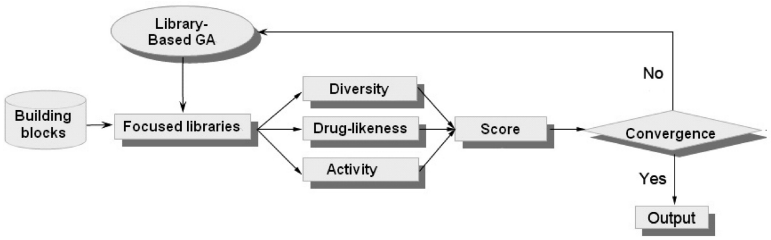
JOURNAL OF combinatorial CHEMISTRY

Article

Focused Combinatorial Library Design Based on Structural Diversity, Druglikeness and Binding Affinity Score

Gang Chen, Suxin Zheng, Xiaomin Luo, Jianhua Shen, Weiliang Zhu, Hong Liu, Chunshan Gui, Jian Zhang, Mingyue Zheng, Chum Mok Puah, Kaixian Chen, and Hualiang Jiang

J. Comb. Chem., **2005**, 7 (3), 398-406• DOI: 10.1021/cc049866h • Publication Date (Web): 18 March 2005



More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 3 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML

Journal of Combinatorial Chemistry is published by the American Chemical Society.
1155 Sixteenth Street N.W., Washington, DC 20036

Focused Combinatorial Library Design Based on Structural Diversity, Druglikeness and Binding Affinity Score

Gang Chen,[†,‡] Suxin Zheng,[†] Xiaomin Luo,[*,†] Jianhua Shen,[†] Weiliang Zhu,[*,†] Hong Liu,[†] Chunshan Gui,[†] Jian Zhang,[†] Mingyue Zheng,[†] Chum Mok Puah,[‡] Kaixian Chen,[†] and Hualiang Jiang[*,†,§]

Drug Discovery and Design Center and State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Shanghai Institutes for Biological Sciences, and Graduate School, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, P. R. China, Technology Centre for Life Sciences, Singapore Polytechnic, 500 Dover Road, Singapore 139651, Singapore, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China

Received August 11, 2004

The advent of focused library and virtual screening has reduced the disadvantage of combinatorial chemistry and changed it to a realizable and cost-effective tool in drug discovery. Usually, genetic algorithms (GAs) are used to quickly finding high-scoring molecules by sampling a small subset of the total combinatorial space. Therefore, scoring functions play essential roles in focused library design. Reported here is our initial attempt to establish a new approach for generating a target-focused library using the combination of the scores of structural diversity and binding affinity with our newly improved druglikeness scoring functions. Meanwhile, a software package, named LD1.0, was developed on the basis of the new approach. One test on a cyclooxygenase (COX)2-focused library successfully reproduced the structures that have been experimentally studied as COX2-selective inhibitors. Another test is on a peroxisome proliferator-activated receptors γ-focused library design, which not only reproduces the key fragments in the approved (thiazolidinedione) TZD drugs, but also generates some new structures that are more active than the approved drugs or published ligands. Both of the two tests took ∼15% of the running time of the ordinary molecular docking method. Thus, our new approach is an effective, reliable, and practical way for building up a properly sized focused library with a high hit rate, novel structure, and good ADME/T profile.

Introduction

The advent of combinatorial chemistry is one of the most exciting developments in medicinal chemistry in the past decade.[1−3] Coupled with automation technologies and high-throughput screening, it offers great potential for discovering new drug leads. This technology allows thousands or even millions of compounds to be synthesized at the same time; however, many products in the huge library are redundant. It also does not make sense to validate and assay millions of compounds. To synthesize a chemical library of reasonable size and considerable hit rate, three-dimensional (3D) structural information and properties of a studied receptor should be taken into consideration to filter out redundant compounds.[4,5] Thus, the critical challenges are, first, to select sets of fragments that have the best potential to be parts of new drug leads for a given target and, second, to set up proper criteria for product judgment (screening).

To overcome the first challenge, three types of virtual libraries have been suggested. They are focused libraries, targeted libraries, and primary screening libraries.[6] A focused library is built on the basis of a lead molecule or pharmacophore and geared toward one particular molecular target. A targeted library is designed for finding drug leads against specific targets. A primary screening libraries is a large combinatorial library used to randomly find new hits or to design novel scaffolds.[6] To solve the second problem, druglikeness and structural diversity have been introduced into the library design to reduce its size and increase its efficiency.[4,7] Initially, the focus in combinatorial library design was on selecting diverse sets of compounds on the assumption that maximizing diversity would result in a broad coverage of bioactivity space and, hence, would maximize the chances of finding drug leads.[7] The creation of diversity by compound libraries has been a central claim and task of combinatorial chemistry since its inception. Suggestions and assumptions on how to assess diversity have been studied during the past decade.[8−18]

Druglikeness is another key factor that needs to be considered during library design.[19−23] It has been estimated that ∼40% of compounds fail to be developed into drugs due to their poor pharmacokinetic properties.[24] Therefore, initial strategies toward this goal should be involved in the use of computational filters to remove compounds deemed to be chemically unsuitable for drug development. This approach has already been applied by several research

* Corresponding authors. Phone: +86-21-50806600, ext 1210. Fax: +86-21-50807088. E-mail (Jiang): hljiang@mail.shcnc.ac.cn.
† Chinese Academy of Sciences.
‡ Singapore Polytechnic.
§ East China University of Science and Technology.

10.1021/cc049866h CCC: $30.25
Published on Web 03/18/2005

groups.[25−27] In 1997, a set of assumptions about necessary features for a "good" drug candidate was suggested and embodied in a so-called "Rule of Five" by Lipinski and co-workers.[28] Then, other scientists put up new methods to improve prediction performance of the "Rule of Five". J Galvez et al. developed a new method to achieve a pattern of general pharmacological activity based on molecular topology.[29] Some other research groups used a neural network to classify chemical compounds into potentially "druglike" and "nondruglike" categories.[30−32] Druglike index (DLI), which is calculated based upon the knowledge derived from known drugs, was introduced by Xu et al.[33] However, the above methods still have their limitations in virtual screening, such as time-consuming and poor performances. Therefore, how to discriminate a druglike compound from nondruglike is still a great challenge in library design.

A focused library is actually a representative sample of the full chemical space. Thus, some stochastic methods for exploring the whole chemical space should be applied, of which genetic algorithms (GA) and simulated annealing are generally used, because of their high efficiency in searching large combinatorial spaces.[34−38] The results of both depend on particular series of pseudorandom numbers, so multiple runs usually need to be done, and there is no guarantee that the global best solution will be found. However, good results are usually found much more quickly than a purely random search or a systematic search.[39] Hence, library-based GA approach was applied to optimize our focused library.

The aim of this study is to establish a new efficient approach that can be used to build, optimize and assess focused libraries based on the 3D structures of target receptor. To reduce the library size and improve its efficiency, we modified and developed the commonly used criteria for evaluating druglikeness. In addition, molecular docking is used to evaluate the binding potential between target protein and candidate compounds to further reduce library size and enhance its hit rate. On the basis of the new approach, a software package, named LD1.0, was successfully developed. Cyclooxygenase (COX)2- and peroxisome proliferator-activated receptors (PPAR)γ-focused libraries were generated by using LD1.0 with the newly modified descriptors for druglikeness. The result shows that LD1.0 and the modified descriptor set are reliable and would have wide implications for future work in discovering and optimizing drug leads.

Methodology

The Software LD1.0. Figure 1 is a schematic diagram of LD1.0. First, building blocks are selected from given fragment databases to create a serial of virtual libraries according to assigned chemical reaction steps. Second, library-based GA is applied to optimize the virtual libraries. Each library is evaluated by certain criteria, such as docking energy, molecular diversity, and druglikeness. Then, according to the evaluating scores of every library, GA retains some libraries with higher scores and meanwhile creates some new libraries to form the next generation of focused libraries. Finally, GA optimization ends once the termination condition is satisfied.

Library-Based GA. GA is a stochastic optimization method that mimics the process of natural evolution by manipulating a population of data structures called chromosomes. The purpose of this study is to design the best focused library, so a library-based GA approach was applied to optimize the focused library. Library-based GA means that the GA's chromosomes are not molecules, but focused libraries. At first, the module "library-based GA" simulates different chemical reactions among the given fragments to produce various molecules. Since multiple reactions have been considered during the program design, this module can cope with a nine-step reaction. For each step, all atomic types, bond lengths, and bond angles of both the reactants and the products are verified to ensure that chemical reactions take place properly. The final products are kept in corresponding libraries. Then each library is evaluated by different descriptor sets of molecular diversity, druglikeness, and potential bioactivity (binding affinity to receptor). The score from each descriptor is normalized. The final score of each focused library is the sum of the normalized scores from each algorithm multiplied by their weights. On the basis of the scores, GA produces the next generation of libraries by copying, crossing, and mutation. GA optimization keeps running in this way until most of the top-ranking libraries include the same building blocks or the optimization reaches the maximum number of genetic generations. The best library is the product: a successfully constructed focused library.

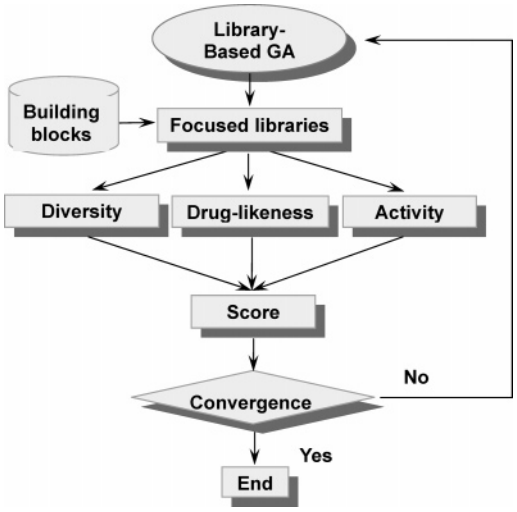


Figure 1. A schematic diagram of the software LD1.0 for the focused library design.

Structural Diversity. It was reported that 2D structural descriptions are fast and accurate for calculating molecular diversity.[12,40] Flower and co-workers found 39 best structural descriptors for calculating molecular diversity,[13] which were also employed in LD1.0. A distance method was selected in this study to estimate molecular structural diversity, for it has considerable physical meanings. First, each descriptor is normalized. Then a Eulide space is described by all descriptors with their weights. So the difference (or similarity), d_{ij}, of two molecules can be presented by the distance between the two molecular points in the space, calculated according to eq 1,

$$d_{ij} = \sqrt{\sum_k (\hat{x}_k^i - \hat{x}_k^j)^2} \qquad (1)$$

Table 1. Descriptor Set for Calculating Druglikeness

descriptor	meaning	range	default weight
XLogP	octanol/water partition coefficient	~−0.5 to 5	0.1
MW	molecular weight	~78 to 500	0.1
HBA	hydrogen bond acceptor	~2 to 10	0.1
HBD	hydrogen bond donor (if no. of HBA is not greater than 10)	≤5	0.1
C3p	ratio of the number of C(sp^3) atoms over the total number of nonhalogen heavy atoms	~0.15 to 0.8	0.2
h_p	ratio of the number of hydrogen atoms over the total number of nonhalogen heavy atoms	~0.6 to 1.6	0.2
unsat_p	ratio of the molecular unsaturation over the total number of nonhalogen heavy atoms	~0.10 to 0.45	0.2

where $\hat{x}_k^j$ is the normalized value of the *k*th descriptor for the *j*th molecule, which can be calculated using eq 2,

$$\hat{x}_k^j = w_k\left(\frac{x_k^j - \bar{x}_k}{\sigma_k}\right) \tag{2}$$

where x_k^j is the value of the *k*th descriptor for the *j*th molecule, $\bar{x}_k$ is the mean value of the descriptor k, σ_k is the variance difference of the descriptor k, and w_k is the weight of the descriptor k. Then a library's diversity, D_k, can be calculated using eq 3,

$$D_k = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{j<i} d_{ij} \tag{3}$$

where *n* is the number of molecules in the library. So D_k is a sum of the distances between any two molecules in a library. The final scores of all libraries in a certain GA generation are normalized using eq 4,

$$D_{k,nor} = \frac{D_k - D_{min}}{D_{max} - D_{min}} \tag{4}$$

where $D_{k,nor}$ is the final score of the *k*th focused library. D_{max} and D_{min} are the maximum and minimum D_k values of a specific generation, respectively. Thus, the highest score of library diversity is 1 and the lowest is 0.

Druglikeness. Usually the molecules for studying druglikeness are taken from molecular databases, such as the MACCS-II Drug Data Report (MDDR), Comprehensive Medicinal Chemistry (CMC), and the Available Chemicals Directory (ACD).[19,34,41,42] In this study, four molecular databases were used, namely, ACD, MDDR, CMC, and the Chinese natural product database(CNPD);[43] however, only those molecules that are composed of C, H, O, N, S, P, and halogen elements with molecular weights from 78 to 600 were chosen for calculating druglikeness. Furthermore, nontherapeutic compounds were removed from the CMC database, and those molecules that are included in MDDR and CMC were also deleted from ACD. Subsequently, there are 294 315, 108 114, 6680 and 41 782 compounds left in the modified ACD, MDDR, CMC, and CNPD databases, respectively. It is expected that the compounds in MDDR, CMC, and CNPD are, on average, more druglike than those in the ACD library.

The selection of descriptors is important in estimating druglikeness. The "rule of five" suggests that a compound is not druglike if it meets two of the four following conditions: the compound has more than five hydrogen bond (H-bond) donors, 10 H-bond acceptors, a molecular weight (MW) greater than 500, and calculated logP (CLogP) greater than 5 (or MlogP > 4.15).[28] When this rule was employed to study the druglikeness of the compounds in the above four libraries, the result showed that CMC is much closer to ACD than to MDDR. For instance, the mean values of MW are 309.9, 325.0, and 392.0 for ACD, CMC, and MDDR, respectively, suggesting that CMC is more ACD-like than MDDR-like. This is different from our expectation, indicating that MW, one descriptor of the rule, could not efficiently separate ACD from CMC. Therefore, there is room to improve the descriptor set for calculating druglikeness. To make a more robust descriptor set, 21 new structural descriptors, which are library- and molecular-size-independent, were designed for developing a new druglikeness descriptor set.[44] Three of them were introduced into LD1.0 for calculating druglikeness. Therefore, the new descriptor set encoded in LD1.0 is composed of the "rule of five" and the three new descriptors of structural ratio. Table 1 summarizes these descriptors, where XlogP is a program developed by Lai et al. for calculating LogP.[45]

Activity Assessment. Molecular activity could be defined by its binding affinity to the target protein. Nowadays, molecular docking is the most commonly used method to evaluate binding strength of a ligand to its target(s). One of the most popular molecular docking programs is DOCK4.0.[46−48] The best energy score from DOCK 4.0 was used in our software package to assess molecular bioactivity. The score could be very different in different systems, suggesting that the ordinary methods for normalization cannot be used in the case of the activity score. To solve this problem, a sigmoid function, eq 5, was introduced into LD1.0 for dealing with the normalization,

$$y = \frac{1 - e^{ax}}{1 + e^{ax}} \tag{5}$$

where *a* is a constant, *x* is the binding energy from the output file of DOCK 4.0, and *y* is the normalized activity score. Because the binding energy in the output file of DOCK4.0 is always no greater than 0.00 kcal/mol, *y* is always between 0 and 1. Therefore, the sigmoid function is a good choice for the normalization of the bioactivity score. Figure 2 demonstrates the plot of the sigmoid function against binding energy when *a* equals 0.05. If the value of *x* is between 0.00 and −80.00 kcal/mol, the normalized activity score, *y*, is between 0.0000 and 0.9640 (Figure 2).

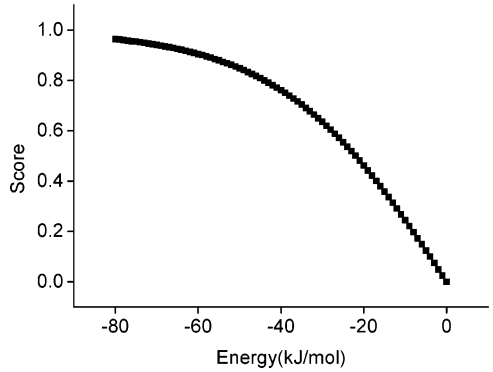


Figure 2. The plot of the sigmoid function *y* against the binding score from DOCK 4.0.

General Score Function for a Focused Library. The general score function for assessing a focused library is composed of three scores: the diversity score, the druglikeness score, and an activity score with weights of 0.1−0.3, 0.1−0.3, and ∼0.5 to 0.9, respectively. The weight of a specific score could be changed for different investigations. For example, if the compounds in a library are known to be very druglike, and what is of most interest in a study is the hit rate, even greater weight might be given to the activity score to improve the hit rate of the final library. The default weights for diversity, druglikeness, and activity scores in LD1.0 are 0.1, 0.2, and 0.7, respectively.

Building Blocks. A successful focus library also relies on the building block database. Our building blocks come from three sources.

1. From Known Drugs or Inhibitors. Extracting fragments from known drugs or ligands (inhibitors or activators) of the studied target is an effective approach for collecting new building blocks. A specific fragment database for a certain target should be constructed on the basis of available structural information of its ligands.

2. From Existing Drug Fragment Libraries. Some fragment libraries, such as the fragment library in the module Ludi of the software Insight II, are often used for de novo drug design.[49] The software LD1.0 has a default fragment database for those targets with little information about their ligands.

3. From Inhibitors of Similar Targets. Homology proteins usually share similar structural features and characteristics, especially at the binding site or active site. Therefore, the ligands for different targets belonging to the same family should share some common fragments in their inhibitors. Thus, the fragment database for a target could be constructed by referring to the structures of the ligands of its homology proteins.

Results and Discussion

Calculating Molecular Diversity. To verify the reasonability of the descriptor set for calculating molecular diversity, two libraries, nos. 1 and 2, were constructed from two sets of molecules randomly selected from the MDDR database with molecular weights less than 350. Library no. 1 has 23 molecules that are composed of only three elements, C, H, and O. Library no. 2, also randomly selected from MDDR, has 23 molecules as well, but without any restraint in their structures; therefore, library no. 1 should possess lower diversity than library no. 2. Then, $D_{k,nor}$ for libraries nos. 1 and 2 should be equal to 0.0 and 1.0, respectively. Meanwhile, two more libraries, nos. 3 and 4, were constructed by mixing some molecules from library no. 1 with some molecules from library no. 2. Table 2 summarized these four libraries. Regarding library no. 3, it has 12 structures from library no. 1 and 11 compounds from library no. 2. Similarly, library no. 4 has 18 structures from library no. 1. Thus, it is expected that library no. 2 should have a higher score than library no. 3. Indeed, the calculation result using the molecular diversity descriptor set (Table 2) shows that the library diversity correlates with the number of the molecules from library no. 2, demonstrating that the structural diversity descriptor set is reasonably good.

Table 2. Molecular Libraries for Testing the Modified Descriptor Set in Evaluating Molecular Diversity

	molecular library			
	1	2	3	4
molecule no. from library 1	23	0	12	18
molecule no. from library 2	0	23	11	5
$D_{k,nor}$	0.0000	1.0000	0.7726	0.3882

Calculating Molecular Druglikeness. 1. Setting Values for the New Descriptors.[44] Taking the descriptor h_p in Table 1 as an example, Figure 3 depicts its distribution in the four modified libraries, ACD, MDDR, CMC, and CNPD.

It shows that when the h_p has values from 0.80 to 1.40, the ACD library has the lowest value of percentages among the four databases, suggesting that a library is more druglike if its h_p value is within the rang from 0.80 to 1.40. Because ACD is not a pure nondruglike library, some of its compounds might be potential drug leads. The default value for h_p was set as 0.6−1.60. According to the definition of h_p in Table 1, it should be related to molecular saturation. Thus, a molecule with a chain structure is not a good drug lead if its carbon atoms are all in sp^3 hybridization. Values of the other two new descriptors were set in a way similar tto that of the descriptor h_p. Their default values are 0.15−0.80 and 0.10−0.45 for C3p and unsat_p, respectively.[44]

2. Validating Druglikeness Descriptor Set. To testify the modified descriptor set for druglikeness, two tests were carried out as shown in Figure 4.

Figure 4a shows the result based on 100 ordinary oral drugs.[50] It illustrates that 58% of them obtained score of 1.0, and 83% have scores >0.6, suggesting that the newly modified druglikeness descriptor set is reasonable. Figure 4b is the testing result for ACD, MDDR, CMC, and CNPD. It demonstrates that more compounds are less druglike in ACD than in MDDR, CNC, and CNPD, suggesting again that our druglikeness descriptor set has a strong ability to discriminate a druglike library from a nondruglike library.

Test Cases for LD1.0. To verify the reliability of the software package LD1.0, two tests were carried out on building COX2-focused and PPARγ-focused libraries.

1. Focused Library Design for COX-2 Inhibitors. Cyclooxygenase is a key enzyme associated with arachidonic acid (AA) metabolism. Inhibitors of COX, such as nonsteroidal antiinflammatory drugs (NSAIDs), have displayed

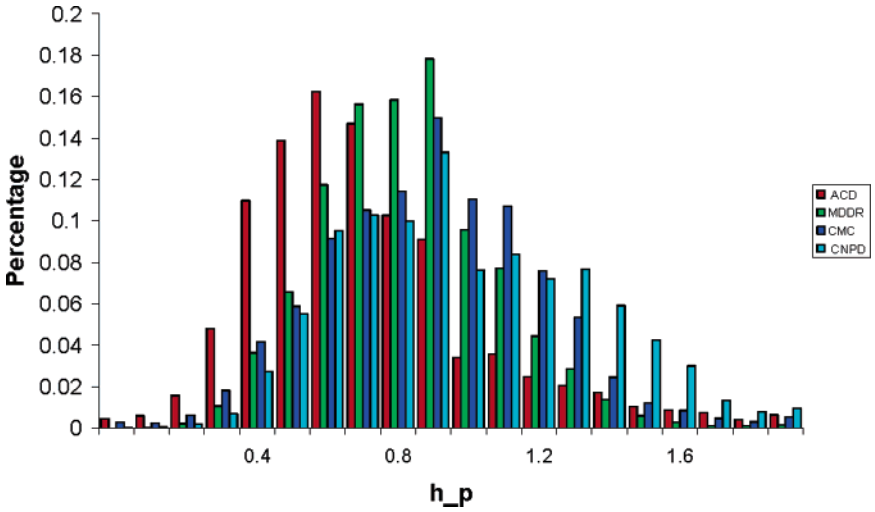
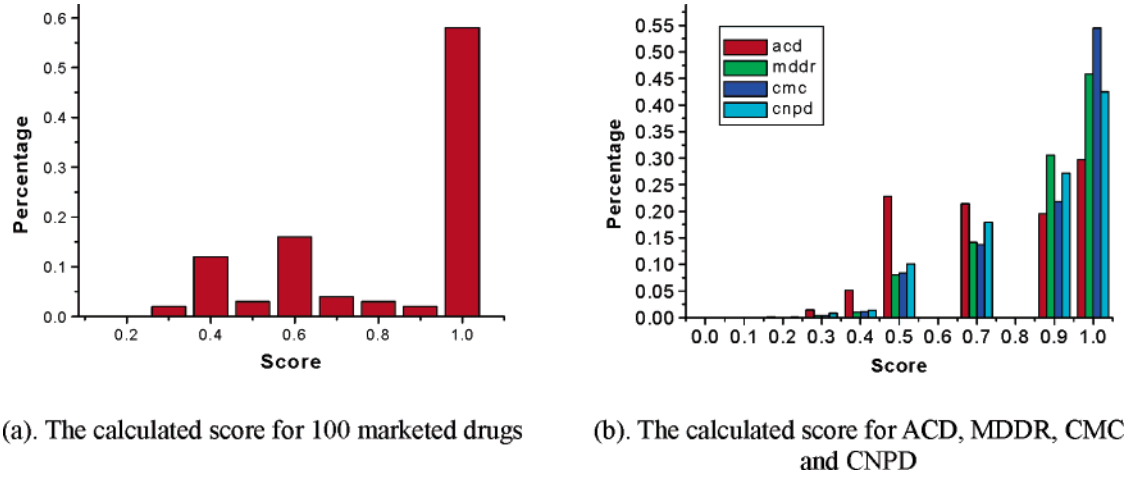


Figure 3. Figure 3. The distribution of the descriptor h_p for the databases of ACD, MDDR, CMC, and CNPD.



(a). The calculated score for 100 marketed drugs

(b). The calculated score for ACD, MDDR, CMC and CNPD

Figure 4. Testing result of the modified druglikeness descriptor set.

their antiinflammatory action.[51−53] However, treatment with NSAIDs, particularly in chronic cases, often leads to disruption of beneficial prostaglandin-regulated processes.[54,55] COX has two isoforms, namely, COX-1 and COX-2.[56−59] It is believed that the inhibition of COX-1 causes the side effects seen with NSAIDs. Therefore, the selective inhibitors of COX-2 would constitute a novel approach to the treatment of inflammation with fewer side effects.[60] This idea has led to the discovery of a family of COX2-selective inhibitors and drugs that are better tolerated than the older NSAIDs. SC58635, an FDA-approved drug for antiinflammation, is one of the drugs.[61] To test LD1.0 and the modified descriptor sets, we decided to build a COX2-focused library by using LD1.0 to see whether the optimized focused library contains the marketed and published COX-2 selective drugs and inhibitors.

1. Fragment Databases. SC58635 could be divided into three parts as shown in Figure 5. Here, the head of SC58635 is part A that is a *p*-aminosulfophenyl; the tail is part C that is *p*-methylphenyl; and the body is part B, 3-trifluoro-1H-pyrazol-1-yl, which acts as a linkage between parts A and C. Referring to the structures of the other published COX inhibitors, we constructed three fragment databases for parts A, B, and C with fragment numbers of 12, 16, and 4, respectively. Figure 6 lists all these fragments.

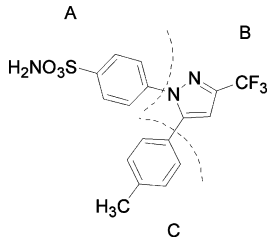


Figure 5. Structure of selective COX-2 drug SC58635 and its fragment division.

These fragments from different parts could react with each other to produce 768 compounds. Starting from these fragments and on the basis of the X-ray crystallographic structure of COX2 (PDB entry 6COX), the software LD1.0 was employed to build a COX2-focued library with a population of $3 \times 3 \times 3$. Default parameters in LD1.0 were used for constructing the testing library. The program terminated normally after running for 28 generations. It took one CPU on a SGI R12000 Origin3800 computer 5.42 h. The final COX2-focused library consists of the fragments A1, A2, A9, B1, B9, B10, C1, C3, and C4.

2. COX2-Focused Library. Experimental results have demonstrated that *p*-aminosulfophenyl and *p*-methylsulfophenyl are key groups for inhibition activity.[62] Indeed, the optimized fragments for part A are A1 (*p*-aminosulfophenyl),

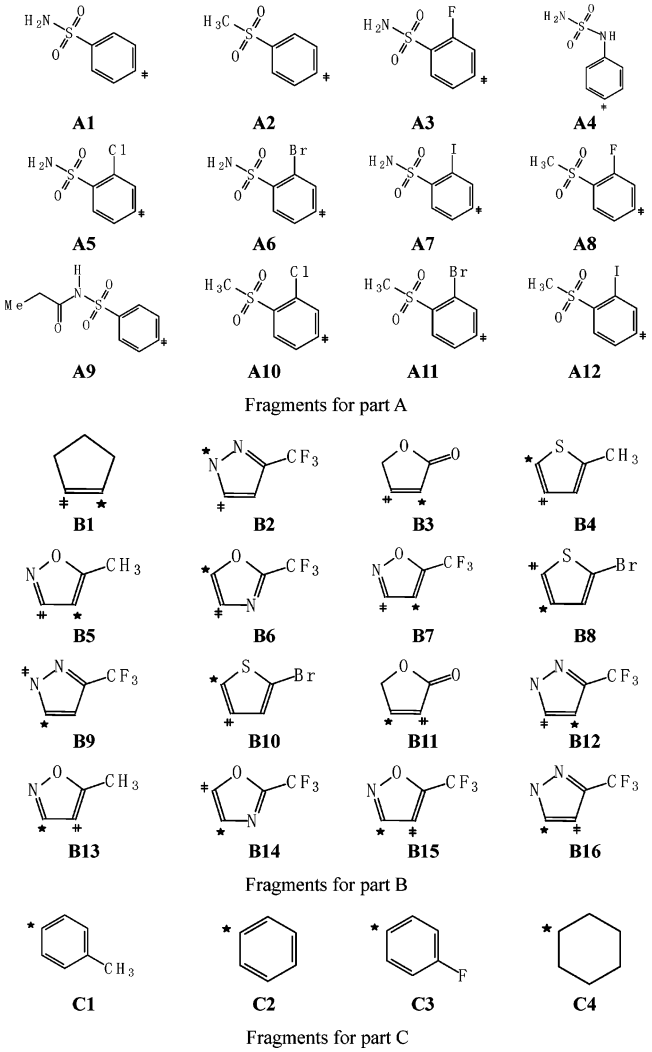


Figure 6. The initial fragments for parts A, B, and C for COX2 inhibitors. The symbols + and ★ represent the sites where fragments connect to each other to form a complete structure.

A2 (*p*-methylsulfophenyl), and A9. The fragment A9 is the head of Parecoxib sodium, a very powerful COX2-selective inhibitor.[63] Figure 7 depicts the binding model between COX2 and SC58635 derived by DOCK4.0.

Figure 7 clearly shows that the head part of the drug interacts with the hydrophilic site of the binding pocket with two hydrogen bonds through its amino and sulfo groups. Therefore, the head part should be the key group for inhibition activity. On the other hand, Figure 7 also suggests that part B interacts with the hydrophobic site of the binding site of COX2. Indeed, the optimized fragments for part B are B1, B9, and B10, which are more likely hydrophobic. B9 is the body part of the drug SC58635, and B10 is the body part of another COX2-selective inhibitor, DuP-697, developed by DuPont Corp.[62] Meanwhile, B1 can be found as the body part in some other COX2 inhibitors.[64] The optimized fragments for part C are also hydrophobic in nature, corresponding to the hydrophobic part of the binding site. Fragment C1 is the tail of SC58635, and C3 is the tail of DuP697. C4 is also a frequently appearing fragment in certain COX2 inhibitors.[65]

On the basis of the optimized fragments, a focused library containing 27 compounds has been constructed. We carried

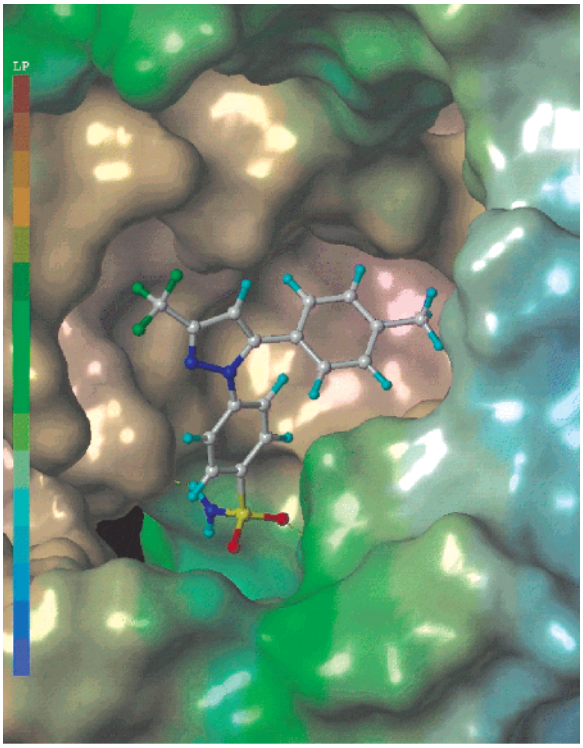


Figure 7. The binding model between COX2 and SC58635 derived from the docking calculation.

out a preliminary literature survey and found that six of them are the reported COX2-selective inhibitors.[64,65] Figure 8 depicts their structures and bioactivities. This result demonstrats that the software LD1.0 could successfully reproduce experimental results. Meanwhile, pure molecular docking was also performed using DOCK4.0 for comparison. The docking process for 768 structures, which were obtained through the combination of the initial fragments in Figure 6, requires more than 40.4 h. Hence, our program LD1.0 (5.42 h) saves more than 86.6% of the running time. Furthermore, no compounds among the top 27 structures identified by DOCK4.0 are found in the LD1.0-optimized COX2-focused library; moreover, pure docking did not place the 6 experimentally tested COX2 inhibitors shown in Figure 8 into the list of the top 27 compounds. Obviously, the combination of druglikeness and structural diversity descriptors with molecular docking is an effective approach for designing focused library with high hit rate and good ADME/T profile. Accordingly, LD1.0 should be a reliable and practical tool for building up a focused library against a specific target structure.

Focused Library Design for PPARγ Agonists. To further verify LD1.0 to see whether it can generate novel structures, another test on designing a focused library of PPARγ agonists was performed. The PPARs form a subfamily of the nuclear receptor superfamily. Three isoforms, encoded by separate genes, have been identified: PPARγ, PPARα, and PPARδ (also named PPARβ). PPARγ, the best-characterized subtype of PPARs, plays a crucial role in adipogenesis, glucose homeostasis, and insulin sensitization.[66,67] Thus, it is an important target for the treatment of type II diabetes, and its agonists are therefore expected to be novel drugs for the disease. Indeed, thiazolidinediones

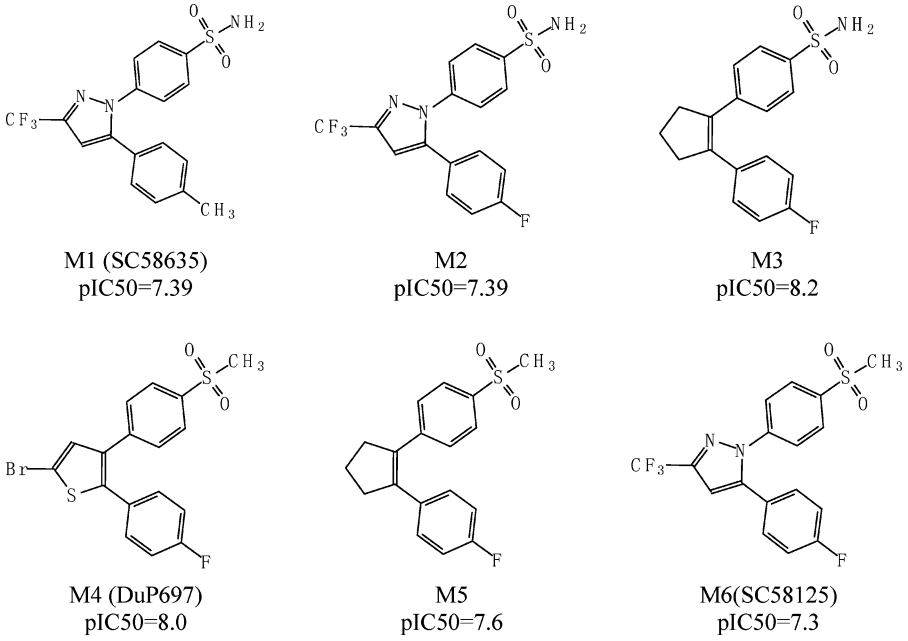


Figure 8. Successfully reproduced inhibitor structures in the COX2-focused library by LD1.0.

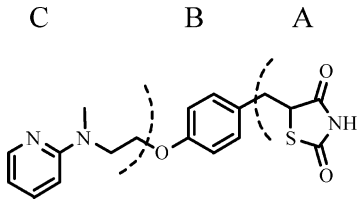


Figure 9. Common structure characteristic of PPARγ agonists and possibly structural fragment division.

(TZD), for example, Rosiglitazone and Pioglitazone (glitazones), approved by the Federal Drug Administration as drugs for type II diabetes, are high-affinity PPARγ agonists. However, the present PPARγ treatment of type 2 diabetes is still inadequate. For instance, a number of TZDs have been dropped from development due to their unacceptable side-effect profile.[68] It remains unclear whether the side effects are caused by the mechanism of action of these compounds or originate within the 2,4-thiazolidinedione chemical structure common to this class. Therefore, we determined to build a PPARγ-focused library that would contain both thiazolidinediones and nonthiazolidinediones. The library was built, on the basis of the X-ray crystallographic structure of PPARγ (PDB entry 2PRG), by using LD1.0.

1. Fragment Databases. It was noticed that most of the PPARγ agonists can be divided into three parts, as shown in Figure 9. Part A is a hydrophilic head, part C is a hydrophobic tail, and part B is a linker between parts A and C. After analyzing a large number of PPAR agonists and referring to other known drug fragment libraries, three fragment databases were constructed for parts A, B, and C with fragment numbers of 118, 88, and 98, respectively. These fragments could react with each other to give $\sim 10^6$ structures. Taking into account the capability of our laboratory for bioassay, we decided to construct a PPARγ-focused library with a population of $10 \times 10 \times 10$ by 3×10 building blocks.

2. Optimizing Building Blocks. The calculation conditions and parameters were set as the same as those for COX2-library construction. After running 434 generations, the program LD1.0 terminated normally, and the optimized building blocks for parts A, B, and C were reported. Some are depicted in Figure 10. This process cost one CPU of the same computer $\sim$84 days.

For part A, all the building blocks in Figure 10a are hydrophilic groups that could form hydrogen bonds with PPARγ. All the building blocks for part C in Figure 10c contain a hydrophobic aromatic ring, but their structures have considerable diversity. According to the X-ray structure, part C is located at the entrance of the PPARγ binding pocket. Because the structure of the entrance is rather flexible, it is understandable that part C has great structural diversity. The PPARγ crystal structure shows that the site where the part B occupied is a flat channel. Hence, a rigid planar fragment structure is expected. Indeed, the optimized building blocks for part B have a common structural feature: a planar five- or six-membered ring structure (Figure 10b).

3. Library Validation. We noticed that parts A and B of TZDs were found among the optimized building blocks for parts A and B. But, the tail of TZDs was not found in the optimized building blocks for part C. This is in agreement with the high structural flexibility of the entrance mouth of the PPARγ binding site. On the other hand, it also hints that there might be some compounds in the library that are more active than the usual TZDs, if our algorithm, including those modified descriptor sets, is reliable. To verify our assumption, a bioassay experiment was then carried out on some of the compounds. Indeed, a few of them were discovered to be more powerful as agonists than those FDA-approved TZD drugs (unpublished data). On the other hand, the fragments used in above focused library construction can generate a general library containing 1 017 632 structures. This library was screened using the pure docking method (DOCK4.0) targeting the binding pocket of PPARγ. The top 1000 structures from DOCK4.0 only reproduced 22 molecules of the 1000 structures in the optimized PPARγ-focused library

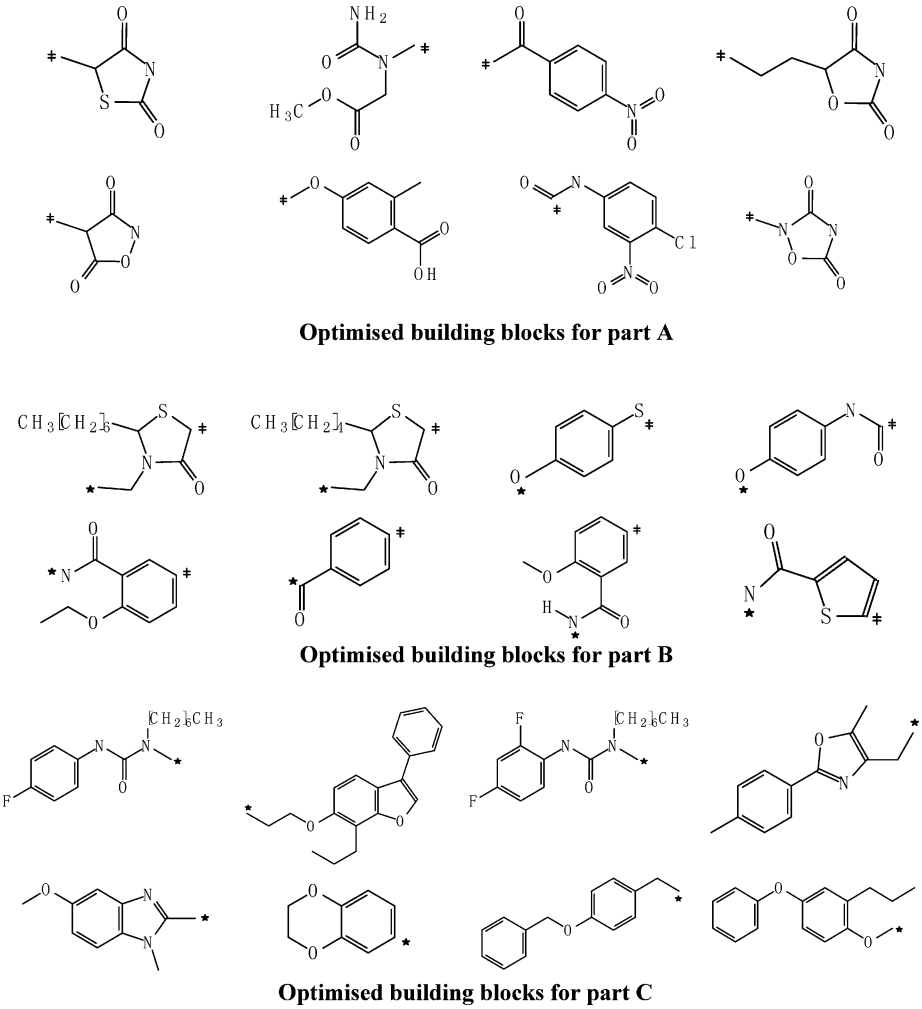


Figure 10. Some of the optimized building blocks for PPARγ focused library construction. The symbols + and ★ stand for the sites where fragments connect to each other to form a complete structure.

generated by LD1.0. Remarkably, LD1.0 saved ~88% of the running time in comparison with DOCK4.0. This testing example demonstrated again that for a combinatorial library design, the software LD1.0 as well as newly modified descriptor sets for druglikeness can optimize the building blocks, reduce the library size, and increase the hit rate. In addition, molecules designed with this approach may have good ADME/T properties.

Conclusions

A new approach for building a target-focused library was developed on the basis of the combination of the descriptor sets for molecular diversity and druglikeness with molecular docking. The library-based GA method was implemented into this method for library optimization, and DOCK4.0 was employed for scoring bioactivity. The application of this approach on building COX2- and PPARγ-focused libraries not only reproduced the important fragments in structures of the FDA approved drugs and in other published ligands, but also generated novel structures with more powerful potential than those available drugs, demonstrating that our new approach is efficient and reliable and would be helpful in discovering new drug leads with good ADME/T properties against various diseases rapidly and cost-effectively.

Acknowledgment. This work was supported by grants from the State Key Program of Basic Research of China (2003CB114401, 2002CB512802, 2004CB518901), the 863 Hi-Tech Program (2001AA235041, 2002AA233011), the National Natural Science Foundation of China (30070891), and Shanghai Key Program of Science and Technology 03DZ19228.

References and Notes

(1) Gallop, M. A.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. *J. Med. Chem.* **1994**, *37*, 1233−1251.
(2) Gordon, E. M.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gallop, M. A. *J. Med. Chem.* **1994**, *37*, 1385−1401.
(3) Terrett, N. K.; Gardner, M.; Gordon, D. W.; Kobylecki, R. J.; Steele, J. *Tetrahedron* **1995**, *51*, 8135−8173.
(4) Salemme, F. R.; Spurlino, J.; Bone, R. *Structure* **1997**, *5*, 319−24.
(5) Stanton, R. V.; Mount, J.; Miller, J. L. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 701−705.
(6) Drewry, D. H.; Young S. S. *Chemom. Intel. Lab. Syst.* **1999**, *48*, 1−20.
(7) Gordon, E. M.; Kerwin, J. F. In *Combinatorial Chemistry and Molecular Diversity in Drug Discovery*; John Wiley & Sons: New York, 1998; pp 17−36.
(8) Jorgensen, A. M.; Pedersen, J. T. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 338−345.
(9) Willett, P.; Winterman, V. *Quant. Struct.-Act. Relat.* **1986**, *5*, 18−25.

(10) Alexander, G. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 414−425.
(11) Brown, R. D.; Martin, Y. C. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.
(12) Matter, H.; Potter, T. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 1211−1225.
(13) Flower, D. R. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 379−386.
(14) Flower, D. R. *J. Mol. Graphics Modell*. **1998**, *16*, 239−253.
(15) Ashton, M. J.; Jaye, M. C.; Mason, J. S. *Drug Discovery Today* **1996**, *1*, 71−78.
(16) Warr, W. A. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 132−140.
(17) Pearlman, R. S.; Smith, K. M. *Perspect. Drug Discovery Des*. **1998**, *9*−*11*, 339−353.
(18) Gillet, V. J.; Willett, P.; Bradshaw, J.; Green, D. V. S. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 169−177.
(19) Clark, D. E.; Pickett, S. D. *Drug Discovery Today* **2000**, *5*, 49−58.
(20) Pickett, S. D.; McLay, I. M.; Clark, D. E. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 263−272.
(21) Darvas, F.; Dorman, G. *Chim. Oggi.* **1999**, *17*, 10−13.
(22) Oprea, T. I.; Zamora, I.; Svensson, P. Qvo vadis, scoring functions? Toward an integrated pharmacokinetic and binding Affinity Prediction Framework. In *Combinatorial Library Design and Evaluation for Drug Design*. Ghose, A. K.; Viswanadhan, V. N. Eds; Marcel Dekker Inc., New York, **2001**, pp 233−266.
(23) Oprea, T. I.; Zamora, I.; Ungell, A. L. *J. Comb. Chem.* **2002**, *4*, 258−266.
(24) Clark, D. E. *Adv. Drug Delivery Rev.* **2002**, *54*, 253−254.
(25) Zheng, W. F.; Cho, S. J.; Tropsha, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 251−258.
(26) Gillet, V. J.; Willet, P.; Bradshaw, J.; Green, D. V. S. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 169−177.
(27) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511−522.
(28) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Delivery. Rev.* **1997**, *23*, 3−25.
(29) Galvez, J.; Garcia-Domenech, R.; Julian-Ortiz, J. V. *J. Mol. Graphics Modell*. **2001**, *20*, 84−94.
(30) Ajay, A.; Walters, W. P.; Murcko, M. A. *J. Med. Chem*. **1998**, *41*, 3314−3324.
(31) Sadowski, J.; Kubinyi, H. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(32) Ghose, A.; Crippen, G. *J. Chem. Inf. Comput. Sci*. **1987**, *27*, 21−35.
(33) Xu, J.; Stevenson, J. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 1177−1187.
(34) Gillet, V. J.; Khatib, W.; Willett, P.; Fleming, P. J.; Green, D. S. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 375−385.
(35) Weber, L.; Wallbaum, S.; Broger, C.; Gubernator, K. *Angew. Chem. Int. Ed. Engl.* **1995**, *34*, 2280−2282.
(36) Weber, L. *Drug Discovery Today* **1998**, *3*, 379−385.
(37) Sheridan, R. P.; San Feliciano, S. G.; Kearsley, S. K. *J. Mol. Graphics Modell.* **2000**, *18*, 320−334.
(38) Sheridan, R. P.; Kearsley, S. K. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 310−320.
(39) Holland, J. H. *Adaptation in Natural and Artificial Systems*, The University of Michigan Press, Ann Arbor, 1975.
(40) Brown, R. D. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 31−49.
(41) Oprea, T. I. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251−264.
(42) Gillet, M. J.; Willett, P.; Bradshaw, J. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−176.
(43) Shen, J. H.; Xu, X. Y.; Cheng, F.; Liu, H.; Luo, X. M.; Chen, K. X.; Zhao, W. M.; Shen, X.; Jiang, H. L. *Curr. Med. Chem.* **2003**, *10*, 2327−2342.
(44) Zheng S. X.; Luo, X. M.; Chen, G.; Zhu, W.; Shen, J. H.; Chen, K. X.; Jiang, H. L. Submitted for publication.
(45) Wang, R. X.; Fu, Y.; Lai, L. H. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 615−622.
(46) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. *J. Mol. Biol*. **1982**, *161*, 269−288.
(47) Oshiro, C. M.; Kuntz, I. D. *J. Comput-Aided Mol. Des.* **1995**, *9*, 113−130.
(48) Ewing, T. J. A.; Kuntz, I. D. *J. Comput. Chem*. **1997**, *18*, 1175−1189.
(49) InsightII 2000.1; Accelrys Inc.: San Diego, CA; 2002.
(50) Raevsky, O. A.; Schaper, K. J.; Artursson, P.; MeFarland, J. W. *Quant. Struct.-Act. Relat.* **2002**, *20*, 402−413.
(51) Dannhardt, G.; Kiefer, W. *Eur. J. Med. Chem*. **2001**, *36*, 19−126.
(52) Carter, J. S. *Exp. Opin. Ther. Pat*. **2000**, *10*, 1011−1020.
(53) Talley, J. J. *Prog. Med. Chem.* **1999**, *36*, 201−234.
(54) Clive, D. M.; Stoff, J. S. *N. Engl. J. Med*. **1984**, *310*, 563−572.
(55) Pirson, Y.; Van Ypersele de Strihou, C. *Am. J. Kidney Dis*. **1986**, *8*, 337−344.
(56) Marnett, L. *Curr. Opin. Chem. Biol*. **2000**, *4*, 545−552.
(57) Garavito, R. M.; Dewitt, D. L. *Biochim. Biophys. Acta* **1999**, *1441*, 278−287.
(58) O'Banion, M. K. *Crit. Rev. Neurobiol.* **1999**, *13*, 45−82.
(59) Marnett, L. J.; Kalgutkar, A. S. *Tips* **1999**, *20*, 465−469.
(60) Masferrer, J. L.; Zweifel, B. S.; Manning, P. T.; Hauser, S. D.; Leahy, K. M.; Smith, W. G.; Isakson, P. C.; Seibert, K. *Proc. Natl. Acad. Sci*. **1994**, *91*, 3228−3232.
(61) Graul, A.; Martel, A. M.; Castaner J. *Drugs Future* **1997**, *22*, 711−714.
(62) Kurumbail, R. G.; Stevens, A. M.; Gierse, J. K.; Mc-Donald, J. J.; Stegeman, R. A.; Pak, J. Y.; Gildehaus, D.; Miyashiro, J. M.; Penning, T. D.; Seibert, K.; Isakson, P. C.; Stallings, W. C. *Nature* **1996**, *384*, 644−648.
(63) Talley, J. J.; Bertenshaw, S. R.; Brown, D. L.; Carter, J. S.; Graneto, M. J.; Kellogg, M. S.; Koboldt, C. M.; Yuan, J.; Zhang, Y. Y.; Seibert, K. *J. Med. Chem*. **2000**, *43*, 1661−1663.
(64) Marot, C.; Chavatte, P.; Lesieur, D. *QSAR* **2000**, *19*, 127−134.
(65) Huff, R.; Collina, P.; Kramer, S. *Inflamm Res*. **1995**, *44* (Suppl. 2), 145−146.
(66) Michalik, L.; Wahli, W. *Curr. Opin. Biotechnol.* **1999**, *10*, 564−570.
(67) Kliewer, S. A.; Umesono, K.; Noonan, D. J.; Heyman, R. A.; Evans, R. M. *Nature* **1992**, *358*, 771−774.
(68) Ekins, S.; Schuetz, E. *Trends Pharmacol. Sci.* **2002**, *23*, 49−50.

CC049866H