Drug and Drug Candidate Building Block Analysis

Junmei Wang*,† and Tingjun Hou‡

Department of Pharmacology, University of Texas Southwestern Medical Center, 6001 Forest Park Road, ND9.136, Dallas, Texas 75390 and Functional Nano & Soft Materials Laboratory (FUNSOM), Soochow University, Suzhou 215123, P. R. China

Received October 13, 2009

Drug likeness analysis is widely used in modern drug design. However, most drug likeness filters, represented by Lipinski's "Rule of 5", are based on drugs' simple structural features and some physiochemical properties. In this study, we conducted thorough structural analyses for two drug datasets. The first dataset, ADDS, is composed of 1240 FDA-approved drugs, and the second drug dataset, EDDS, is a nonredundant collection of FDA-approved drugs and experimental drugs in different phases of clinical trials from several drug databases (6932 entries). For each molecule, all possible fragments were enumerated using a brutal force approach. Three kinds of building blocks, namely, the drug scaffold, ring system, and the small fragment, were identified and ranked according to the frequencies of their occurrence in drug molecules. The major finding is that most top fragments are essentially common for both drug datasets; the top 50 fragments cover 52.6% and 48.6% drugs for ADDS and EDDS, respectively. The identified building blocks were further ranked according to their relative hit rates in the drug datasets and in a screening dataset, which is a nonredundant collection of screening compounds from many resources. In comparison with the previous reports in the field, we have identified many more high-quality building blocks. The results obtained in this study could provide useful hints to medicinal chemists in designing drug-like compounds as well as prioritizing screening libraries to filter out those molecules lack of functional building blocks.

INTRODUCTION

It is assumed that the chemical space exceeds 10^{60} molecules, and it is impossible for mankind to make all of those molecules. So far, only about 27 million compounds have been registered.[1] Recently, Blum and Reymond reported a virtual screening database, GDB-13, which has 970 million drug-like small molecules.[2] The authors claimed that GDB-13 is the largest publicly available database of virtual molecules ever reported. On the other hand, the chemical compounds used by biological systems represent an amazingly small fraction of the entire chemical space: less than 10 000 small molecules have been approved to treat all kinds of diseases.[1]

With the advances of combinatorial chemistry and high-throughput screenings (HTS), obtaining one or several inhibitors (IC_{50} better than 1 μM) is not very difficult for some protein targets in drug discovery. However, the challenge of choosing proper drug leads to optimize and turning promising drug candidates into real drugs is still enormous. As a matter of fact, the number of new molecular entities (NMEs) approved by the Federal Food and Drug Administration (FDA) each year has ranged from 9 to 58 from 1980 to 2008. Moreover, there is no trend indicating that the number of approved NMEs has been soaring in recent years.[3]

Why is it so challenging to turn a NME into a real drug? First of all, a developable drug lead of the right target must be discovered. A developable drug lead implies it is expandable and optimizable. It is not easy to cherry-pick a developable drug lead from many HTS hits since the activities of HTS hits are typically low and there are many false positives in HTS assays.

After a developable drug lead has been discovered, it is further optimized to improve its activity of inhibition, selectivity against other targets, and its ADME-Tox properties. To conduct lead optimization fruitfully, it is challenging for medicinal chemists to make an optimal choice of the most likely compounds to succeed from the infinite number of possible variations of the drug lead. Once one or several promising drug candidates are obtained, there remains the challenge of making the right clinical decisions. Only a small fraction of experimental drugs have the chance to be approved by the FDA. According to Kola and Landis,[4] the poor drug efficacy, toxicology, safety, and physicochemical characteristics are responsible for most of the drug attritions.

To identify developable drug leads and increase the successful rate of potential drugs to be approved, drug-like analysis has been applied in drug discovery.[5−7] The drug-likeness of a molecule can be characterized by some drug-like properties derived from databases of known drugs. The most famous drug-likeness filter is the "Rule of 5" suggested by Lipinski,[8] which states that a good drug candidate should have a molecular weight equal to or smaller than 500, a calculated log *P* (ClogP) equal to or smaller than 5.0 (or 4.15 for Moriguchi logP),[9] and numbers of hydrogen-bond donors and acceptors equal to or less than 5 and 10, respectively. Those four rules were derived from an analysis of 2245 drugs from the World Drug Index (WDI) believed

* To whom correspondence should be addressed. E-mail: junmei.wang@utsouthwestern.edu.
† University of Texas Southwestern Medical Center.
‡ Soochow University.

10.1021/ci900398f
Published on Web 12/18/2009

to have entered phase II clinical trials. Another popular molecular property for characterizing drug likeness is polar surface area. Kelder et al. found that 90% of orally bioavailable non-CNS drugs have a polar surface area (PSA) below 120 Å^2, and this number dropped to 80 Å^2 for CNS drugs. Wenlock et al. compared the physiochemical property profiles of developed and marketed drugs to identify the trends in physiochemical properties that favor a drug's successful passage through clinical development and on to the market.[10] Vieth and Sutherland made a comparison of two molecular properties, molecular weight and clogP, for different families of FDA-approved drugs, and they suggested that the modified rules of drug likeness should be adopted for certain target classes.[11]

A number of ADME-Tox properties have also been applied to describe the drug likeness of a compound.[7,12] We found that real drugs are about 20-fold more soluble than the so-called drug-like molecules in the ZINC database, which have no violation of Lipinski's "Rule of 5" at all. Specifically, oral drugs are about 16-fold more soluble, while injection drugs are 50−60-fold more soluble.[13] We believe that ADME-Tox models that predict molecular physiochemical, physiological, and pharmacokinetic properties can be utilized as filters to prioritize compound libraries prior to high-throughput screenings or as descriptors to construct classification models to distinguish drugs from nondrugs.[14]

Besides molecular properties discussed above, the drug likeness of a compound can also be characterized by its building blocks. Drug building-block analysis has been conducted in a set of notable studies.[15−21] The following is a simple review of their findings. The early work on drug building-block analysis goes back to the past century. Bemis and Murcko studied 5120 drugs selected from the Comprehensive Medicinal Chemistry (CMC) database (v. 94.1) and identified the molecular frameworks that frequently occurred in the drugs.[15] The molecular framework here is constituted by ring systems (cycles or cycles sharing an edge in graphic representation of molecules) and linkers (atom paths connecting two ring systems). A total of 1179 frameworks were identified among the 5120 compounds, and the top 42 frameworks accounted for 24% of the drugs. It should be noted that the "drugs" mentioned in this paper are not necessarily the FDA-approved drugs. As a matter of fact, there are less than 1400 new molecular entities approved by the FDA so far. In contrast, it is estimated that more than 3000 developing drugs are in different phases of clinical trials.[22,23] In this work, all of the biologically active compounds, pharmaceutical agents, and experimental drugs entering clinical trials are generally referred to as drugs. Another type of building block, the side chains which consist of nonring and nonlinker atoms, were also analyzed by Bemis and Murcko in another work.[16] One major finding is that the CMC database contained about 15 000 side chains, of which 11 000 belonged to one of the top 20 types. The average number of side chains per molecule is four, and the average number of heavy atoms per side chain is two.

Using a similar strategy to that of Bemis and Murcko's to dissect a molecule, Wang and co-workers analyzed 58 682 chemicals from RTECS (the Registry of Toxic Effects of Chemical Substances) to identify structure patterns, which are pertinent to some kind of activity, to account for the specific toxicity.[17]

In 2001, Lee and Schneider made side by side comparisons of two datasets, one 5757-drug dataset taken from the Derwent World Drug Index[24] and 10 495 natural products compiled from the BioscreenNP database (http://www.ibscreen.com/natural.html).[18] Two basic molecular properties, molecular weight and clogP, are essentially the same for the two datasets. However, a drug molecule contains approximately twice the number of nitrogens than a natural product. The ring systems were identified in an exhaustive manner for molecules in both datasets. There are 1748 and 807 different ring systems identified for the BioscreenNP and WDI databases, respectively. Approximately 35% of the ring systems in the drugs are also present in the natural products. On the other hand, only 17% of the ring systems occurring in natural products have an identical counterpart in the drugs. In 2007, Grabowski and Schneider conducted a similar study to identify the top scaffolds for the drug database of COBRA (Collection of Bioactive Reference Analogues) and a set of natural product datasets using a similar approach to Bemis and Murcko's.[21]

In 2007, Siegel and Vieth analyzed the chemical structures of 1386 marketed drugs. They found that 15% of drugs are contained within other drugs differing by one or more continuous chemical fragments and about 30% of drugs contain other drugs as building blocks.[19] Their findings justify the effort of performing a drug building-block analysis to help discover new drugs. In another recent paper, Sutherland et al. utilized a recursive algorithm to generate possible fragments within a molecule. Then, fragment fingerprints were constructed using those top occurring fragments (at least one occurrence per 1000 molecules) in the datasets. Naive Bayes models were then generated to calculate the probability of observing activity—either active or inactive, given the presence of a fragment. Their results demonstrated that protein similarity can be assessed by comparing the fragment fingerprints of ligands of the two receptors.

Besides the physical properties and molecular building blocks for describing drug likeness, there was a set of classification models published in recent years that distinguishes drugs from nondrugs. Most of those models were constructed by neural networks, genetic algorithms, and recursive partitioning using less physical descriptors such as atom types. Those models were nicely reviewed by Muegge.[6]

It is pointed out that the three types of methods for drug-likeness analysis are complementary to each other, and the combination of them can much better describe the drug likeness of a molecule. Recently, we have successfully applied a combination of molecular fragments and a few molecular properties to model human oral bioavailability, plasma protein binding, and urinary excretion.[12] On the other hand, no valid model can be developed for those ADME properties purely using molecular properties as descriptors.[25,26]

In this work, we set out to perform a drug building-block analysis to identify different kinds of building blocks using a recursive algorithm for two drug datasets, an approved drug dataset (ADDS) and an extended drug dataset (EDDS). The first drug dataset is composed of 1240 marketed drugs. The second drug dataset includes all of the entries in the first dataset and a set of developing drugs collected from several drug databases. The reasons that motivated us to conduct a

drug building-block analysis are listed as follows. First of all, most reports in the field are almost 10 years old, and the databases (CMC, WDI, etc.) being used in the analysis are obsolete. It is estimated that 250 new molecules are added to CMC per year, and WDI has also been greatly expanded recently (http://www.daylight.com/products/wdi.html). Siegel and Vieth's work focused on 1386 marketed drugs, and only real drugs which are wholly incorporated in other drugs were identified. This extremely strict searching criterion screened out many molecular fragments that occur in the drug dataset.

Second, most previous work on drug building-block analysis only studied one type of building block, either the ring system or the framework or the side chain. We believe another type of building block, which is called "structure pattern" in Wang et al.'s structural analysis on a toxic substance database, is essential in drug building block analysis. Structure pattern is defined as the combination of frameworks and side chains. In this work, we proposed a different method to classify building blocks. Three types of building blocks have been defined: drug scaffold (resembling structure pattern), ring system (resembling framework), and small fragments (resembling linkers and side chains). We found that our nonexclusive classification is suitable to exhaustively identify high-quality building blocks.

Third, few previous studies further label the identified building blocks as drug-like or not drug-like. The drug-like building blocks, which may be called functional building blocks, are typically the relevant groups of a molecule that interact with a receptor and are responsible for its activity, in other words the pharmacophore. On the other hand, the nondrug-like fragments, which may be called the composition building blocks, are typically the auxophore of a molecule. It is difficult to, without bias, measure the drug-likeness of building blocks and to classify them into the functional and composition categories. In this work, we attempt to propose a classification scheme to group a building block into different categories (I, II, III, IV, and V) based on the ratio of its occurrence frequency in the drug dataset to that in the screening dataset. The larger the ratio is, the higher the category. For a building block, if the frequency of occurrence in the drug dataset is smaller than that in the screening dataset, it belongs to either category I or category II. Otherwise, it belongs to category III, IV, or V. One may find that the classification scheme is useful if he or she believes that the hypothesis "the larger the ratio is, the more drug-like the building block is" is valid to some degree.

Finally, to the best of our knowledge, no building block analyses have been performed for both the approved and the extended drug datasets using the same algorithm. It is interesting to compare the analysis results for the two kinds of drug datasets.

2. METHODOLOGY

2.1. Data Sets. Two drug datasets and one screening dataset were prepared for this study. The first drug dataset, ADDS, is composed of FDA-approved drugs in all kinds of dosage forms. While the second drug dataset, EDDS, is a collection of FDA-approved drugs and experimental drugs at different phases of clinical trials from several resources, which include DrugBank (http://www.drugbank.ca), World Drug Index (WDI), Prous (http://www.prous.com), and so forth. The third dataset, SDS (the screening dataset), is a subset of the "clean-drug-like" molecules in the ZINC database (http://zinc.docking.org).[27]

A standard procedure was employed to clean up the three datasets. Too small (molecular weight smaller than 50) or too large (molecular weight larger than 1000) molecules were excluded. Duplicated entries and those containing elements other than C, H, O, N, S, P, and halogens were eliminated. To avoid some fragments being over-represented by the extremely similar molecules, redundancy was reduced for all three datasets so that the Tanimoto similarity between any two molecules was less than 0.99 for ADDS and EDDS and 0.95 for SDS. The Optisim algorithm was applied to conduct diverse selection to reduce the redundancy.[28] For EDDS, the approved drugs have highest priority to enter the dataset, and the experimental drugs in the later phases of clinical trials have higher priority to be selected. Finally there are 1240, 6932, and 1.95 million entries in the approved drug, the extended drug, and the screening datasets, respectively.

The average molecular weights are 345.58 $\pm$ 144.20, 321.75 $\pm$ 127.87, and 396.74 $\pm$ 63.69 for ADDS, EDDS, and SDS, respectively. As to ClogP, the averages are 1.87 $\pm$ 2.94, 1.98 $\pm$ 2.93, and 3.182 $\pm$ 1.13 for the three mentioned datasets, correspondingly. It should be pointed out that the extended drug set contains all of the entries in the approved drug set, and there is no overlap between the two drug datasets and the screening dataset.

2.2. Fragmentation. In the second step, a brute force approach was utilized to recursively cut every molecule in the drug dataset into fragments. A valid cutting happened at every cuttable bond. A fragment, if it has at least one cuttable bond, is then taken as a parent molecule for further dissection. All of the fragments generated by this procedure were collected and sorted, and duplicated entries were eliminated. An example of this cutting mechanism is shown in Figure 1. In this study, we defined a cuttable bond to be a single, nonring, nontail ending bond. Specifically, the two atoms of a cuttable bond cannot occur in the same ring, no matter if it is an aromatic or aliphatic ring; both atoms of the cuttable bond must be nontail ending atoms. A tail ending atom is defined as an atom having only one bonded atom, such as hydrogen, halogens, and the ending oxygen in a delocalized bond. In this work, however, an atom is also considered as a tail ending atom when it has multiple bonded atoms but only one bonded atom has a bonding connectivity larger than 2. For example, in Figure 1, bonds *a*, *b*, and *c* are all cuttable bonds. On the other hand, bond *d*, the single bond between methyl and cyclopropane of the fragment methylcyclopropane, should not be considered a cuttable bond since the methyl carbon is a tail ending atom according to our definition. In Figure 1, the exemplary drug molecule is in layer 0, and it has three offsprings in layer 1 after cutting one cuttable bond each time. The fragments in layer 1 are further cut to generate fragments in layer 2. This procedure is iteratively run until no fragments have valid cuttable bonds.

2.3. Fragment Classification. In the third step, all of the fragments of the drug molecules were merged and sorted. For each fragment, the frequency of its occurrence in the drug molecules was calculated. Those fragments that showed up in less than five drug molecules were eliminated. The top fragments that extensively appeared in drug molecules were then identified. It is notable that up to 1 million

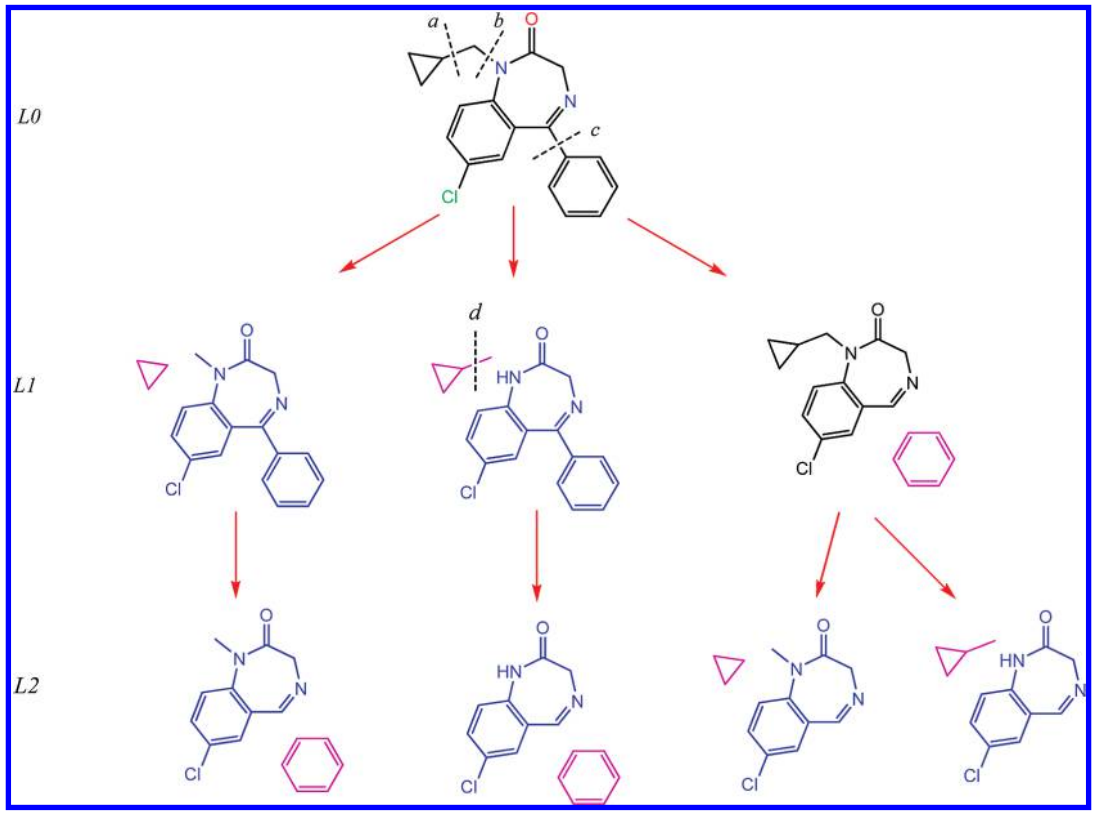

Figure 1. How to cut a molecule into fragments using a brute force algorithm? The parent molecule in L0 (Layer 0) is cut into fragments from its cuttable bonds (bonds *a*, *b*, and *c*), and its three direct offspring are stored in L1 (Layer 1). All offspring fragments in L1 are further cut into fragments from their cuttable bonds, and their offspring fragments are stored in L2 (Layer 2). This procedure is iteratively performed until no fragment has any cuttable bond. All fragments in magenta are valid small fragments (SF), and those in blue are valid ring systems (RS). According to our definition, bonds *a*, *b*, and *c* are cuttable bonds, but bond *d* is not.

fragments can be generated for a couple thousand molecules using the brute force algorithm. Therefore, a frequency of occurrence cutoff is needed to filter out those unimportant fragments.

Different researchers have different strategies to dissect a molecule into building blocks. Bemis and Murcko exclusively dissected a molecule into three types: ring systems, linkers, and side chains. Another type of building block, framework, is defined as the union of ring systems and linkers. All of their structural analysis of the CMC drugs was based upon counting frameworks.[15] Wang et al. adopted the same approach to dissect their molecules into units. Another building block, which was called structure pattern, was defined as a structure unit composed of a framework and some given functional groups.

To be consistent with our recursive dissection algorithm, we adopted a nonexclusive grouping strategy to classify the molecular fragments. Three types of building blocks were defined, namely, drug scaffold (DS), ring system (RS), and small fragment (SF). The definition of drug scaffold is not very rigorous in this work. We loosely defined a drug scaffold to be a molecular fragment having at least nine heavy atoms. Generally, the more heavy atoms a scaffold has, the less frequent it shows up in the drug dataset, though the more appropriate it is for it to serve as a core structure in combinatorial chemistry or a scaffold in drug lead optimization. For a RS building block, the number of heavy atoms is equal to or larger than eight, and it has at least one ring and no more than two rotatable bonds in any branches. In view of most drugs and screening molecules having at least one five- or six-membered ring, a five- to seven-membered single ring is not considered as a ring group according to our definition. As to SF building blocks, the heavy atoms ranged from 3 to 12, and the occurrences must be equal to or larger than 10. Considering that a large percentage of drug molecules (66.2%) and screening molecules (90.7%) have at least one benzene ring, the phenyl group is also eliminated from the SF set. It is notable that Siegel and Vieth's building blocks, which are actually real drugs, are similar to our DS type except that we required a DS building block to have at least nine heavy atoms. Certainly, different criteria may be applied to define building blocks and to collect fragments to perform analysis, and we found that the above criteria are suitable to retrieve appropriate numbers of building blocks for DS and RS as well as SF. Examples of different types of building blocks are shown in Figure 2. It should be pointed out that many building blocks identified in this study and in the previous work have open valences. In other words, those building blocks are stable only if they are linked to other building blocks or atoms.

2.4. Classification on Drug Building Blocks. In the last step, Smarts strings were generated for the top fragments to query the screening dataset using the OEChem programming library of OpenEye Scientific Software.[29] The frequencies of occurrence of the identified fragment in the screening dataset were then calculated. The building blocks were grouped into one of five categories based on the ratio of the frequency of occurrence in the drug dataset to that in the screening dataset. The following is the classification scheme:

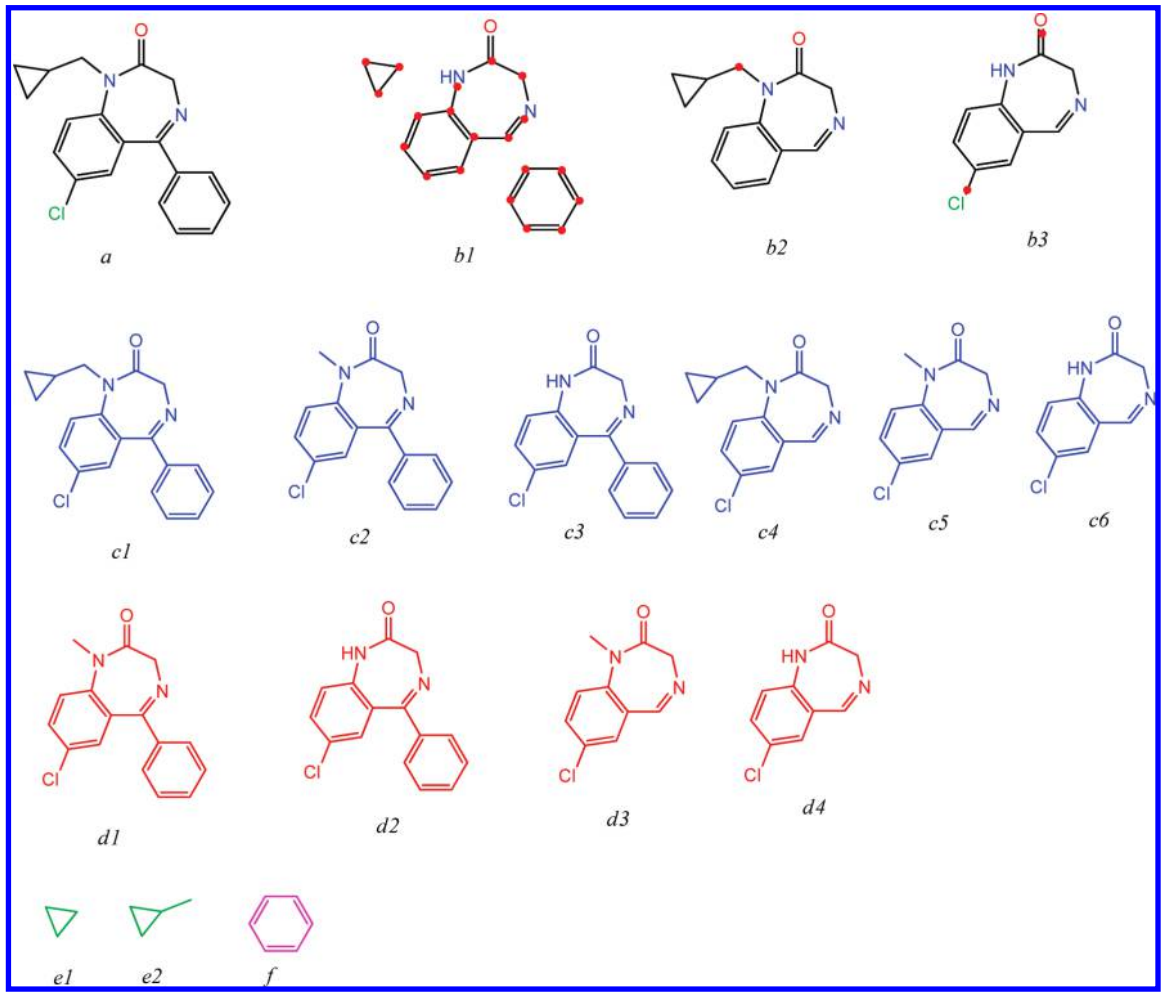

Figure 2. Examples of different types of building blocks. The parent molecule *a* is dissected into a set of fragments. According to Bemis and Murcko's scheme,[15,16] the red-dotted atoms in *b1* are atoms for three ring systems, the red-dotted atom in *b2* is a linker atom, and the two red-dotted atoms in *b3* are side-chain atoms. According to our definition, all fragments *c1* to *c6* are drug scaffolds (DS), all fragments *d1* to *d4* are ring systems (RS), and *e1* and *e2* are valid small fragments (SF). The phenyl group *f* is not considered as a SF or RS.

category I, if the ratio of is smaller than 0.5; category II, if the ratio is equal to or larger than 0.5 but smaller than 1.0; category III, if the ratio is equal to or larger than 1.0 but smaller than 2.0; category IV, if the ratio is equal to or larger than 2.0 but smaller than 5.0; and category V, if the ratio is equal to or larger than 5.0.

3. RESULTS AND DISCUSSION

3.1. Datasets. In this work, two drug datasets, an approved drug dataset and an extended drug dataset, have been applied to perform building block analysis. It is arguable what kind of dataset should be used to conduct drug building-block analysis. A point of view is that only the marketed drugs are suitable for scaffold analysis. The problem is that the FDA-approved drugs are less than 1400 and the drug analysis based on such a small dataset may lead to biased results. Moreover, the result of drug building block analysis based on approved drugs may be less helpful in developing new drugs targeted on novel protein targets. Another point of view is to use both the approved and the experimental drugs, even bioactive compounds to conduct the scaffold analysis. As a matter of fact, most drug building block analysis used extended drug datasets. Lipinski's "Rule of 5", a popular drug-likeness filter, was also derived from an analysis of 2245 drugs from the World Drug Index (WDI) believed to have entered phase II clinical trials. The building block analysis using the extended drug set is more useful in developing new drugs targeted on novel protein targets. In this study, both an approved drug dataset and an extended drug set were prepared to conduct building block analysis.

What is the background frequency of occurrence of a given building block in chemical space? The background frequency can be estimated using the hit rate of this building block in a diverse screening database supposing the database is big enough to cover all kinds of chemicals. In this study, the screening database, SDS, was prepared from 6.2 million ZINC "clean-drug-like" molecules through diverse selection. The background frequencies are used to perform classification for the building blocks.

3.2. Top Drug Scaffolds. The accumulated occurrences in percentage as a function of the number of top fragments are shown in Figure 3 for the DS building blocks. It is shown that the trends of the accumulated occurrence in a function of top scaffolds are essentially the same for the two drug datasets. The top 50 DSs cover 52.6% and 48.6% of drugs for ADDS and EDDS, respectively. For the next 50 scaffolds, only 5.1% and 7.6% of new drug entries are covered for the ADDS and EDDS, respectively. In Table 1, the accumulated occurrences of the top 50, 100, 500, and 1000 hits are listed for drug scaffolds with different definitions. Again, the trends

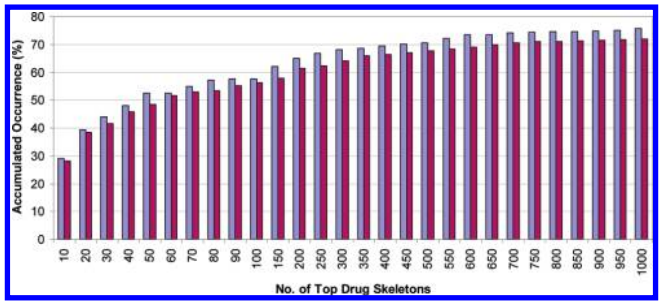

Figure 3. Statistics on the accumulated occurrence (%) for the top 1000 drug scaffolds identified in two drug building block analyses. Blue, the approved drug dataset; violet, the extended drug dataset.

of accumulated occurrences are similar for the two drug datasets. It should be pointed out that the drug scaffold definition of a molecular fragment having at least nine heavy atoms is always applied except when DS is redefined specifically. It is understandable that the larger a drug scaffold is, the less frequent it shows up in the drug database.

The top 100 drug scaffolds are shown in Figures 4 and 5 for the building block analysis using ADDS and EDDS, respectively. The DS building blocks are grouped into five categories according to the ratio of the frequency of occurrence in the drug dataset to the background frequency estimated with SDS. The following is the color code for each category: category I, red; category II, magenta; category III, black; category IV, blue; and category V, green. The same color scheme is used for two other kinds of building blocks, RS and SS. The Smarts annotations of the top DS as well as their hit rates and building block categories are listed in Tables S1 and S2 of the Supporting Information. If one wants to define his or her own drug scaffolds, he or she can retrieve the top building blocks with Tables S1 and S2. We want to point out that none of the nonring atoms in those fragments match any ring atoms during the database searches.

As shown in Figures 4 and 5, there is a large overlap between the top DSs of the two drug sets. For the top 50 DSs identified with ADDS, 41 are ranked as the top 100 DSs for EDDS. On the other hand, only six top 50 DSs of EDDS do not show up in the top 100 DSs of ADDS. Compared to previous work in drug scaffold analysis, the top DSs identified in this study cover many more drug molecules. For example, in Bemis and Murko's study, the top 42 frameworks only account for 24% of their CMC dataset.[15] It should be pointed out that most of Bemis and Murko's frameworks are ring systems, and thus it is more appropriate to make a comparison using the RS set below.

From Figures 4 and 5, one can find some large fragments having similar chemical structures, such as DSs 33−38 of Figure 4a. Those fragments are likely from a set of similar drugs, such as antibiotics. When the identified building blocks are used in combinatorial chemistry and drug design, one can skip those fragments if his or her purpose is to develop drugs other than antibiotics.

3.3. Top Ring Systems. The accumulated occurrences in percentage as a function of the number of top fragments are shown in Figure 6 for the ring systems. A total of 76 and 206 ring system building blocks were identified for the approved drug set and the extended drug set, respectively. A total of 76 RSs cover 41.4% of ADDS, while 206 RSs cover 44.3% of EDDS. Specifically, the top 40 RSs cover 36.5% and 32.6% drug molecules for ADDS and EDDS, respectively. The percents of coverage of the top 40 RSs are much higher than that covered by the 42 frameworks in the Bemis and Murcko's work (24%). The top 50 ring systems of both drug datasets are shown in Figure 7. It is shown that most RSs belong to categories IV and V for both the drug sets. For example, 9-(tetrahydrofuran-2-yl)-9H-purine (#33 of Figure 7a and #22 of Figure 7b) is the common structure of 11 approved deoxyguanosine drugs, which include adenosine monophosphate, vidarabine, cladribine, clofarabine, didanosine, fludarabine, nelarabine, and so forth. The Smarts annotations of the top RS as well as their hit rates and building block categories are listed in Tables S3 and S4 of the Supporting Information.

Similar to DS, there is a large overlap between the top RSs identified in both drug datasets. Though the sequence orders are slightly different, 20 common building blocks show up in the top 25 RSs of both drug datasets. Moreover, all of the top 25 RSs of ADDS show up in the top 50 RSs of EDDS, and only one top 25 RS of EDDS does not show up in the top 50 RSs of ADDS.

3.4. Top Small Fragments. The accumulated occurrences in percentage as a function of the number of top fragments are shown in Figure 8 for the SF set. It is notable again that the phenyl group is not considered as a SF. The top 50 SF building blocks cover 92.8% and 90.8% for ADDS and EDDS, respectively. In Bemis and Murcko's work,[16] the average number heavy atoms of a side chain is two, much smaller than our SF building blocks, which have at least three heavy atoms. The small fragments in this work are more like the linking atoms than the side chains in Bemis and Murcko's definitions. The top 50 small fragments for both drug datasets are shown in Figure 9 for both ADDS and EDDS. Again, the top small fragments are essentially the same for the approved drug dataset and the extended drug dataset. Only one top 25 SF of ADDS does not show up in the top 25 SFs of EDDS. The Smarts annotations of the top

Table 1. Accumulated Occurrences (%) of Top 50, 100, 500, and 1000 Drug Fragments for Different Definitions of Drug Scaffolds

	approved drug dataset				extended drug dataset			
no. heavy atoms	50	100	500	1000	50	100	500	1000
≥7	74.7	84.0	88.3	90.8	74.2	80.6	85.6	87.7
≥8	55.2	63.8	78.1	82.5	52.7	61.7	74.7	78.7
≥9	52.6	57.7	70.7	75.6	48.6	56.2	67.7	72.0
≥10	40.9	49.0	62.8	68.2	38.8	45.9	59.1	63.5
≥11	21.7	32.8	52.1	61.0	21.0	33.5	46.6	54.6
≥12	19.6	26.0	40.1	50.6	17.7	24.1	36.1	43.6
≥13	17.4	17.4	33.5	42.3	13.4	18.1	29.0	34.0
≥14	11.5	15.5	27.7	35.8	8.8	11.5	23.0	27.9
≥15	11.2	13.2	22.2	30.9	6.1	8.4	18.3	22.4

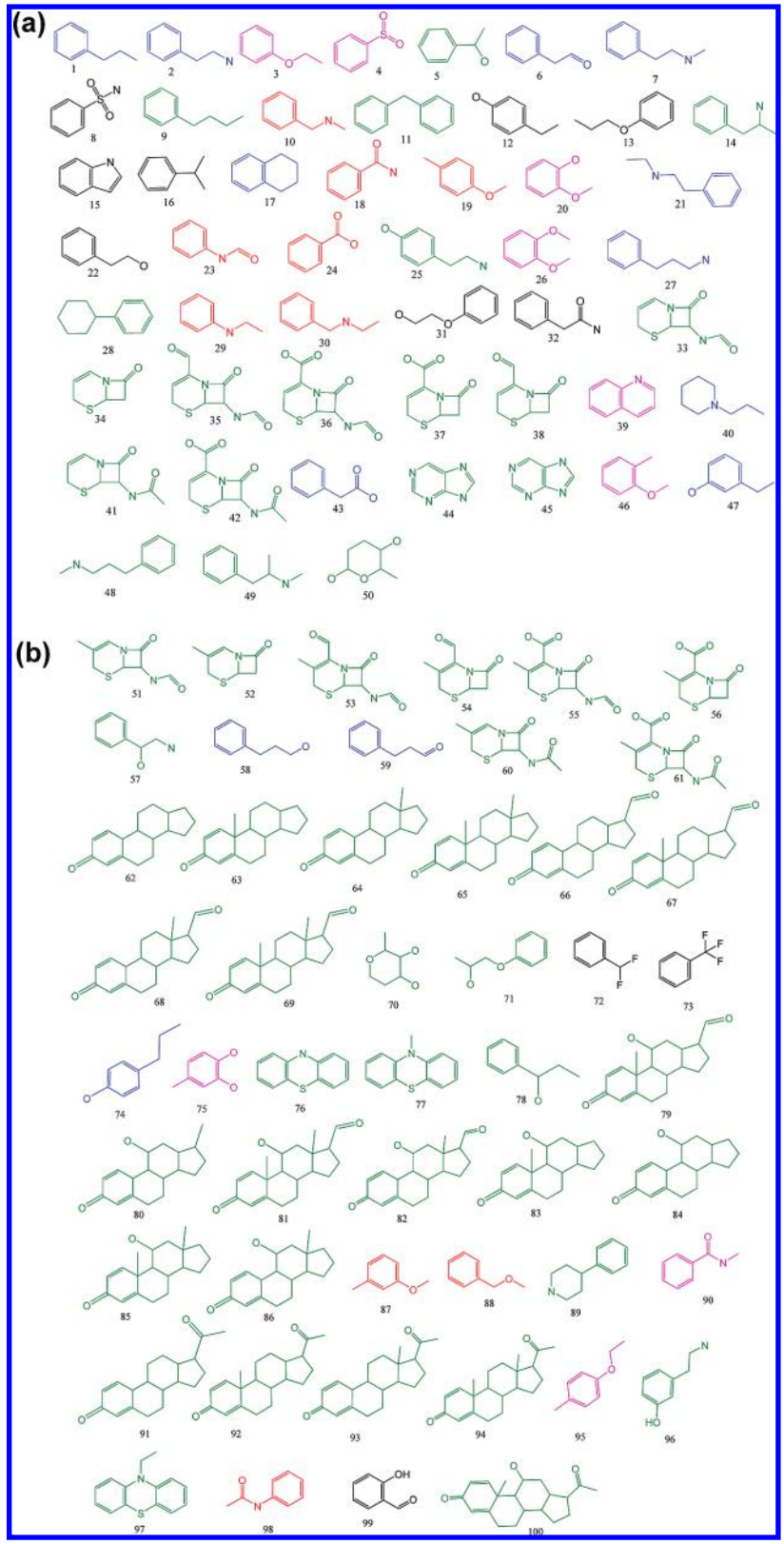

Figure 4. The top 100 drug scaffolds identified from the 1240 FDA approved drugs. The drug scaffolds are colored according to their building block categories: red (category I), magenta (category II), black (category III), blue (category IV), and green (category V). (a) 1−50 and (b) 51−100.

SFs as well as their hit rates and building block categories are listed in Tables S5 and S6 of the Supporting Information.

With an examination of the top building blocks of DS, RS, and SF, one can find that some large building blocks are made of some smaller building blocks. It is pointed out that this is a natural feature of building block analysis when an iterative algorithm of cutting rotatable bonds is utilized. Almost all of the drug scaffold analysis has this feature.

3.5. Building Block Classification. Hundreds and thousands of building blocks can be generated using the recursive dissection algorithm even for a small dataset like ADDS. Certainly those building blocks cannot be treated equally or

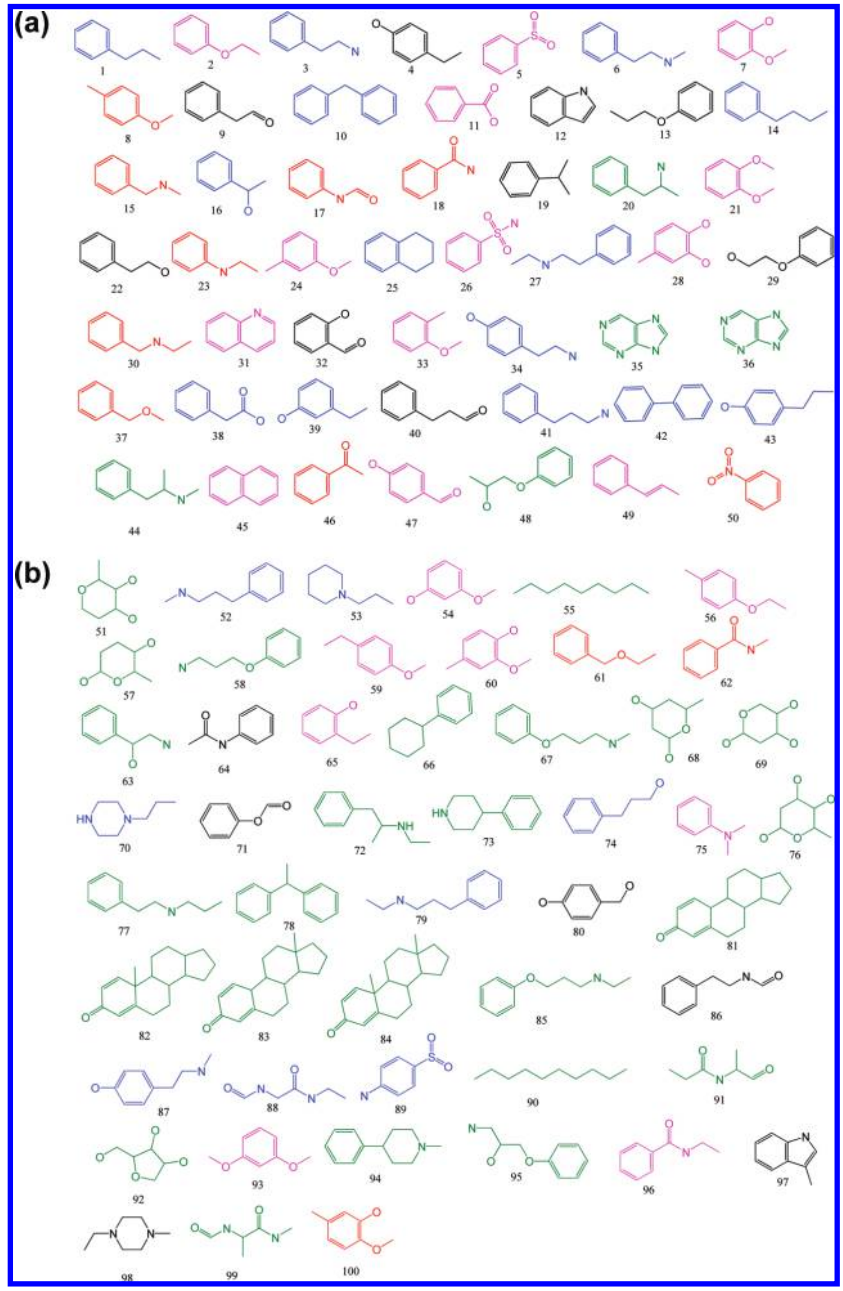

Figure 5. The top 100 drug scaffolds identified using the extended drug dataset (6932 entries). The drug scaffolds are colored according to their building block categories: red (category I), magenta (category II), black (category III), blue (category IV), and green (category V). (a) 1−50 and (b) 51−100.

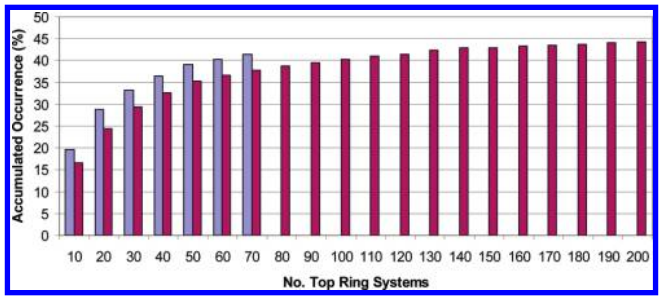

Figure 6. Statistics on the accumulated occurrence (%) for the top ring systems identified in two drug building block analyses. Blue, the approved drug dataset; violet, the extended drug dataset.

purely on the basis of their occurrence in the drug dataset. If so, we can conclude that the phenyl functional group is one of the most important building blocks since most drugs (66.2%) contain it. However, we know that the above conclusion is not valid or useful given the fact that more than 90% of screening compounds contain at least one phenyl group. To the best of our knowledge, there is no parameter from the chemical structure aspect to describe how drug-like a building block is or whether it is functional or compositional.

The ratio of the frequency occurrence of a building block in a drug dataset to that in the screening dataset may serve as a measure of drug-likeness of the building block. However, this ratio is very sensitive to the composition of the screening dataset. Take the nitro functional group as an example; at least two marketed drugs, ranitidine and nizatidine, have a nitro functional group, while no compound in the 1.9 million-compound SDS has this functional group at all. That is to say, the background frequency of nitro is 0, and the calculated frequency ratio is ridiculously large. To attenuate

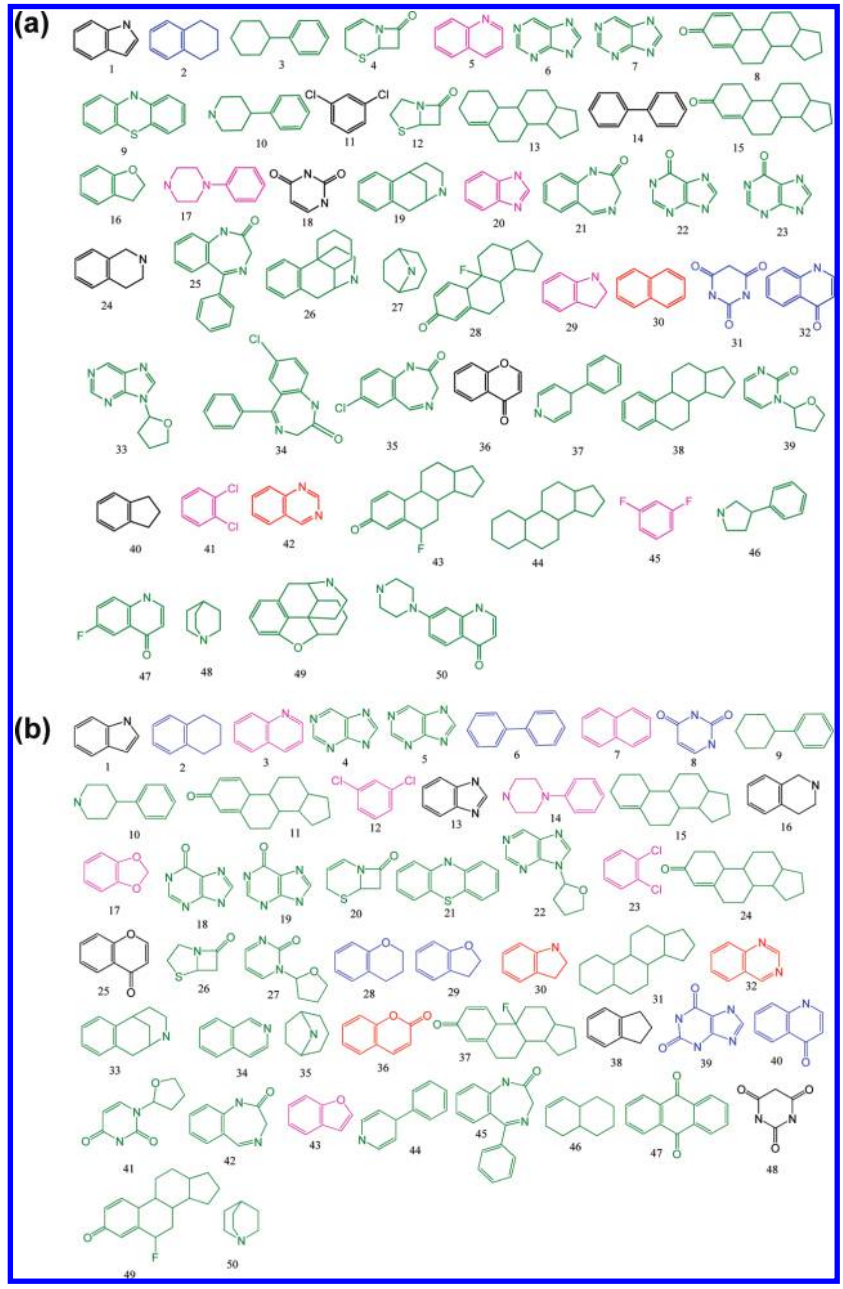

Figure 7. The top 50 ring systems identified using the two drug datasets, (a) the approved drug dataset and (b) the extended drug dataset. The drug scaffolds are colored according to their building block categories: red (category I), magenta (category II), black (category III), blue (category IV), and green (category V).

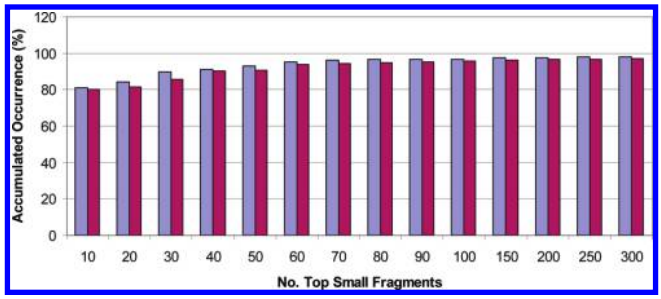

Figure 8. Statistics on the accumulated occurrence (%) for the top small fragments identified in two drug building block analyses. Blue, the approved drug dataset; violet, the extended drug dataset.

the sensitivity of the frequency ratio to the composition of the screening dataset, we proposed a classification scheme. All of the building blocks belong to one of five categories. The building blocks in categories IV and V are more drug-like than those in categories I, II, and III. However, it should be emphasized that the wrong class assignment likely occurs when the background frequency of a building block is not accurate. Although it may still be disputable, the above classification scheme is our first attempt to describe the drug-likeness of the building blocks of drugs. It is our hope that other more appropriate approaches come into being to describe the drug likeness of building blocks inspired by our work.

3.6. Comparison to Previous Studies. A striking difference of the numerous studies on drug building block analysis is that different databases were utilized: CMC by Bemis and Murcko,[15,16] WDI by Lee and Schneider,[18] marketed drugs by Siegel and Vieth,[19] and a collection of marketed drugs and experimental drugs in our current work. Other differences

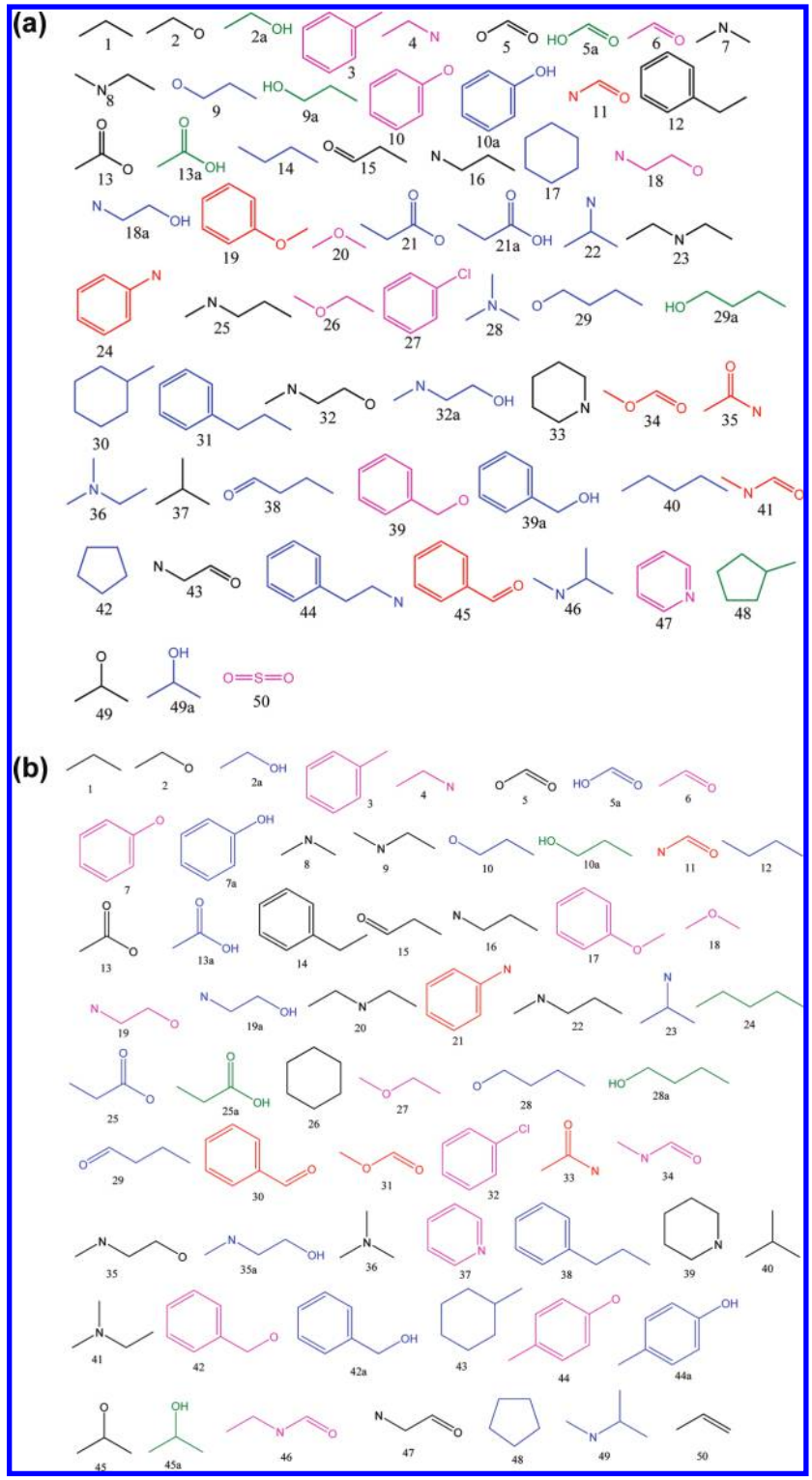

Figure 9. The top 50 small fragments identified using the two drug datasets, (a) the approved drug dataset and (b) the extended drug dataset. The drug scaffolds are colored according to their building block categories: red (category I), magenta (category II), black (category III), blue (category IV), and green (category V).

include the types of building blocks, the molecular dissection algorithm, and so forth. The side-by-side comparison of the drug building block analysis work is presented in Table 2.

It is interesting to investigate how many top building blocks in one study also show up in the others. To our surprise, none of the top 42 atomic frameworks (Chart 2 of Bemis and Murcko's work[15]) show up in Siegel and Vieth's DIODs (drugs in other drugs, listed in the Table 2 of the paper[19]) and vice versa. In the Lee and Schneider work,[18] a set of ring systems found in drugs but not in the collection of natural products were listed in Figure 3 of the paper.[18] Not surprisingly, none of those ring systems shows up in Bemis and Murcko's frameworks or Siegel and Vieth's DIODs.

We then compared our three types of building blocks with those identified by others. Encouragingly, 12 of the top 15 frameworks, except benzene, imidazole, and hydropyrimidine in Chart 2 of Bemis and Murcko's paper,[15] show up in either

Table 2. Comparison of Several Drug Scaffold Analysis Studies

studies	database	no. of entries	dissection algorithm	building block type	drug likeness assessment
Bemis and Murcko[14]	CMC	5120	deep-first-search	framework	kno
Bemis and Murcko[15]	CMC	6990	iteratively remove tail-ending atoms	side chain	no
Wang et al.[16]	RTECS	58 682	pattern recognition using Wiswesser line notation (WLN)	framework and structure pattern	FC[a] was used to measure the importance of a framework to a specific toxicity
Lee and Schneider[17]	WDI	4998	identification of cylic molecular subgraphs followed by removal of side chains and linkers	ring system	topological pharacophore pattern was used to discriminate drugs from no-drugs; no assessment for the identified building blocks
	ACD	4282			
Siegel and Vieth[18]	marketed drugs	1386	substructure search	marketed drug	no
Wang and Hou (this work)	marketed drugs	1240	recursive dissecting the rotatable bonds of the parent molecule and the generated fragments	drug scaffold	building blocks were grouped into one of five categories based on the ratios of frequencies of occurrence in the drug and screening datasets
	market drugs experimental drugs from WDI, DrugBank	6932		ring system	
	subset of ZINC	1 950 000		small fragment	

[a] Functional coefficient, which is defined as the ratio of hit rates of a framework in a specific toxicity subset and the whole toxicity database, respectively.

the top 50 RS sets, the top 100 DS sets, or the top 50 SF sets in this work. As pointed out above, benzene is not considered as a RS or SF; instead, it is a functional group which frequently occurs in both drug datasets and the screening dataset. In other words, only two top 15 frameworks identified by Bemis and Murcko are missing in our top building blocks. The following are the symbolized correspondences: 1 (benzene) ∼, 2 ∼ RS [13, 15], 3 ∼ DS [11, 10], 4 ∼ RS [8, 11], 5 ∼ SF [47, 37], 6 ∼ RS [30, 7], 7 ∼ RS [25, 45], 8 ∼ RS [9, 21], 9 ∼ SF [17, 26], 10 ∼ RS [7, 5], 11 ∼ RS [33, 22], 12 ∼ RS [5, 3], 13 (hydropyrimidine)∼, 14 (imidazole) ∼, and 15 ∼ DS [11, 10]. Here, "2 ∼ RS [13, 15]" denotes that the second framework in Chart 2[15] corresponds to the 13th RS for ADDS and 15th RS for EDDS. The other symbols can be explained in a similar way. In contrast, only eight top 15 ring systems of ADDS (RSs 1, 5, 7, 8, 9, 10, 13, and 14 in Figure 7a) and seven top 15 ring systems of EDDS (RSs 1, 3, 5, 6, 7, 11, and 15 in Figure 7b) in this work show up in the top 42 frameworks of Bemis and Murcko.

When comparing our building blocks to Siegel and Vieth's DIODs,[19] we observed a very similar phenomenon. Most of the 18 DIODs listed in Table 2 of the paper have the exact or very similar counterparts in the top building blocks (top 100 of DS and top 50 of RS sets) of this work. The following are the symbolized correspondences: 1 ∼ DS [2, 3], 2 ∼ DS [25, 34], 3 ∼ DS [24, 11], 4 ∼ DS [18, 18], 5 (hydroquinone) ∼, 6 ∼ DS [25, 34], 7 ∼ DS [8, 26], 8 ∼ DS [50, 69], 9 ∼ RS [26, 33], 10 ∼ DS [69, 83], 11 ∼ DS [96, 63], 12 (ceftizoxime) ∼ [42,], 13 (sodium oxybate, the total number of heavy atoms is less than eight, cannot be recognized as a DS or RS in this work) ∼, 14 (cefetamet) ∼ [42], 15 ∼ DS [50, 76], 16 ∼ RS [34, 45], 17 (cyclizine lactate) ∼, and 18 ∼ DS [49, 72]. On the contrary, for both ADDS and EDDS, only three top 18 DSs and one top 18 RS of our work have counterparts in DIODs.

In Lee and Schneider's paper,[18] a set of ring systems (five clusters, 16 ring systems) which show up in the drug database but not in the natural products were listed. It is understandable that those ring systems are not necessarily the top fragments occurring in the drug dataset. Even though, we found exactly the same or very similar counterparts for three ring systems belonging to three different clusters in our top RS sets for both ADDS and EDDS.

On the basis of the fact that there were no overlapping building blocks among the previous studies, and most of the top building blocks in each study have counterparts in our top DS, RS, and SF sets but not the opposite, we conclude that our drug database is adequate for identifying drug scaffolds and our recursive dissection algorithm performs well. Moreover, we identified many more high-quality building blocks than the previous studies.

3.7. Application of Drug Building Blocks in Drug Discovery. We think many individual building blocks in the DS set and RS set as well as their combinations can serve as core structures for combinatorial chemistry experiments. In combinatorial chemistry, side fragments are linked to core structures to form final products. It is ideal if the side fragments, the small building blocks, are drug like. In this study, we also identified a set of small fragments that are highly frequently used by drugs. The functional SFs and FGs can be utilized as either linkers or side chains to extend the core structures or to splice two cores.

The structural analysis of the drug building blocks (drug scaffolds, ring systems, small fragments) should have great applications in drug discovery. First of all, it provides useful hints for medicinal chemists to wisely select scaffolds, ring systems, and fragments to design potential drugs in their efforts of drug discovery. The advances of generating new chemicals through scaffold hopping was recently reviewed by Zhao.[30] Second, as discussed above, the DS and RS building blocks or their combinations can serve as the core structures and the SF and FG building blocks as side chains in combinatorial chemistry. Third, both DS and RS building blocks can be used as anchors in fragment-based drug design.[31−33] Finally, a structural drug likeness analysis can

be carried out for screening compounds as well as virtual compounds designed by computational combinatorial chemistry. A structural drug likeness score is calculated according to how many drug-like scaffolds and fragments occur in the compound in question. The structural drug likeness in combination with the molecular property based on drug likeness provides a general filter to prioritize screening libraries prior to experimental HTS.

As discussed above, it is problematic to use the frequency ratio of a building block to quantitatively measure its drug likeness directly. Large errors can occur when the background frequencies cannot be estimated correctly. In this work, we have introduced a classification scheme to effectively attenuate the sensitivity of the frequency ratio to the composition of the screening dataset. In order to improve the background frequency calculation, we have used a large screening dataset that is a collection of screening libraries from many vendors to eliminate the bias caused by one or several vendors. Though great effort has been put to reduce the background calculation errors, misclassification can inevitably occur for some building blocks when vendors intentionally eliminate some types of compounds. However, we should not demand perfection for such kind of classification scheme since it is better than none. Even for the widely used drug-likeness filter, "Rule of 5", about 7.6% drugs violate at least two rules of the "Rule of 5" in our approved drug dataset. Specifically, 932 drugs satisfy four rules, 213 satisfy three rules, 67 violate two rules, 27 violate three rules, and one drug violates four rules. It should be pointed out that to determine how important a building block is, both the frequency of occurrence in the drug dataset and the category the building block belongs to should be taken into consideration.

4. CONCLUSION

In this study, a brute force approach has been applied to analyze the molecular structures in two drug datasets, an approved drug dataset (1240 entries) and an extended drug dataset (6932 entries including both the FDA-approved drugs and experimental drugs). Three types of building blocks, that is, drug scaffolds, ring systems, and small fragments, were identified and ranked according to their frequencies of occurrence in the drug dataset. It is found that drug molecules utilize a limited number of fragments as building blocks to construct biologically functional entities. For both drug datasets, the top 50 drug scaffolds cover about half of the drugs in that dataset. Each building block was further grouped into one of five categories using the ratio of the hit rate in a drug dataset to that in the screening dataset. The application of the structure-based building block analysis is also discussed in the text.

ABBREVIATIONS

IC_{50}, the concentration of an inhibitor that inhibits 50% of the enzyme; ADME-Tox, Absorption, Distribution, Metabolism, Excretion, and Toxicity; DS, drug scaffold; RS, ring system; SF, small fragment; NME, new molecular entity; HTS, high-throughput screening; WDI, World Drug Index; CMC, Comprehensive Medicinal Chemistry; ACD, Available Chemicals Directory; RTECS, the Registry of Toxic Effects of Chemical Substances; DIODs, drugs in other drugs; ADDS, the approved drug dataset; EDDS, the extended drug dataset; SDS, the screening dataset.

ACKNOWLEDGMENT

We are grateful for the research support from OpenEye Scientific Software, Santa Fe, New Mexico.

Supporting Information Available: The Smarts annotation of the drug fragments used for database searches, hit rates, and the building block categories are listed in Tables S1−S6. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Dobson, C. M. Chemical space and biology. *Nature* **2004**, *432*, 824–828.
(2) Blum, L. C.; Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732–8733.
(3) Drugs@FDA. http://www.accessdata.fda.gov/Scripts/cder/DrugsatFDA/ (accessed Dec 2009).
(4) Kola, I.; Landis, J. Can the pharmaceutical industry reduce attrition rates. *Nat. Rev. Drug Discovery* **2004**, *3*, 711–715.
(5) Leeson, P. D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discovery* **2007**, *6*, 881–890.
(6) Muegge, I. Selection criteria for drug-like compounds. *Med. Res. Rev.* **2003**, *23*, 302–321.
(7) Clark, D. E.; Pickett, S. D. Computational methods for the prediction of 'drug-likeness'. *Drug Discovery Today* **2000**, *5*, 49–58.
(8) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
(9) Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I.; Matsushita, Y. Simple method of calculating octanol water partition-coefficient. *Chem. Pharm. Bull.* **1992**, *40*, 127–130.
(10) Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. A comparison of physiochemical property profiles of development and marketed oral drugs. *J. Med. Chem.* **2003**, *46*, 1250–1256.
(11) Vieth, M.; Sutherland, J. J. Dependence of molecular properties on proteomic family for marketed oral drugs. *J. Med. Chem.* **2006**, *49*, 3451–3453.
(12) Wang, J. M.; Krudy, G.; Xie, X. Q.; Wu, C. D.; Holland, G. Genetic algorithm-optimized QSPR models for bioavailability, protein binding, and urinary excretion. *J. Chem. Inf. Model.* **2006**, *46*, 2674–2683.
(13) Wang, J. M.; Krudy, G.; Hou, T. J.; Zhang, W.; Holland, G.; Xu, X. J. Development of reliable aqueous solubility models and their application in druglike analysis. *J. Chem. Inf. Model.* **2007**, *47*, 1395–1404.
(14) Wang, J.; Hou, T. Recent Advances on in silico ADME Modeling. *Ann. Rep. Comput. Chem.* **2009**, *5*, 101–127.
(15) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
(16) Bemis, G. W.; Murcko, M. A. Properties of known drugs. 2. Side chains. *J. Med. Chem.* **1999**, *42*, 5095–5099.
(17) Wang, J. S.; Lai, L. H.; Tang, Y. Q. Structural features of toxic chemicals for specific toxicity. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1173–1189.
(18) Lee, M. L.; Schneider, G. Scaffold architecture and pharmacophoric properties of natural products and trade drugs: Application in the design of natural product-based combinatorial libraries. *J. Comb. Chem.* **2001**, *3*, 284–289.
(19) Siegel, M. G.; Vieth, M. Drugs in other drugs: a new look at drugs as fragments. *Drug Discovery Today* **2007**, *12*, 71–79.
(20) Sutherland, J. J.; Higgs, R. E.; Watson, I.; Vieth, M. Chemical fragments as foundations for understanding target space and activity prediction. *J. Med. Chem.* **2008**, *51*, 2689–2700.
(21) Grabowski, K.; Schneider, G. Propertises and Architecture of Drugs and Natural Products Revisited. *Curr. Chem. Biol.* **2007**, *1*, 115–127.
(22) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008**, *36*, D901–D906.
(23) Gao, Z. T.; Li, H. L.; Zhang, H. L.; Liu, X. F.; Kang, L.; Luo, X. M.; Zhu, W. L.; Chen, K. X.; Wang, X. C.; Jiang, H. L. PDTD: a web-

accessible protein database for drug target identification. *BMC Bioinf.* **2008**, 9.
(24) *The Derwent World Drug Index*, version 3; Derwent Publications, Ltd: London, 1999.
(25) Hou, T. J.; Wang, J. M.; Zhang, W.; Xu, X. J. ADME evaluation in drug discovery. 6. Can oral bioavailability in humans be effectively predicted by simple molecular property-based rules. *J. Chem. Inf. Model.* **2007**, *47*, 460–463.
(26) Hou, T. J.; Wang, J. M.; Zhang, W.; Xu, X. J. ADME evaluation in drug discovery. 7. Prediction of oral absorption by correlation and classification. *J. Chem. Inf. Model.* **2007**, *47*, 208–218.
(27) Irwin, J. J.; Shoichet, B. K. ZINC - A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(28) Clark, R. D.; Langton, W. J. Balancing representativeness against diversity using optimizable K-dissimilarity and hierarchical clustering. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1079–1086.
(29) Stahl, M.; Guba, W.; Kansy, M. Integrating molecular design resources within modern drug discovery research: the Roche experience. *Drug Discovery Today* **2006**, *11*, 326–333.
(30) Zhao, H. Y. Scaffold selection and scaffold hopping in lead generation: a medicinal chemistry perspective. *Drug Discovery Today* **2007**, *12*, 149–155.
(31) Zartler, E. R.; Shapiro, M. J. Fragonomics: fragment-based drug discovery. *Curr. Opin. Chem. Biol.* **2005**, *9*, 366–370.
(32) Erlanson, D. A.; McDowell, R. S.; O'Brien, T. Fragment-based drug discovery. *J. Med. Chem.* **2004**, *47*, 3463–3482.
(33) Rees, D. C.; Congreve, M.; Murray, C. W.; Carr, R. Fragment-based lead discovery. *Nat. Rev. Drug Discovery* **2004**, *3*, 660–672.

CI900398F