteristic with the known imidazothiazole analog. The structures of three compounds and the known template are illustrated in Figure 6. It is remarkable that three fused heterocyclic rings superimposed well with one another (two indole rings and one purine ring), while different sizes of the pyrrole, oxazole and six-membered heterocyclic rings led to different depth of insertion of the 4-methylsulfone and sulphonamide into the pocket. However, there is no doubt that all selective parts extend into the selective pocket exactly, which was further confirmed by the visualization of the three compounds binding in the active sites of COX-2 with Val^{523} and Ile^{523} presented (Figure 7).

Table 6. Candidate compounds ordered with respect to binding energies (4 building blocks)

	Rank[a]	E_{tot}	E_{bind}	Explicit+grid (kT/mol)		SASA (Å^2)	
				E_{ele}	E_{lj}	Bind	Free
Structural motif similar to SC-558							
Cmp433	7	−84.957	−295.263	−78.236	−217.027	0.868	592.277
Cmp138	16	−385.711	−278.516	−101.506	−177.010	2.273	547.421
Cmp336	17	−60.714	−277.146	−82.375	−194.771	2.127	522.863
Cmp48	51	−158.770	−259.467	−73.249	−186.217	1.860	525.423
Cmp448	53	−28.184	−259.048	−49.031	−210.018	1.942	544.981
Cmp206	59	104.013	−256.793	−102.233	−154.560	1.907	521.675
Cmp169	94	−143.160	−246.763	−65.281	−181.482	2.159	525.304
Cmp42	114	−454.559	−240.769	−61.003	−179.766	2.827	548.058
Cmp7	138	−300.105	−235.382	−85.492	−149.890	2.054	470.995
Cmp150	146	−104.235	−233.474	−53.404	−180.070	2.115	577.307
Cmp80	176	−89.993	−226.070	−58.069	−168.001	1.998	527.729
Cmp314	180	−189.479	−225.427	−44.442	−180.985	2.282	550.725
Cmp103	192	−132.168	−223.244	−52.108	−171.136	2.293	532.675
Cmp32	211	461.770	−218.629	−36.539	−182.090	3.680	577.972
SC-558[†]		−417.332	−232.765	−35.914	−196.851	2.270	693.801
Structural motif similar to imidazothiazole analog							
Cmp68	23	324.507	−275.274	−41.831	−233.442	2.832	610.869
Cmp315	36	−594.928	−266.066	−38.100	−227.966	2.978	606.191
Cmp59	100	−642.041	−245.423	−30.145	−215.278	1.037	575.067

[†]The average properties of recovered SC-558.
[a]These compounds are named after the original sequence number of them in the resulting compounds, however, the rank numbers are from the remaining compounds after the first screening.

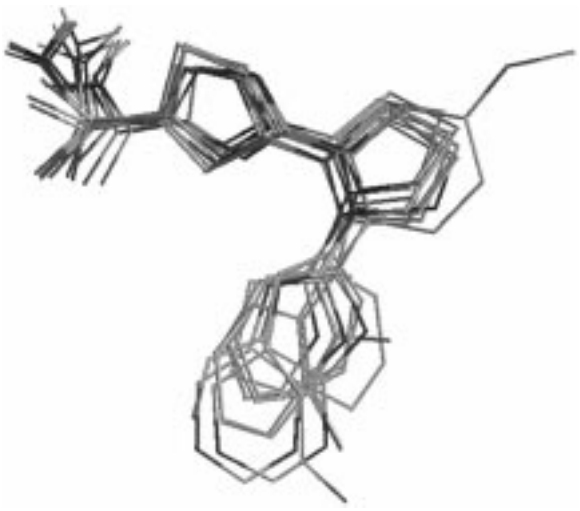
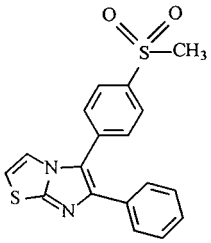


Figure 5. Superposition of ten ligands chosen from a known class of slow, tight-binding and selective inhibitors for COX-2. They are characterized by a selective substituent – sulfone or sulfonamide and diarylheterocycles.

It is notable that the known inhibitors, such as SC-58092 and SC-58076, which consist of a thiazole and three phenyl rings, express different selectivity for COX-1 and COX-2 when the *R* changes from SO_2 to S [2]. In our results, we found one compound with five building blocks and four compounds with six building blocks, all are characterized by a central pyrazole ring and three phenyl or heterocyclic rings around. It is interesting that all selective parts appeared in these compounds are 4-methylsulfones, namely, *R* is SO_2, except that the seventh compound with a chemically similar group (Figure 8). The results are consistent with known experimental results. The results of detailed comparison are presented in Table 7.

Prediction of structure of COX-2 in complex with known selective inhibitors

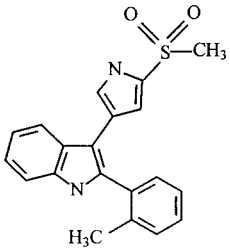
At present, the structures of COX-2 complexed with flurbiprofen, indomethacin and SC-558 – the selective COX-2 inhibitor – have been determined, but the binding modes of other ligands especially the selective inhibitors are still unknown. Our method provides a



Imidazothiazole analog

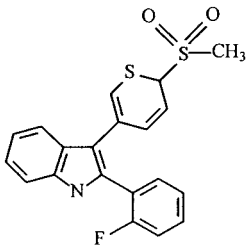
IC_{50} (COX-2) ~ 0.016 μ M

IC_{50} (COX-2) > 50 μ M



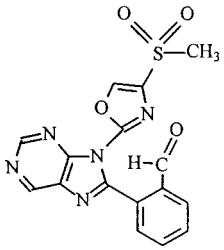
Cmp315

Seq : 36



Cmp59

Seq : 100



Cmp68

Seq : 307

Figure 6. The structures of three compounds and the known template with fused heterocyclic rings in a cluster with high similartiy, which share this characteristic with the known imidazothiazole analog. The three compounds are named after the original sequence number of all the compounds before screening. The sequence numbers of them after screening and ordering are presented. Cmp315, Cmp59 and Cmp68 are from the compounds with 4 building blocks.

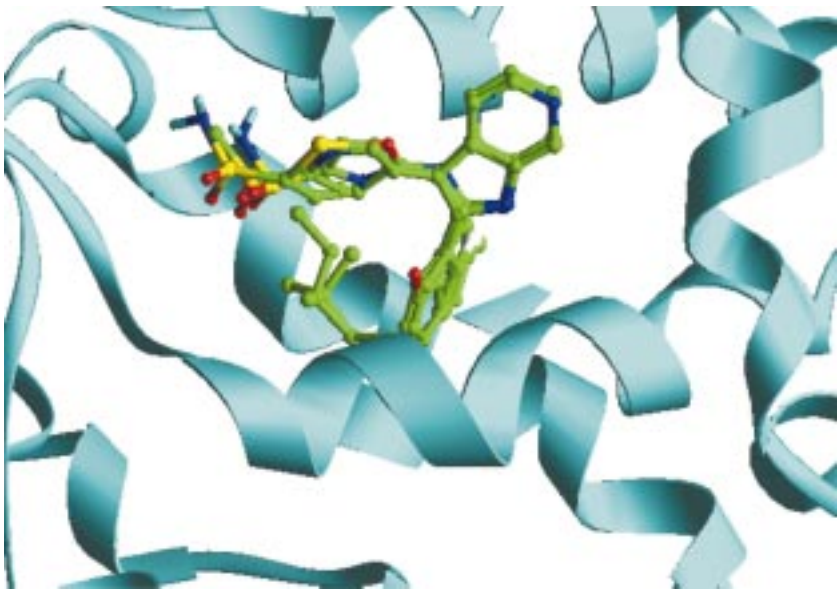


Figure 7. Three generated compounds with four building blocks (with sequence number 59, 68 and 315) complexed within the binding region of COX-2. The Ile^{523} in COX-1 and Val^{523} in COX-2 are explicitly shown. The selective groups of three compounds all insert into the pocket, bordered by Val^{523}.

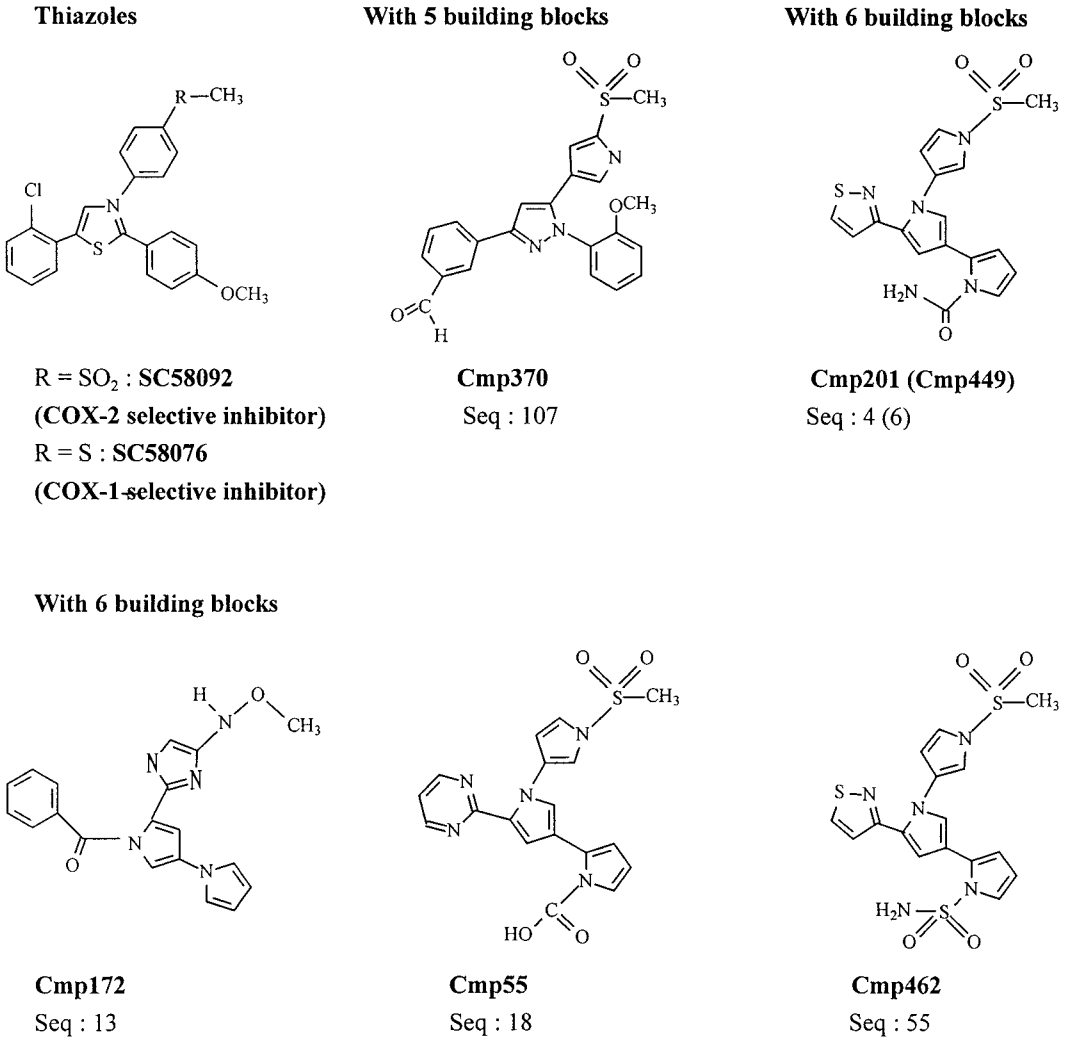


Figure 8. Superposition of the compound (Cmp370) with five building blocks and four compounds with six building blocks Cmp201 (Cmp449), Cmp172, Cmp55, Cmp462 and their template. They are characterized by a central pyrazole rings and three phenyl or heterocyclic rings around.

Table 7. Candidate compounds ordered with respect to binding energies

				Explicit+grid (kT/mol)		SASA (Å^2)	
	Rank[a]	E_{tot}	E_{bind}	E_{ele}	E_{lj}	Bind	Free
Cmp370[b]	107	−121.933	−263.963	−55.964	−207.998	3.362	706.683
Cmp201	4	−417.954	−366.388	−137.807	−228.581	3.100	654.648
Cmp449	6	−395.763	−357.639	−126.839	−230.800	3.042	651.303
Cmp172	13	159.410	−343.627	−149.652	−193.975	4.924	664.488
Cmp55	18	−376.531	−334.385	−103.704	−230.681	3.114	654.043
Cmp462	55	407.809	−301.276	−89.891	−211.385	3.730	664.243

[a]These compounds are named after the original sequence number of them in the resulting compounds, however, the rank numbers are from the remaining compounds after the first screening.
[b]Cmp370 is made up of five building blocks. Because it shares the same character of four rings with other compounds of six building blocks, it is also listed in this table.

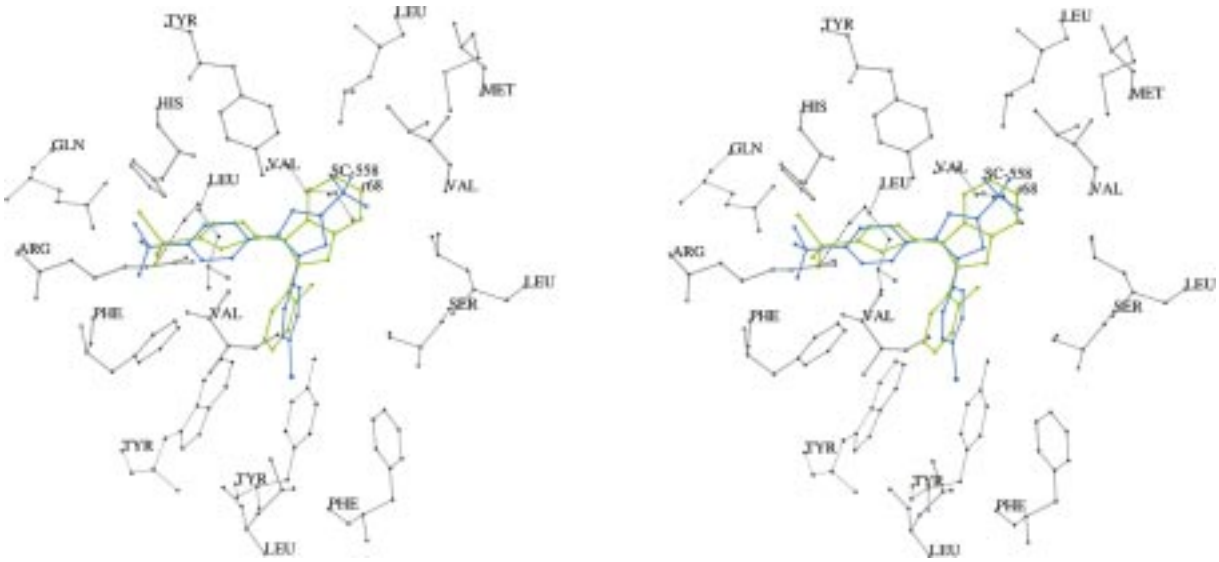


Figure 9. The stereoview of complex structure of SC-558 and Cmp68 with four building blocks that represent the known imidazothiazole analog. The SC-558 is presented with the color of cornflowerblue while Cmp68(named r68) is presented with chartreuse. The dashed line represents the hydrogen bond formed between Cmp68 and Leu^{352}.

useful tool to study structural basis of inhibition and binding modes, and to understand the mechanism of inhibition.

The compounds in which structural motifs and binding modes are similar to the known ligands will be good argument for the diversity of this method, therefore, they are of special interest to us. Additionally, the co-crystal structures of them with the enzyme are still unknown. The interaction between SC-558 and the active site of COX-2 has been studied. The structures of SC-558 and Cmp68 with four building blocks that represent the known imidazothiazole analog are compared (Figure 9). The bromophenyl ring of SC-558 is bound in a hydrophobic cavity formed by Phe^{381}, Leu^{384}, Tyr^{385}, Trp^{387}, Phe^{513} and Ser^{530}. The toluene ring of Cmp68 is also bound in this pocket and the methyl inserts into a small cavity, the bottom of which is formed by Val^{344} and Val^{349} and surrounded by Tyr^{348}, Tyr^{385}, Ser^{530}, Leu^{531}, Leu^{534} and Phe^{205}. The trifluoromethyl group of SC-558 is bound in an adjacent pocket formed by Met^{113}, Val^{116}, Val^{349}, Tyr^{355}, Leu^{359} and Leu^{531}. Due to the large cavity of this pocket, the purine ring of Cmp68 can be accommodated. The entire phenylsulphonamide binds in a cavity, where the phenyl ring interacts with the backbone of $Ser^{353}$and hydrophobic residues like Leu^{352}, Tyr^{355}, Phe^{518}, Val^{523}, and the selective sulphonamide interacts with His^{90}, Gln^{192} and Arg^{513}. One of the oxygen atoms forms a hydrogen bond to His^{90}; the other oxygen is linked by a hydrogen bond to Arg^{513}. The amide nitrogen forms a hydrogen bond to the carbonyl oxygen of Phe^{518}. Incorporating a pyrrole ring, which is smaller than the phenyl ring in SC-558, leads to the observation that the 4-methylsulfone just extends into the selective pocket. The depth of insertion is not the same as in SC-558. One hydrogen bond is also formed between the oxygen atoms of 4-methylsulfone in Cmp68 and Leu^{352} in the active sites. Moreover, the bulkier methyl group makes the 4-methylsulfone stable in the pocket. It is clear that Cmp68 and SC-558 share basically the same binding sites though different building blocks form the two compounds.

Conclusion

Some de novo design methods have been described in the literature, such as CONCERTS [8] and HOOK [7]. Compared to other methods, some new features can be found in DycoBlock. First, the method of linking the functional groups by bridges was not adopted. This avoids the introduction of a specific skeleton database. Moreover, as the number of combinational possibilities increases, the diversity of the compounds is dramatically enhanced. The most important feature of DycoBlock is that the whole design process is dynamic. It is a dynamic process to search for binding

sites during a long MCSMD + LSA simulation. At the second step, these separate functional groups are not fixed at their binding sites during the connecting process. For DycoBlock the assembling is also a dynamic process, which will result in optimal compounds. Functional groups can move and adjust their conformations during the connecting process. Different linking pathways will be tried in a single step within each CO unit.

In this work, we make special efforts to enhance the design efficiency, feasibility, and the reasonableness as well. Regenerating a selective inhibitor in the active site with the correct binding mode is a rigorous criterion for de novo design methods. It is very encouraging that with the least number of types (5 types) and building block copies (200 copies for each type), the selective inhibitor SC-558 of COX-2 was regenerated with a high ratio. Not only the structures but also the binding positions are consistent with the experimental data. It is worth noting that the selective functional group is bound in the specific pocket with correct binding mode.

A clustering algorithm has been applied in this study to classify the resulting compounds. Binding energy, as calculated by a molecular force field, and solvent accessible surface area were used to evaluate the binding affinity and further rank the compounds, and recognize the favorable binding modes.

There are still some drawbacks in this version of DycoBlock. First, the solvent effect and flexibility of the receptor not considered in this study, however, actually play an important role in the design procedure. Recently, an implicit solvent model based on the stochastic dynamics (SD) and solvent accessible-surface area (SASA) model developed in this group seemed to provide a possible treatment of solvent effect in the DycoBlock. Moreover, the flexibility of receptor has been considered in a new version of DycoBlock, the results will be reported elsewhere. Although some algorithms has been developed in this work to screen the resulting compounds, a fast and reliable method for calculating the binding free energy is still necessary and under development.

Acknowledgements

This work was supported by Chinese National High Technology Project (Grant number: 863-103-13-03-01), Chinese National Fundamental Research Project (Grant number: G1999075605) and Chinese National Natural Science Foundation (Key Project. Grant number: 39990600). We gratefully thank van Gunsteren W.F. (Department of Chemistry, ETHZ) for GROMOS96 force field.

References

1. Liu, H.-Y., Duan, Z.-H., Luo, Q.-M. and Shi, Y.-Y., Proteins, 36 (1999) 462.
2. Marnett, L.J. and Kalgutkar, A.S., Curr. Opin. Chem. Biol., 2 (1998) 482.
3. Böhm, H.J., Curr. Opin. Struct. Biol., 7 (1996) 433.
4. Bohacek, R.S. and McMartin, C., J. Am. Chem. Soc., 116 (1994) 5560.
5. Böhm, H.J., J. Comput. Aid. Mol. Des., 6 (1992) 61.
6. Miranker, A. and Karplus, M., Proteins, 11 (1991) 29.
7. Eisen, M.B., Wiley, D.C., Karplus, M. and Hubbard, R.E., Proteins, 19 (1994) 199.
8. Pearlman, D.A. and Murcko, M.A., J. Med. Chem., 39 (1996) 1651.
9. Smith, W.L. and Mamett, L.J., in Sigel, H. and Sigel, A. (Eds), Metal Ions in Biological Systems. Marcel Dekker, New York, NY, 1994, pp. 163–199.
10. Otto, J.C. and Smith, W.L., J. Lipid Mediat. Cell Signal., 12 (1995) 139.
11. DeWitt, D.L., El-Harith, E.A., Kraemer, S.A., Andrews, M.J., Yao, E.F., Armstrong, R.L. and Smith, W.L., J. Biol. Chem., 265 (1990) 5192.
12. Seibert, K., Zhang, Y., Leahy, K., Hauser, S., Masferrer, J., Perkins, W., Lee, L. and Isakson, P., Proc. Natl. Acad. Sci. USA, 91 (1994) 12013.
13. Masferrer, J.L., Zweifel, B.S., Manning, P.T., Hauser, S.D., Leahy, K.M., Smith, W.G., Isakson, P.C. and Seibert, K., Proc. Natl. Acad. Sci. USA, 91 (1994) 3228.
14. Kurumbail, R.G., Stevens, A.M., Gierse, J.K., McDonald, J.J., Stegeman, R.A., Pak, J.Y., Gildehaus, D., Miyashiro, J.M., Penning, T.D., Seibert, K., Isakson, P.C. and Stallings, W.C., Nature, 384 (1996) 644.
15. Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P., Science, 220 (1983) 671.
16. Elber, R. and Karplus, M., J. Am. Chem. Soc., 112 (1990) 9161.
17. Ryckaert, J.P., Ciccotti, G. and Berendsen, H.J.C., J. Comput. Phys., 23 (1977) 327.
18. van Gunsteren, W.F., Billeter, S.R., Eising, A.A., Hunenberger, P.H., Kruger, P., Mark, A.E., Scott, W.R.P. and Tironi, I.G., Biomolecular simulation: the GROMOS96 manual and user guide. University of Groningen, the Netherlands, ETH Zürich, Switzerland: Biomos; 1996.
19. van Gunsteren, W.F. and Berendsen, H.J.C., Groningen molecular simulation (GROMOS) laboratory manual, University of Groningen, the Netherlands: Biomos; 1987.
20. Richmond, T.J., J. Mol. Biol., 178 (1984) 63.
21. Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Zakrzewski, V.G., Montgomery, J.A., Jr., Stratmann, R.E., Burant, J.C., Dapprich, S., Millam, J.M., Daniels, A.D., Kudin, K.N., Strain, M.C., Farkas, O., Tomasi, J., Barone, V., Cossi, M., Cammi, R., Mennucci, B., Pomelli, C., Adamo, C., Clifford, S., Ochterski, J., Petersson, G.A., Ayala, P.Y., Cui, Q., Morokuma, K., Malick, D.K., Rabuck, A.D., Raghavachari, K., Foresman, J.B.,

Cioslowski, J., Ortiz, J.V., Baboul, A.G., Stefanov, B.B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Gomperts, R., Martin, R.L., Fox, D.J., Keith, T., Al-Laham, M.A., Peng, C.Y., Nanayakkara, A., Gonzalez, C., Challacombe, M., Gill, P.M.W., Johnson, B., Chen, W., Wong, M.W., Andres, J.L., Gonzalez, C., Head-Gordon, M., Replogle, E.S. and Pople, J.A., Gaussian 98, Revision A.7, Gaussian, Inc., Pittsburgh PA, 1998.

22. van Gunsteren, W.F., Berendsen, H.J.C. and Rullmann, J.A.C., Mol. Phys., 44 (1981) 69.

23. Kalgutkar, A.S., Kozak, K.R., Crews, B.C., Phillip Hochgesang, G., Jr. and Marnett, L.J., J. Med. Chem. 41 (1998) 4800.

24. Puig, C., Crespo, M.I., Godessart, N., Feixas, J., Ibarzo, J., Jiménez, J.-M., Soca, L., Cardelús, I., Heredia, A., Miralpeix, M., Puig, J., Beleta, J., Huerta, J.M., López, M., Segarra, V., Ryder, H. and Palacios, J.M., J. Med. Chem., 43 (2000) 214.