A New Rapid and Effective Chemistry Space Filter in Recognizing a Druglike Database

Suxin Zheng,[†] Xiaomin Luo,*[,†] Gang Chen,[†] Weiliang Zhu,*[,†] Jianhua Shen,[†] Kaixian Chen,[†] and Hualiang Jiang*[,†,‡]

Shanghai Institute of Materia Medica, Shanghai Institutes of Biological Sciences, Chinese Academy of Sciences, 555 Zu Chong Zhi Road, Shanghai 201203, China, and School of Pharmacy, East-China University of Science and Technology, Shanghai 200237, China

Received January 27, 2005

To develop a new chemistry space filter with high efficiency and accuracy, an analysis on distributions of as many as 50 structural and physicochemical properties was carried out on both druglike and nondruglike databases, viz. MACCS-II Drug Data Report (MDDR), Comprehensive Medicinal Chemistry (CMC), and Available Chemicals Directory (ACD). Based on the analysis results, a chemistry space filter was developed that can effectively discriminate a druglike database from a nondruglike database. The filter is composed of two descriptors: one is a molecular saturation related descriptor, and the other is associated with the proportion of heteroatoms in a molecule. Both are molecular size independent. Therefore, the profiles of a druglike database could be characterized as proper molecular saturation and proper percentage of heteroatoms, revealing direct indices for designing and optimizing combinatorial libraries. The application of the new filter on the Chinese Natural Product Database (CNPD) suggested that CNPD is, as expected, a potential druglike database, testifying that the new filter is reliable. Therefore, this newly developed chemistry space filter should be a potent tool for identifying druglike molecules, thus, it would have potential applications in the research of combinatorial library design and virtual high throughput screening using computational approaches for drug discovery.

1. INTRODUCTION

Recently, one of the essential trends in the drug research has been the integration of what has traditionally been considered as 'development' activities into the early phases of drug discovery.[1] For this purpose, an in vitro ADME/T (absorption, distribution, metabolism, excretion, and toxicity) screen has been implemented into the early stage of drug discovery for discarding compounds that are likely to fail further down the line, i.e., researchers are trying their best to identify and/or design a subset of druglike molecules from the vast expanse of what could possibly be synthesized.[1−3] However, the experimental filters of ADME/T have several limitations, such as time-consuming, resource-intensive, and requiring compound samples.[4] Thus, great efforts have been contributed to the development and application of computational methods for predicting drug likeness.[1,4−6] Nowadays, drug likeness prediction has already been, to some extent, integrated into computational drug design paradigm (e.g. virtual screening) in identifying and/or designing compounds not only with good biological activity but also with a good druglike profile.[7−9]

Some major progress has been achieved in the development of computational methods for druglike predication.[1,5,6] In general, these methods can be categorized into two classes: (1) methods for predicting general molecular drug likeness by computational filters constructed based on a set of available drug and/or druglike molecules and (2) methods for predicting special pharmacological profiles, such as solubility, permeability, intestinal absorption, BBB penetration, and pharmacokinetic properties, by using quantitative structure− (physicochemical) property relationship (QSPR) models generated based on a series of known drugs or druglike molecules.[4] The former is usually applied in combinatorial library design for differentiating druglike from nondruglike molecules, and the latter is normally used in lead optimization.[6] The best-known method of drug likeness prediction is the "rule of 5" developed by Lipinski and coworkers[10] by analyzing 2245 available drugs from the World Drug Index (WDI).[11] According to the "rule of 5", compounds should have poor absorption if they have any two of the following characters: hydrogen bond (H-bond) donors (the sum of O−H and N−H groups) are more than 5; H-bond acceptors (the sum of N and O atoms) are more than 10; the molecular weight (MW) is greater than 500; and the calculated octanol/water partition coefficient (ClogP) is larger than 5. But it is much qualitative and cannot describe the significant differences between the MACCS-II Drug Data Report (MDDR)[12] and the Available Chemicals Directory (ACD)[13] databases.[14] Using the Comprehensive Medicinal Chemistry (CMC)[15] and the MDDR as representatives of druglike molecule databases and the ACD as a surrogate of nondruglike molecules, Ajay et al.,[16] Sadowski et al.,[17] and Frimurer et al.[18] have constructed neural network models to classify druglike and nondruglike molecules. In the neural network analyses, both one-dimensional parameters, including molecular weight, ISIS keys (topological indexes),[19]

* Corresponding author phone: 86-21-50807188; fax: 86-21-50807088; e-mail: hljiang@mail.shcnc.ac.cn. Corresponding author address: Shanghai Institute of Materia Medica, 555 Zu Chong Zhi Road, Shanghai 201203, China (H.J.).

[†] Chinese Academy of Sciences.

[‡] East-China University of Science and Technology.

10.1021/ci050031j CCC: $30.25
Published on Web 06/10/2005

Ghose and Crippen atom types,[20] and two-dimensional parameters, e.g. functional groups, were used as descriptors. Thus, in nature these methods belong to the first class of methods. A genetic algorithm-based approach has been developed by Gillet et al.,[21] to distinguish between druglike and nondruglike compounds. Simple descriptors (such as MW, the numbers of H-bond donors and acceptors, rotatable bonds, and aromatic rings) and a topological shape descriptor have been used in model construction. Compounds from the WDI were assumed to be a druglike data set, and compounds from the SPRESI database[22] were assumed to be a nondruglike data set. Wagener et al.[23] have established a model of decision trees to discriminate between drugs and nondrugs. Although these approaches (e.g. neural network approaches, genetic algorithm-based approaches, and decision tree approaches) may distinguish between compounds that are druglike and nondruglike, they are database-dependent, and the training data sets impact the analysis models.[24] Accordingly, they can only recognize those compounds that resemble existing drugs as druglike, compounds from completely new classes could be misclassified, holding back their application in lead optimization.[1,6]

In consequence, it is desirable to develop a new filter that is database independent for distinguishing drugs from nondrugs. To this end, Muegge et al.[24] have developed a pharmacophore point filter to recognize druglike molecules. Some other simple descriptors could also be useful in classifying druglike and nondruglike molecules. For instance, Pareto (80/20 principle) analyses[14] on MDDR and ACD databases revealed that 70% of the druglike compounds have characters of $0 \leq$ HBD (hydrogen bond donors) ≤ 2, $2 \leq$ HBA (hydrogen bond acceptors) ≤ 9, $2 \leq$ RGB (rotatable bonds) ≤ 8, and $1 \leq$ RNG (number of rings) ≤ 4. Obviously, pharmacophore and simple descriptor filters that are database independent are different from neural network, genetic algorithm-based, and decision tree approaches and thus may be used in druglike prediction for new compounds. However, larger compounds bear, on average, more functional groups, which may produce more HBAs, HBDs, RGBs, and pharmacophore points. Accordingly, these filters are likely to prefer larger molecules. Muegge et al.[24] have pointed out that the capability of pharmacophore filter and the neural network approach in discriminating the ACD data set from druglike data sets decreases if the ACD data set is constructed based on molecules with greater molecular weights. This weakness may result from their used descriptors that are, to a certain extent, correlated with the molecular size.

Considering the advantage and disadvantage of the available computational methods of predicting drug likeness, we developed a new chemistry space filter, which is molecular size independent, based on the ratios of different molecular descriptors. Testing analyses on the ACD, MDDR, and CMC indicated that it is capable of differentiating drugs from nondrugs. Because the filter was developed based on simple, database-independent, and molecular size uncorrelated rules, it not only can be used to discriminate between druglike and nondruglike databases but also has the potential for identifying a novel druglike molecular subset from a newly designed combinatorial library. Application of the filter to the CNPD (Chinese Natural Product Database)[8,9,25] showed that the CNPD has a score close to CMC, indicating that the CNPD is a potential source for drug discovery. This result is expected, for natural products have been the major molecular structural resources for lead discovery,[9,26,27] and CNPD in fact contains a lot of drug and druglike molecules, such as anti-Alzheimer's disease drug huperzine A[28] and antimalarial drug artemisinin.[29]

Table 1. Number of Database Entries

	ACD	MDDR	CMC
initial	379808	122729	8474
modified database with molecular weight cutoffs of 78 and 600	289665	80486	6465
modified database with molecular weight cutoffs of 78 and 750	293487	87266	6678

2. METHODS

Preparation of Databases. For developing a new filter, the CMC and MDDR databases were used as druglike data sets. ACD is a database including the compounds commercially available, MDDR is a database containing those compounds reported with biological activity, and CMC is a database composed of those compounds launched as a drug. There is no completely nondruglike database available to date, so ACD is taken as a nondruglike data set in this study as other studies did.[14,16,18,24] However, ACD, which is a common database for virtual screening, certainly contains some druglike compounds. We removed from ACD those compounds that are also found in MDDR or CMC to reduce the influence from overlapping data. In details, all the databases have undergone several steps below to exclude the following compounds from each database: compounds with missing or invalid structures; compounds with atoms other than C, H, N, O, S, P, and X (halogen, F, Cl, Br, I); all the compounds have entries in either MDDR or CMC were removed from ACD. In addition, we eliminated all compounds with no therapeutic activity from CMC,[30] including radiopaque agents, contrast agents, solvents, anesthetics, disinfectants, flavoring agents, pharmaceutical aids, surgical aids, dental, surfactants, ultraviolet screens, emetics, preservatives, aerosol propellants, chelators, keratolytics, insecticides, astringents, laxatives, sweeteners, dental caries prophylactics, adhesives, dentistry, pharmaceutic aids, veterinary, buffers, scabicides, and ectoparasiticides. Antineoplastic drugs and anticancer drugs were also removed from the databases because they are often highly cytotoxic and are likely to react with protein targets.[30,31] Opera[14] have demonstrated that the subsets of ACD and MDDR with reactive compounds removed have a difference of less than 4% in drug likeness in comparison with the complete databases, so we did not remove the molecules with reactive functional groups. Since almost all the molecules with MW less than 78 do not have enough functional groups, and too large molecules may have poor absorption property, we set the low and high cutoffs for the molecular weight as 78 and 600, respectively. Furthermore, to estimate the influence of molecular weight cutoff on the drug likeness of the remainder data sets, 3 more databases with the molecular weight from 78 to 750 were constructed, namely ACD_1, MDDR_1, and CMC_1, respectively. Table 1 summarized the information of these data sets.

All the studied molecules were converted into the format of Sybyl mol2 using the program "sdf2mol2" in the toolkit

Table 2. Definitions of the Newly Developed Descriptors

descriptor	definition
	Single
RGB	number of rigid bonds
RNG	number of five-, six-, and seven-membered rings
C3	number of sp^3 hybridized C atoms
A3	number of sp^3 hybridized C, O, S, and N atoms
N	number nitrogen atoms
O	number of oxygen atoms
C	number of carbon atoms
UNC	number of sp, sp^2 and aromatic C atoms
UNA	number of atoms with type of 'C1', 'C2', 'Car', 'O2', 'Oco2', 'So', 'So2', 'N1', 'N2', 'Nam', 'Nar', and 'Npl3'
BD2	number of double bonds
BD3	number of triplet bonds
BDAR	number of aromatic bonds
AUH	number of atoms rather than H and X
BDUH	number of the bonds which do not contain H and X atoms
UNSAT	RNG+BD2+2*BD3+(BDAR+1)/2
	Ratio
C3P	C3/AUH
UNSATP	UNSAT/BDUH
UNC_C3	UNC/C3
UNA_A3	UNA/A3
A3_C	A3/C
NO_C3	(N+O)/C3

of DOCK4.0,[32] A ring, which is purely composed of X.2 or X.ar atoms (X = C, O, N, S and P) but without X.2 of exocyclic double bond, is considered as aromatic. Programs for data analyses were developed with C++ language in our center (http://www.dddc.ac.cn).

Descriptors. As many as 50 initial descriptors about molecular physicochemical properties were designed. After testing and optimization, 15 of them were found to be of value. Table 2 lists their names and definitions. Based on the 15 optimized simple descriptors, 6 new descriptors, viz. NO_C3, UNC_C3, UNA_A3, C3P, UNSATP, and A3_C, were developed, which are ratios of the 15 descriptors (Table 2). The descriptor NO_C3 stands for the ratio of the total number of oxygen and nitrogen atoms to the number of carbon atoms with sp^3 hybridization. This descriptor is essential as the oxygen and the nitrogen atoms constitute almost all the common pharmacophores in drugs and play important roles in biological activity and absorption. The descriptor UNC_C3 represents the ratio of the number of unsaturated carbon atoms to the number of sp^3 carbon atoms and, therefore, is related to molecular saturation and rigidity. UNA_A3 is the ratio of the number of all aromatic atoms to the number of sp^3 carbon atoms and, thus, is associated with molecular aromaticity. These two descriptors were designed because the aromatic and unsaturated atoms are ubiquitous in drug molecules. The other three descriptors UNSATP, C3P, and A3_C are associated with molecular insaturation and flexibility as well.

3. RESULTS AND DISCUSSIONS

Setting Values for New Descriptors. Opera[14] demonstrated that the "rule of 5" is not a good filter for differentiating druglike and nondruglike databases, for the "rule of 5" could not detect the difference between the ACD and MDDR data sets. To overcome the shortage of the "rule of 5", Opera used a set of filters including RNG and RGB (the definitions of these two descriptors are described in Table 2) for screening chemistry spaces. They demonstrated that 63% of ACD compounds have $0 \le \text{RNG} \le 2$ and $\text{RGB} \le 17$, while 29% of ACD compounds have $3 \le \text{RNG} \le 13$ and $18 \le \text{RGB} \le 56$. In contrast, 61% of the MDDR compounds are in the space of dimension with $\text{RNG} \ge 3$ and $\text{RGB} \ge 18$, and only 25% of the MDDR compounds are located in the range of $0 \le \text{RNG} \le 2$ and $\text{RGB} \le 17$. We also applied RNG and RGB to filter the ACD, MDDR, and CMC databases. The results are shown in Table 3, which indicate that the RNG and RGB, as indicated by Opera,[14] can differentiate the ACD and MDDR data sets; however, they cannot efficiently discriminate the ACD from CMC data sets (Table 3). It is unreasonable, for molecules in CMC should be more druglike than those in MDDR. Therefore, RNG and RGB are also not effective descriptors in discriminating druglike and nondruglike data sets.

To explore the possible cause that the RNG and RGB are not effective, distributions of molecular weight (MW) between 78 and 600 for these data sets were analyzed, and the result is shown in Figure 1. The peak of MW distribution for ACD is located at 350, while those for CMC and MDDR are situated at 300 and 400, respectively. Meanwhile, the average MW was also calculated for these data sets, and the result is listed in Table 4. It suggests that compounds in MDDR are on average larger than those in ACD and CMC. This indicates that it is difficult, from the MW distributions, to discriminate druglike and nondruglike data sets. However, in general, larger compounds have high possibility with larger values of RNG and RGB than smaller ones have. Accordingly, Opera's RNG and RGB filters are compound-size dependent. This may be one of the reasons that Opera's filters could not discriminate CMC and ACD but could discriminate MDDR and ACD, because the molecular size of ACD is averagely close to CMC but relatively far apart from MDDR (Table 4).

Different from Opera's descriptors, our new descriptors, viz. NO_C3, UNC_C3, UNA_A3, C3P, UNSATP, and A3_C, are MW independent as they are the ratios of the molecular physicochemical properties (Table 2). Figure 2 depicted their distributions in the ACD, MDDR, and CMC databases.

Figure 2a is the C3P distribution of molecules in the databases of ACD, MDDR, and CMC, which are asymmetrical Gaussian distributions. It is clear that the characteristic of the C3P distribution in ACD is different from that in other databases. For instance, the distribution peak is located at C3P = 0.05 in ACD, while it is at 0.25 and 0.30 in MDDR and CMC, respectively. Furthermore, Figure 2a shows that ACD library has the lowest distribution percentages among the 3 databases if C3P has values ranging from 0.20 to 0.65, suggesting that a library is more druglike if its C3P value ranges from 0.20 to 0.65. Table 5 shows that more than 66% compounds in both the MDDR and CMC have C3P values within this range, whereas only ~35% compounds in the ACD database do.

As shown in Table 2, C3P is the ratio of the number of sp^3 hybridized C atoms to the number of the total heavy atoms except halogen atoms. Thus, for druglike molecules, their C3P values should not be too small or too great. According to its distributions in different databases (Figure 2a), the initial value for C3P may be set as 0.20−0.65. Too small a C3P value means too high percentages of heteroatoms

Table 3. Distributions Based on Opera's "Druglike" Chemistry Space Filters

limitation	ACD	ACD_1	MDDR	MDDR_1	CMC	CMC_1
0 ≤ RNG ≤ 2 RGB ≤ 17	56.03%	55.54%	23.43%	21. 90%	47.01%	46.20%
RNG ≥ 3 RGB ≥ 18	34.76%	35.22%	63.37%	65.08%	40.11%	40.97%

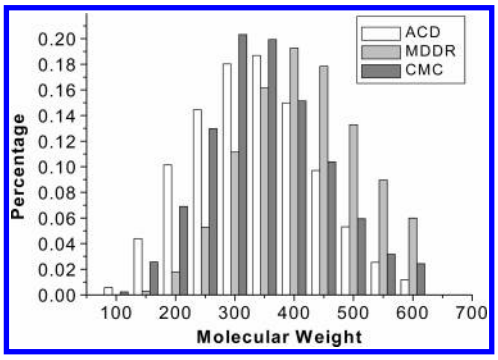

Figure 1. Distribution of molecular weight.

Table 4. Average Molecular Weights of Data Sets

	ACD	MDDR	CMC
average MW (78−600)	309.9	392.0	325.4
average MW (78−750)	314.5	413.7	336.1

(O, N, S, P) or/and unsaturated atoms such as aromatic carbon atoms. These molecules are usually nondruglike due to their over hydrophilicity (molecules with too many heteroatoms cannot penetrate the cell wall), rigid (molecule with a lot of multiple bonds and/or aromatic rings are not beneficial to binding to receptors), or toxicity (molecules with too many aromatic rings, especially condensed aromatic ring structure, are noxious to human being). On the other hand, a molecule with too large a C3P value is most probably nondruglike, because they are more alkanelike.

Distributions of UNSATP for the ACD, MDDR, and CMC are shown in Figure 2b, which is asymmetrical as well. When the UNSATP value is between 0.15 and 0.40, the ACD library has the lowest distribution among the 3 databases (Figure 2b). Thus, a library may be more druglike if its UNSATP value is located in the region from 0.15 to 0.40. Table 5 suggests that ∼57% and 62% of the compounds in MDDR and CMC are found within this region, respectively, while only ∼32% of the compounds in ACD are in this region. UNSATP stands for molecular insaturation. The larger the UNSATP value, the greater the molecular insaturation. Therefore, the value of UNSATP for a good drug candidate should be within a certain range. Based on the distribution profiles of UNSATP in different databases, initial value for UNSATP could be assigned between 0.15 and 0.40.

Figure 2c shows distributions of UNC_C3 for ACD, MDDR, and CMC. By using the same method as for C3P and UNSATP, the UNC_C3 was initially set between 0.4 and 2.4 (Table 5). Table 5 shows that druglike databases have a much higher UNC_C3 distribution within this range than a nondruglike database. For the descriptor is the ratio of the number of unsaturated carbon atoms to the number of sp^3 carbon atoms; it is somewhat similar to the descriptor UNSATP in nature.

Distributions of UNA_A3 in the databases are shown in Figure 2d, which indicates that UNC_A3 has a similar profile to the UNC_C3 distribution. Its initial value can also set to be between 0.4 and 2.4 for it has almost the same characteristic as UNSATP. Apparently, UNA_A3 distribution in the ACD is always less than the MDDR and CMC within this range. In addition, its physicochemical meaning is similar to that of UNC_C3, i.e., it is associated with molecular insaturation.

Distributions of A3_C for ACD, MDDR, and CMC are shown in Figure 2e and Table 5. Accordingly, its initial value was established as 0.3−0.8 based on its distribution characteristics for the databases. More than 62% of the compounds in druglike databases are included, but only 32% of the compounds in ACD were found in the range, respectively. The descriptor A3_C is related to molecular saturation according to its definition (Table 2). Therefore, this descriptor is also similar to C3P in nature.

Distributions of NO_C3 in ACD, MDDR, and CMC are shown in Figure 2f, based on which the initial value of NO_C3 could be set as 0.20−1.20. More than 68% of the MDDR and 63% of the CMC compounds, but only about 39% of the ACD compounds, can pass this criterion (Table 5), indicating that oxygen and nitrogen atoms are important for druglike compounds. It is expected because both oxygen and nitrogen are the most general elements of various pharmacophores and may interact with targets through hydrogen bonds. This descriptor should be related to molecular hydrophobicity as the introduction of N and O elements to hydrocarbon compounds will significantly affect molecular partition coefficient. However, molecules with too high or too low a ratio of heteroatoms to carbon atoms are usually not good drug candidates either. One of the reasons might be that it is difficult for the molecules with too high a NO_C3 value to permeate various types of membranes, whereas it is very alkanelike for the molecules with too low a NO_C3 value.

It is noticeable that the distributions of the new descriptors for the databases ACD_1, MDDR_1, and CMC_1 are almost the same as those for the databases of ACD, MDDR, and CMC (Table 5). More impressively, these descriptors are capable of discriminating the drug likeness among the databases with similar molecular weight distribution, such as CMC and ACD (Table 4), demonstrating again that all the six descriptors are molecular size independent. Apparently, these descriptors have several advantages: (1) They are primarily capable of distinguishing between druglike and nondruglike databases (Figure 2 and Table 5), because each descriptor has its own criterion value. (2) Each of the new descriptors has its clear definition and explicit physicochemical meaning. For example, C3P is the ratio of the number of sp^3 hybridization C atoms to the number of the total heavy atoms except X atoms (Table 2). Therefore, C3P value stands for molecular saturation. (3) As indicated above, the descriptors are ratio properties and are molecular size independent, thus they are able to be used in filtering a newly designed compound library for lead optimization.

However, as discussed above, 5 of the new descriptors, e.g., C3P, UNSATP, UNA_A3, UNC_C3, and A3_C, are all related to the molecular saturation. Thus, only one of them

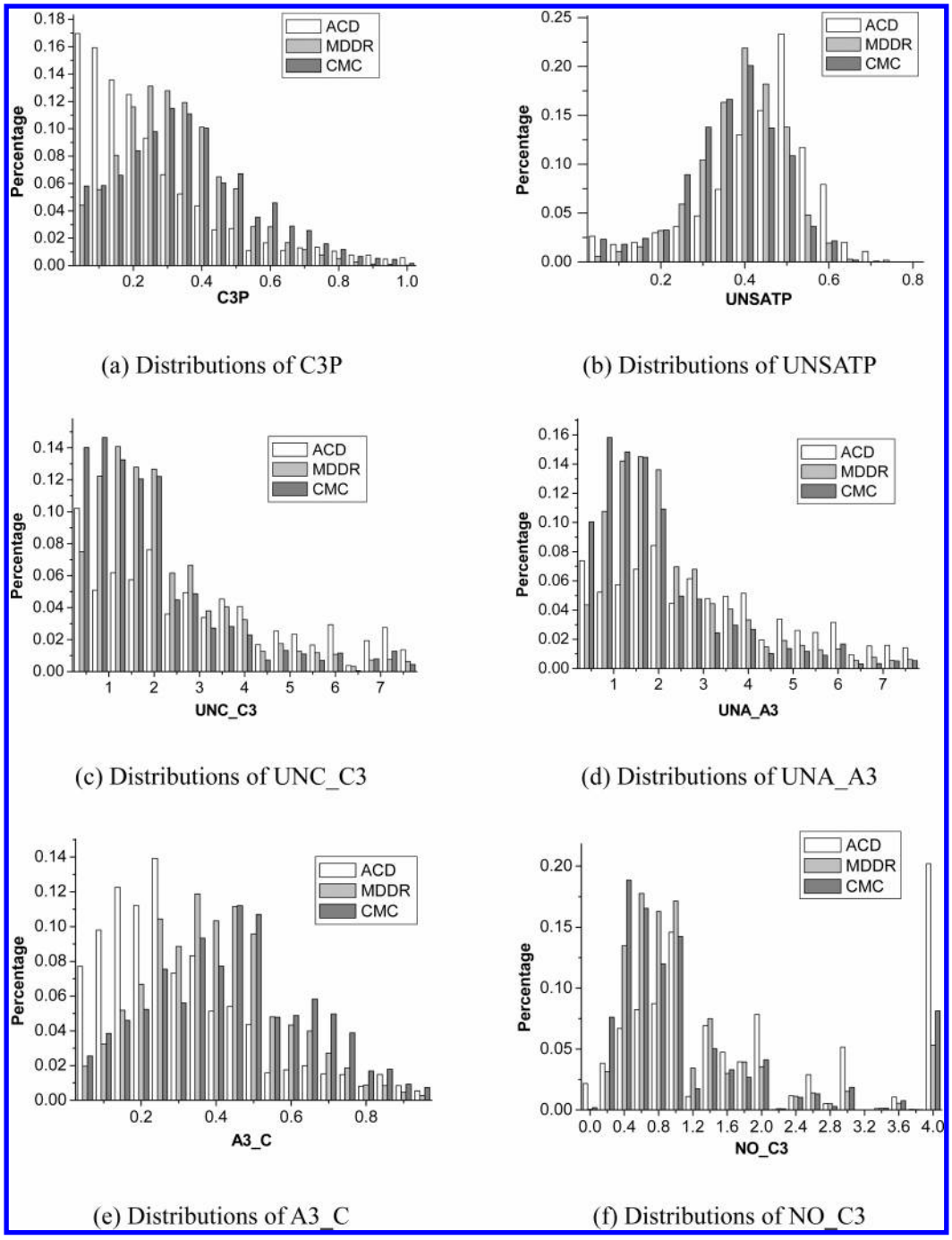

Figure 2. Distributions of the six descriptors for ACD, MDDR, and CMC. The *x*-axis is the value of descriptor; the *y*-axis is the percentage of molecules past the corresponding descriptor over all molecules in the corresponding database. All the analyses were performed for the molecules with a molecular weight between 78 and 600 in the databases. The columns at NO_C3=4.0 in (f) is a sum of distributions for all the molecules with NO_C3 greater than 3.9.

Table 5. Initial Values and Distributions for the 6 New Descriptors and the New Chemistry Space Filter

	initial value	ACD	MDDR	CMC	ACD_1	MDDR_1	CMC_1
C3P	$0.2 < C3P \leq 0.65$	34.74%	67.43%	66.19%	34.95%	68.30%	66.32%
UNSATP	$0.15 < UNSATP \leq 0.40$	31.76%	57.17%	62.73%	32.00%	59.21%	63.15%
UNA_A3	$0.4 < UNA\_A3 \leq 2.4$	30.63%	60.03%	60.99%	30.80%	60.76%	60.93%
UNC_C3	$0.4 < UNC\_C3 \leq 2.4$	28.22%	57.90%	56.62%	28.38%	58.79%	56.74%
A3_C	$0.3 < A3\_C \leq 0.80$	32.31%	61.53%	65.01%	32.48%	62.28%	64.99%
NO_C3	$0.2 < NO\_C3 \leq 1.2$	39.35%	68.09%	63.38%	39.54%	68.81%	63.57%
chemistry space filter	$0 \leq UNSATP \leq 0.43$ $0.10 \leq NO\_C3 \leq 1.8$	39.09%	69.05%	70.39%	39.38%	70.18%	70.77%

is necessary to dealing with molecular saturation. We studied their correlations with the heteroatom proportion descriptor NO_C3 and decided to use UNSATP as a representative descriptor for molecular saturation as it is mostly an independent descriptor from NO_C3 (Supporting Information, Table S1). Thus, two descriptors, NO_C3 and UNSATP,

Table 6. Distribution of Rings

	ACD	MDDR	CMC
have at least one ring	92.63%	98.24%	95.14%
have at least one aromatic ring	86.73%	91.88%	82.23%
have at least one nonaromatic ring	26.55%	62.54%	56.15%

were finally selected for scoring drug likeness of a compound database. To further improve the descriptors' performance and robustness, we combined the UNSATP with NO_C3 to create a new chemistry space filter. Taken into account that ACD is not a pure nondruglike database, the final value for the new descriptors was reset and optimized to get maximum capability of the chemistry space filter in differentiating druglike and nondruglike compound databases (Table 5). The result indicates that this chemistry space filter can distinguish the MDDR and CMC more efficiently than individual NO_C3 or UNSATP. The cutoff of molecular weight has almost no influence on the result. About 70% of CMC and MDDR compounds pass through the new chemistry space filter, while just about 39% of ACD molecules are screened out by the filter. Furthermore, the distribution of CMC in that space is slightly higher than that of MDDR, which is inconsistent with the common sense that CMC should be more druglike than MDDR. Therefore, this filter is suitable as a chemistry space filter to discriminate drug likeness and nondrug likeness.

As the new filter is atom type based, it shows a favorable computational speed. The evaluation of drug likeness can be completed within 2 h for the MDDR database (~80 000 compounds) and 4 h for the ACD database (~300 000 compounds) on a computer with a CPU of 733 MHz and RAM of 256M.

Aromatic and Nonaromatic Rings in Druglike Molecules. The new chemistry space filter reveals a lower unsaturation ($0 \leq$ UNSATP ≤ 0.43) characteristic of compounds in a druglike database than in a nondruglike database, indicating that druglike molecules generally contain more sp^3 atoms. This seems reasonable, because molecules with high values of UNSATP contain more unsaturated bonds and aromatic rings, leading to bad ADME/T properties. However, as a common sense, a rigid structure is favorable for binding a ligand to its target in terms of entropy change, because docking a very flexible structure to the binding site of the target protein would lead to a significant loss of entropy. Thus, cyclization of sp^3 atoms might be an effective way to reduce entropy loss for a molecule binding to a receptor.[33] We performed a statistic analysis on the ring numbers for the compounds in the ACD, MDDR, and CMC databases (Table 6). The result indicates that more than 82% of the molecules in all the databases contain aromatic rings. More interestingly, $\geq$56% of the compounds in MDDR and CMC have at least one nonaromatic ring, while only 26% compounds in ACD encode nonaromatic (saturated) rings. This statistic analysis demonstrates that nonaromatic rings may be one of the causes of the difference between a nondruglike database, ACD, and a druglike database, MDDR and CMC.

Application in Druglike Analysis of CNPD. Natural products have some different structure properties in comparison with synthetic compounds and available drugs, such as less nitrogen, halogen, and sulfur atoms while more oxygen atoms.[27,34] Thus the difficulty has been recognized in characterizing the drug likeness of a natural product database based on the knowledge of known drugs.[35,36] The Chinese Natural Product Database (CNPD)[25] is a typical natural product database, which contains around 45 000 compounds isolated mostly from Chinese herbs. To evaluate its drug likeness, the new filter was applied which releases a score of 72.91%, very close to CMC, indicating that CNPD is a druglike database. This makes sense, for CNPD in fact contains a lot of drugs or druglike molecules, such as the antimalarial drug *artemisinin* and the anti-Alzheimer's disease drug *huperzine* A.[9,28,29] In addition, natural products have a higher possibility than synthetic compounds to be drugs or lead compounds. Among the 520 new drugs approved between 1983 and 1994, 157 were natural products or derived from natural products and more than 60% of antibacterials and anticancer drugs originated from natural products.[37−39] Recently, using a virtual screening approach in conjunction with an electrophysiological assay, we have found 4 natural compounds showing higher inhibitory activities to the potassium ion channel, and their I_K block activities are 20−1000 times higher than that of tetraethylammonium (TEA).[8] All of these suggest that the CNPD database would be a good source for new drug discovery.

4. CONCLUSIONS

We have designed and tested as many as 50 newly designed descriptors, 15 of which were discovered to be appropriate descriptors in differentiating between druglike and nondruglike databases. Six new descriptors, viz. NO_C3, UNC_C3, UNA_A3, C3P, UNSATP, and A3_C, have been developed as ratios of the 15 optimized descriptors, which are more potent than the originals and are molecular size independent. Notably, these descriptors have comprehensible definitions and physicochemical meanings for druglike properties and drug-receptors interactions, which overcome the problem of the "black box" shortcomings of the neural network methods. These six descriptors could be divided into two classes: molecular saturation-related descriptors and heteroatom proportion descriptors. Among these descriptors, it was found that UNSATP, as a representative of the molecular saturation-related descriptor, is mostly independent of the heteroatom proportion descriptor NO_C3. Thus, a new chemistry space filter was successfully developed by combining the two descriptors. The new filter demonstrates that a druglike molecule should have appropriate proportions of the nitrogen, oxygen, aromatic atom, sp^3 hybridization atom, aromatic, and nonaromatic rings. By using the new filter, the Chinese Natural Product Database (CPND) was tested to be very close to CMC in terms of drug likeness, indicating this database is a good source for drug discovery by using computational virtual screening.

In summary, we have successfully developed a new chemistry space filter for distinguishing a druglike database from a nondruglike database. It is molecular size independent, fast, and more efficient in comparison with the existing filters. Therefore, it should be very useful in combinatorial library design and virtual screening for drug discovery.

ACKNOWLEDGMENT

This work was supported by grants from the National Basic Research Program of China (2003CB11401,

2002CB512802, and 2004CB518901), the 863 Hi-Tech Program (2002AA233061, 2003AA235010, and 2004AA104270). We thank Professor Lai for providing the XLogP program for partition coefficient calculations.

Supporting Information Available: Correlation coefficient between heteroatom proportion descriptor NO_C3 and the 5 molecular saturation-related descriptors for the databases of ACD, MDDR, and CMC (Table S1). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Clark, D. E.; Pickett, S. D. Computational methods for the prediction of 'drug likeness'. *Drug Discovery Today* **2000**, *5*, 49−58.
(2) Di, L.; Kerns, E. H. Profiling druglike properties in discovery research. *Curr. Opin. Chem. Biol.* **2003**, *7*, 402−408.
(3) Smith, D. A.; van de Waterbeemd, H. Pharmacokinetics and metabolism in early drug discovery. *Curr. Opin. Chem. Biol.* **1999**, *3*, 373−378.
(4) Walters, W. P.; Murcko, A.; Murcko, M. A. Recognizing molecules with druglike properties. *Curr. Opin. Chem. Biol.* **1999**, *3*, 384−387.
(5) Walters, W. P.; Murcko, M. A. Prediction of 'drug likeness'. *Adv. Drug Deliver. Rev.* **2002**, *54*, 255−271.
(6) Muegge, I. Selection criteria for druglike compounds. *Med. Res. Rev.* **2003**, *23*, 302−321.
(7) Good, A. C.; Krystek, S. R.; Mason, J. S. High-throughput and virtual screening: core lead discovery technologies move towards integration. *Drug Discovery Today* **2000**, *5*, 61−69.
(8) Liu, H.; Li, Y.; Song, M.; Tan, X.; Cheng, F., et al. Structure-Based Discovery of Potassium Channel Blockers from Natural Products: Virtual Screening and Electrophysiological Assay Testing. *Chem. Biol.* **2003**, *10*, 1103−1113.
(9) Shen, J.; Xu, X.; Cheng, F.; Liu, H.; Luo, X. et al. Virtual Screening on Natural Products for Discovering Active Compounds and Target Information. *Curr. Med. Chem.* **2003**, *10*, 2327−2342.
(10) 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 Deliver. Rev.* **1997**, *23*, 3−25.
(11) World Drug Index is available from Derwent Informatin, London, U.K. Website www.derwent.com.
(12) MACCS-II Drug Data Report is available from MDL Information Systems Inc., San Leandro, CA, 94577 and contains biologically active compounds in the early stages of drug development.
(13) The Available Chemicals Database was provided by MDL Information Systems, Inc., 1999.
(14) Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251−264.
(15) Comprehensive Medical Chemistry is available from MDL Information Systems Inc., San Leandro, CA, 94577 and contains drugs already in the market.
(16) Ajay; Walters, W. P.; Murcko, M. A. Can We Learn To Distinguish between "Druglike" and "Nondruglike" Molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.
(17) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(18) Frimurer, T. M.; Bywater, R.; Narum, L.; Lauritsen, L. N.; Brunak, S. Improving the Odds in Discriminating "Druglike" from "NonDruglike" Compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1315−1324.
(19) SSKEYS, MDL Information Systems Inc., San leandro, CA.
(20) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic physicochemical parameters for three-dimensional structure directed quantitative structure−activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.
(21) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.
(22) The SPRESI database is produced by the All-Union Institute of Scientific and Technical Information of the Academy of Science of the USSR (VINITI) in Moscow and the Central Information Processing for Chemistry (ZIC) in Berlin. This database consists of data extracted from 1000 journals and patents, books, and other sources from 1975 to 1990. SPRESI is distributed by Daylight Chemical Information Systems, Inc., Mission Viejo, CA.
(23) Wagener, M.; van Geerestein, V. J. Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 280−292.
(24) Muegge, I.; Heald, S. L.; Brittelli, D. Simple Selection Criteria for Druglike Chemical Matter. *J. Med. Chem.* **2001**, *44*, 1841−1846.
(25) Chinese Natural Product Database is available from the Shanghai Institute of Material Medica and contains the Chinese natural compounds published.
(26) Nisbet, L. J.; Moore, M. Will natural products remain an important source of drug research for the future? *Curr. Opin. Biotech.* **1997**, *8*, 708−712.
(27) Henkel, T.; Brunne, R. M.; Müller, H.; Reichel, F. Statistical Investigation into the Structural Complementarity of Natural Products and Synthetic Compounds. *Angew. Chem., Int. Ed.* **1999**, *38*, 643−647.
(28) Jiang, H.; Luo, X.; Bai, D. Progress in Clinical, Pharmacological, Chemical and Structural Biological Studies of Huperzine A: A Drug of Traditional Chinese Medicine Origin for Treatment of Alzheimer's Disease. *Curr. Med. Chem.* **2003**, *10*, 2231−2252.
(29) Li, Y.; Wu, Y. An over four millennium story behind qinghaosu (artemisinin)- -a fantastic antimalarial drug from a traditional chinese herb. *Curr. Med. Chem.* **2003**, *10*, 2197−2230.
(30) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(31) Wang, J.; Ramnarayan, K. Toward Designing Druglike Libraries: A Novel Computational Approach for Prediction of Drug Feasibility of Compounds. *J. Comb. Chem.* **1999**, *1*, 524−533.
(32) DOCK 4.0. Molecular Design Institute, San Francisco, CA, U.S.A.
(33) Xu, J.; Stevenson, J. Druglike Index: A New Approach To Measure Druglike Compounds and Their Diversity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177−1187.
(34) 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.
(35) Feher, M.; Schmidt, J. M. Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218−227.
(36) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308−1315.
(37) Nielsen, J. Combinatorial synthesis of natural products. *Curr. Opin. Chem. Biol.* **2002**, *6*, 297−305.
(38) Cragg, G. M.; Newman, D. J.; Snader, K. M. Natural products in drug discovery and development. *J. Nat. Prod.* **1997**, *60*, 52−60.
(39) Newman, D. J.; Cragg, G. M.; Snader, K. M. The influence of natural products upon drug discovery. *Nat. Prod. Rep.* **2000**, *17*, 215−234.

CI050031J