Model-Free Drug-Likeness from Fragments

Oleg Ursu and Tudor I. Oprea*

Division of Biocomputing, Department of Biochemistry and Molecular Biology and University of New Mexico (UNM) Center for Molecular Discovery, UNM School of Medicine, MSC11 6145, Albuquerque, New Mexico 87131

Received May 20, 2010

We developed a drug-likeness filter (DLF), starting from molecular fragments and molecular weight (MW), a key property relevant in drug design. The molecular fragments were selected from extended connectivity atom environments based on their occurrence ratio in our collection of drugs and "nondrugs". The DLF recalls 87.05% of compounds from DRUGS ($N = 3823$) and 40.25% of compounds from the Available Chemicals Directory, (ACD, $N = 178\ 0\ 11$), using molecular fragments only. By adding MW (under 600) as an additional filter, 78.81% of DRUGS and 40.17% of ACD are recalled. The DLF procedure was externally validated using the MDL Drug Data Report (MDDR) data set ($N = 169\ 277$): 78.45% of compounds were recalled using the molecular fragments only, while 65.64% pass the DLF-MW filter. Over 50% of a pesticides collection ($N = 1482$) passed the DLF, as these chemicals share molecular fragments with known drugs. Developed as a model-free filter, DLF is perhaps less useful in discriminating drugs from nondrugs but more likely to rapidly eliminate those chemicals rich in nondrug-like fragments. Since almost 40% of ACD, the standard reference set for nondrugs, contain drug-like molecules, by using a rule-based system such as DLF, one is less likely to mislabel nondrugs due to overfitting. Reliable benchmarks for nondrugs are not likely to exist since medicinal chemistry catalogs tend to be biased toward existing drugs.

INTRODUCTION

The increasing effort to transfer the chemical space map into biological space and the quest for high quality leads in drug discovery requires an improved effort of developing high-quality chemical libraries for high-throughput screening (HTS). The key to the design of high-quality HTS libraries is the "drug-likeness" of individual compounds. Conceptually, drug-likeness captures those "molecules which contain functional groups and/or have physical, chemical, and biological properties consistent with a majority of known drugs",[1] which has been subject to many studies starting with Pfizer's rule of five.[2] A variety of machine learning (neural networks, decision trees, support vector machines, etc.) tools and molecular descriptors as well as molecular property/functional groups filters have been applied on several drug/nondrug data sets (MDDR, WDI, CMC, ACD, etc.) in an attempt to develop a drug-likeness scoring procedure/filter with various levels of success.[3−23] These papers are reviewed elsewhere.[24] As a general trend, machine learning methods outperform filter-based methods, according to literature, by a margin of 10−20%. However, when confronted with a large number of input variables and large training sets, these machine learning tools might lead to overtraining/overfitting[25−27] with poor prediction accuracy beyond used training/validation sets. Drug-likeness prediction models published in the literature[3,4,6,13,20] employing machine learning tools have a prediction accuracy ranging from ∼75 to 90%. Such models often use more than half of the available data for training and hundreds of descriptors. Our results seem to indicate (see Results and Discussion Section) that almost 40% of the Available Chemicals Directory (ACD), after removing drug structures, contains chemical structures with a high content of drug fragments, i.e., similar to drug structures, which indicates that between 10 and 30% of the ACD structures which are similar to drugs are incorrectly classified, most likely due to the model ability to "memorize" data. To assess the true predictive power of such models, *independent validation sets* for drugs/nondrugs are required, which is beyond the scope of this study. In contrast, filter-based methods use fewer classes of variables and rules and have comparatively good performances while being less prone to overestimates. We refer to rule-based methods as "model-free" in order to emphasize that their utility resides in their suitability for very large data set evaluation, as decisions (pass/fail) are rooted in the statistical distribution of input variables as opposed to machine learning methods that employ (non) linear transformations on input variables. The rules underlying such filters are based on statistical evidence, e.g., the occurrence of fragments or the distribution of properties, and on chemical expertise, with the advantage of being linked to chemical moieties and direct interpretation. Perhaps this ease-of-use contributed to the wide adoption of filter-based techniques (e.g., the rule of five, lead-likeness filter) in compound selection for high-throughput and virtual screening.

Among many categories of two-dimensional (2D) molecular descriptors, a set that performs well in virtual screening is referred to as circular fingerprints or atomic signatures.[28−30] These circular fingerprints, computed using the Morgan algorithm, produce canonical atomic environments and features that encode enough structural information to compute any 2D descriptors.[31−33] The information-rich content

* Corresponding author. E-mail: toprea@salud.unm.edu. Telephone: (505) 272-3694.

10.1021/ci100202p © 2010 American Chemical Society
Published on Web 07/22/2010

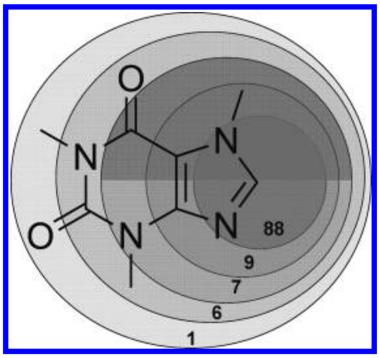

Figure 1. Fragmentation of the caffeine molecule up to five bonds radiuses. Fragment occurrence ratios are computed based on fragmentation of all chemical structures from DRUGS.

of these descriptors is attributed to systematic exploration of the chemical space describing a set of chemical structures. Use of atomic signatures to characterize drug/nondrug chemical space should provide a comprehensive mapping of all possible molecular fragments present in drug/nondrug libraries.

METHODS

Data Set Preparation. Our in-house collection composed of active principles only (no salts, formulations, or combinations included) for over 10 000 approved drugs worldwide served as the drugs data set ("DRUGS"; $N = 3{,}823$). The ACD from MDL/SYMYX (version 2002.1) was used as the "nondrugs" data set. After removing duplicates and compounds present in DRUGS, ACD contained 178 011 compounds. For external validation, we used the MDDR database from MDL/SYMYX (release 2006.2). After removal of duplicates and structures present in both DRUGS and ACD, MDDR contained 169 277 compounds.

Fingerprint Calculation. Circular fingerprints were computed using an in-house program written using Java and the JChem[34] application programming interface (API). Atomic environments of up to five bond radiuses were collected and saved for further processing (Figure 1).

While most of the drug molecules in DRUGS/ACD have a diameter below 10, some drug-like scaffolds, such as steroids, require a diameter of 10, hence the five bond radiuses were used to capture larger molecular fragments, which makes the filter more sensitive to larger drugs. The fingerprints generation algorithm stops fragment expansion before reaching the radius of five limit when all atoms from the input molecular structure are covered. A total of 1 418 622 unique atomic environments in both data sets were generated during fingerprint calculations.

Molecular Fragments Selection. The selection of those atomic environments or molecular fragments that are most relevant for drugs/nondrugs was performed using an occurrence-based scheme, as follows: Each newly generated fragment was assigned two probability values: one associated with DRUGS and another one associated with ACD. Only molecular fragments with an occurrence ≥ 3 ($\sim 0.1\%$) for the DRUGS data set and ≥ 100 ($\sim 0.1\%$) for the ACD data set were processed further. The following types of fragments were discarded: (i) fragments for which the probability values were equal; ii) fragments for which the maximum probability value for one of the data sets was less than twice the minimum probability value for the other data set. A total of 1 360 790 unique fragments were generated from the ACD data set. Of these, 7215 fragments have an occurrence ratio ≥ 100, and 4954 fragments have a relevant probability ratio, i.e., $p_{\mathrm{ACD}} \geq 2^{*}p_{\mathrm{DRUGS}}$. Out of the DRUGS data set, 88 037 unique fragments were generated, of which 12 970 have an occurrence ratio ≥ 3, and 11 016 fragments have a relevant probability ratio, i.e., $p_{\mathrm{DRUGS}} \geq 2^{*}p_{\mathrm{ACD}}$. The DRUGS data set is much smaller in size when compared to the ACD data set; however, the number of unique fragments derived from DRUGS that passes the above constrains is at least twice as large as the number of fragments derived from ACD. One possible explanation for this is that the DRUGS data set contains rather large series of molecules with, e.g., the cyclopentanoperhydrophenanthrene (steroid) scaffold (27), the beta-lactam (penicillin) core (45), and the cephem (cephalosporin) nucleus (59) as well as substituted biaryls in nonsteroidal anti-inflammatory drugs (17), etc. This reflects the "me-too" strategy in pharmaceutical research and development, as first-in-class drugs are quickly followed by molecules with similar scaffolds. These recurring structural patterns in the DRUGS data set nonetheless contribute to higher fragments occurrence ratios, whereas the ACD data set serves as a diversity source for chemicals and chemical reagents, where compound series with recurring patterns are not as numerous, and certain fragments have lower occurrence.

The final list of molecular fragments selected for DLF contains only fragments that have a probability of occurrence $p \geq 0.1\%$, a discrimination capability $p_{\mathrm{DRUGS}}/p_{\mathrm{ACD}} \geq 2$ for fragments derived from DRUGS, and a $p_{\mathrm{ACD}}/\mathrm{p}_{\mathrm{DRUGS}} \geq 2$ for fragments derived from ACD, respectively. The distribution of the fragment sizes and diameter and the discrimination capability are depicted in Figure 2. Most of the ACD-like fragments have 6−12 atoms (81%) compared to DRUG-like fragments which tend to have a more uniform distribution (Figure 2a); 76% of the ACD-like fragments have a diameter between 4 and 6, compared to DRUGS fragments, where 80% of the fragments have a diameter between 2 and 8 (Figure 2b). The discriminant power of DLF fragments increases rapidly with the diameter radius (Figure 2c). Thus, fragments with only one atom (diameter 0) have the lowest discriminant power, around 2, compared to that of large fragments (diameter 10), which have the highest discriminant power (≥ 500). This is not unexpected; as molecular fragments grow in size, the signal and attribute abilities become more specific for a particular data set. All 15 970 fragments were stored as SMARTS,[35] along with probability values for both the ACD and DRUGS data sets.

Filtering Process. Each of the three data sets (DRUGS, ACD, and MDDR) was submitted to the drug-like filter (DLF). For each molecular fragment where there was a substructure match, the probability values associated with that fragment were summed up, and the final sums were compared. Molecules pass the DLF if the sum of probability values for drug fragments is higher than that of nondrugs; it fails the DLF otherwise. The filtering process is similar to the Naïve Bayes classifier, the difference between our procedure and the former being related to the way probabilities are used. In our procedure, probabilities are summed, whereas in Naïve Bayes probabilities are multiplied. We found by trial and error that the fragment-based filter performs much better when rule fragment probabilities are

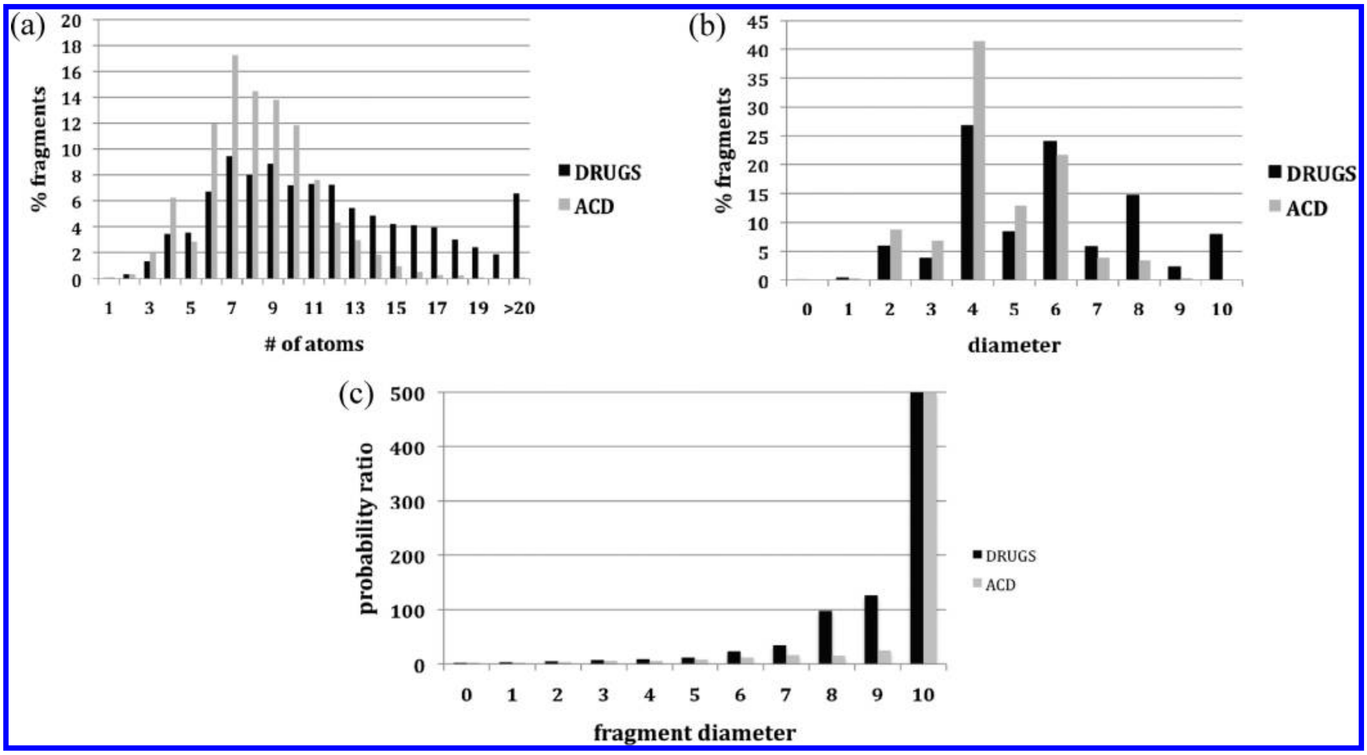

Figure 2. Distribution of properties in fragments selected for DLF, number of atoms (a), diameter calculated as maximum eccentricity (b), and fragment discrimination power calculated as p_{DRUGS}/p_{ACD} (c).

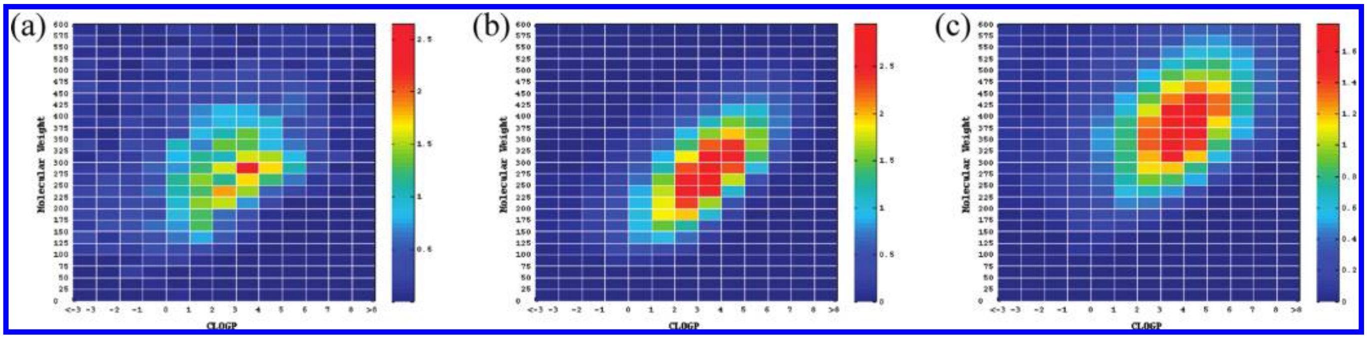

Figure 3. Distribution of MW vs CLogP for (a) DRUGS, (b) ACD, and (c) MDDR.

used in an additive, not multiplicative manner. The filter program is implemented in Java programming language with substructure matching based on the JChem API.

RESULTS AND DISCUSSION

Filter Performance. The DRUGS, ACD, and MDDR data sets were filtered via DLF using the procedure described in the Methods Section, with the following outcome: 87.05% of DRUGS, 39.65% of ACD, and 78.45% of MDDR structures passed. Within the DRUGS data set, 94 compounds (2.46%) have MW ≥ 1000; of these, 93 passed the DLF. From MDDR, 4544 compounds (2.68%) have MW ≥ 1000, of which 4510 passed the filter. These high MW compounds (e.g., cyclosporine) have a high content of drug-like fragments to pass DLF; however they represent only a small fraction of drug molecules. MW and the octanol/water partition coefficient, Log *P* (estimated, e.g., by CLogP)[36] are important properties for pharmacokinetics. We examined the distribution of these two properties in order to improve DLF performance.

The molecules from the DRUGS, ACD, and MDDR data sets exhibit different distributions for MW and CLogP (Figure 3). The 90th percentile for DRUGS was MW = 562.04 compared to 417.28 for ACD and 646.07 for MDDR, respectively. For CLogP, the 90th percentile for DRUGS was 5.48 compared to 5.55 (ACD) and 6.68 (MDDR), respectively. Based on these observations, most of the DRUGS have MW < 600, and because fragment occurrence is not size dependent, we decided to include the MW ≤ 600 rule (optional) to the DLF procedure. Given certain trends in combinatorial library design, it is not uncommon to find structures with MW > 600. By focusing on structures with MW ≤ 600, the user can focus on molecules that are more likely to be represented in the DRUGS data set. The source code for DLF (available in Supporting Information) gives the user the ability to disable this rule or to include additional rules based on other properties. Indeed, CLogP has been extensively used as a filter for drug-likeness. However, we aimed to evaluate the presence of chemical substructures in the chemical space defined by DRUGS, rather than use a property-based filter as a basis for distinguishing drugs from “nondrugs”.

Here, we observe that without using any threshold values for CLogP and MW, we filter a significant subset (40%) of

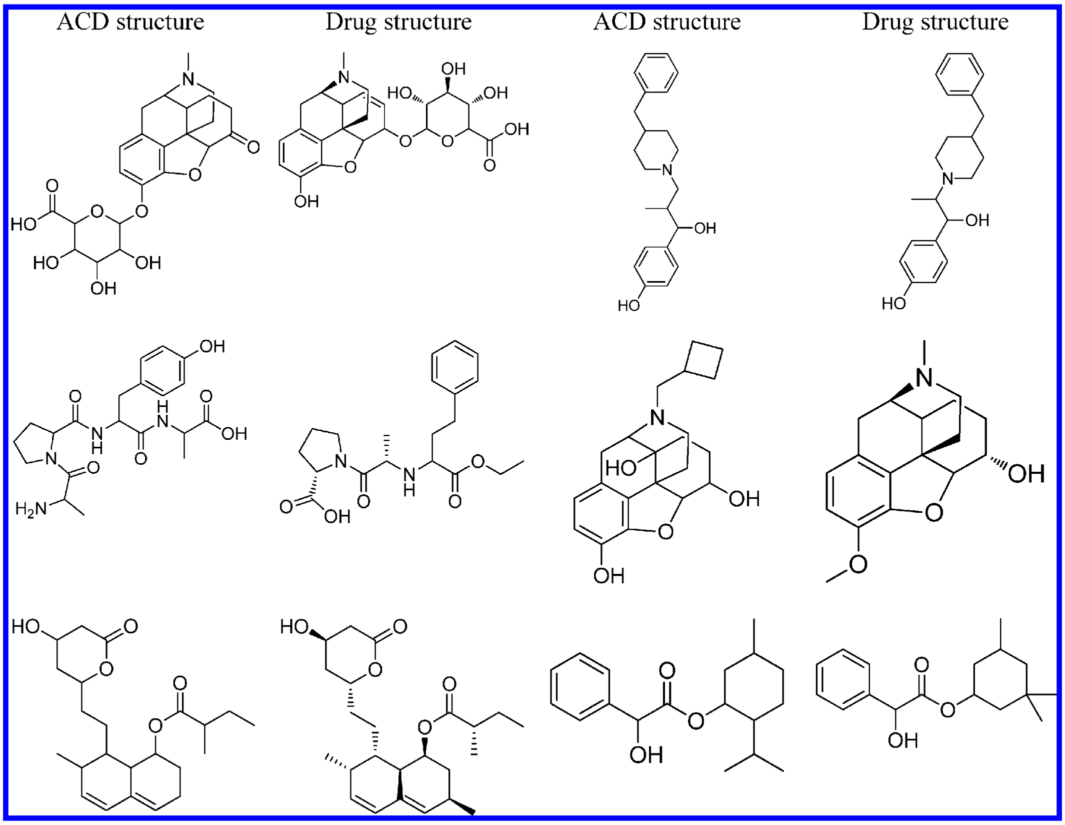

Figure 4. Compounds in ACD similar to approved drugs.

Table 1. Distribution of Molecular Properties in ACD Compounds Which Pass SMARTS Filter

molecular property	threshold	% ACD[a]	ACD[a] median value	% DRUGS	DRUGS median value	%MDDR	MDDR median value
molecular weight (MW)	≤500	98.50	278.24	86.42	318.52	71.61	423.46
partition coefficient (CLogP)	≤5	88.82	2.58	86.16	2.34	71.80	3.70
hydrogen-bond donor count (HDO)	≤5	99.29	1	92.81	1	93.02	2
hydrogen-bond acceptor count (HAC)	≤10	99.97	2	94.59	3	95.70	3
ring count (RNG)	1 ≤ RNG ≤ 4	87.91	2	82.29	2	75.12	4
rotatable-bond count (RTB)	2 ≤ RTB ≤ 8	84.85	4	67.38	5	56.98	7
rigid-bond count (RGB)	≥18	29.17	13	40.62	16	72.74	22
polar surface area (PSA)	≤120	94.06	60.22	79.91	68.74	75.69	82.28

[a] Fraction of ACD compounds which pass the SMARTS filter.

ACD that exhibits a distribution of molecular properties that is quite similar to that observed from drugs. After adding MW ≤ 600 as an additional DLF rule, 78.81% of DRUGS, 40.17% of ACD, and 65.64% of MDDR passed through (DLF + MW). The ~10% decrease for DRUGS and MDDR can be explained by the fact that there are 319 (8.34%) drug molecules with MW > 600, of which 315 (8.24%) actually passed. After adding the MW ≤ 600 criterion, these molecules fail the (DLF + MW). In the MDDR set, there are 23 037 (13.61%) molecules with MW > 600; out of these, 21 697 (12.82%) passed the DLF.

Drug-like ACD Compounds. Almost 40% of ACD structures pass the DLF, which proves that *ACD is far from perfect as a surrogate for nondrugs*. Indeed, there are many ACD compounds that have very similar structures to known drugs (Figure 4). The property distribution of certain descriptors used in estimating drug-likeness was examined for the ACD compounds that pass the DLF. This distribution (Table 1 and Figure 5) appears to be similar to thresholds observed and discussed elsewhere.[2,7,14,37,38]

The DLF-compliant ACD subset has a property distribution closer to that of DRUGS than the MDDR set. In comparing ACD with DRUGS, the larger positive differences are observed for PSA, RTB, and MW, where at least 10% more ACD compounds are inside those thresholds, while the largest negative difference is observed for RGB, where there are at least 10% more DRUGS inside the threshold. A likely reason for the RGB shift to lower values is that most of the ACD compounds have a lower MW (and lower complexity) compared to MDDR, which was used to set the threshold value. More than 94% of the ACD subset has PSA ≤ 120 Å^2 compared to DRUGS, where almost 80% of the molecules have PSA ≤ 120 Å^2.

Property distributions, combined with the chemical structural features for the DLF-compliant ACD data set indicate that this subset is more drug-like compared to that of ACD subset (60.35%) that fails the DLF. Drug-like molecular fragments have a higher occurrence rate in the DLF-compliant ACD subset compared to nondrug-like molecular fragments (Figure 6), which is consistent with a "drug-like"

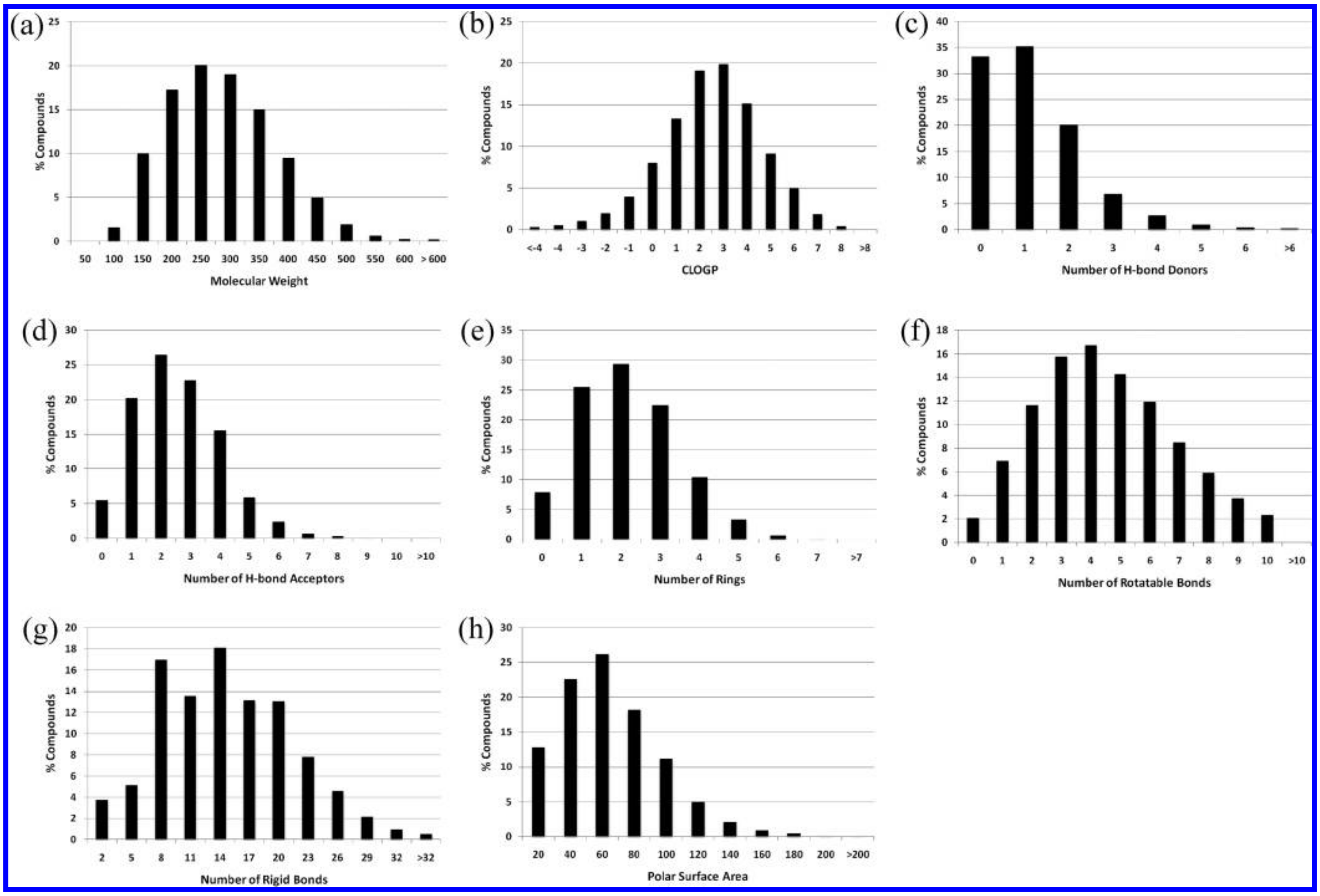

Figure 5. Distribution of MW (a), CLogP (b), HDO (c), HAC (d), RNG (e), RTB (f), RGB (g), and PSA (h) in the DLF-compliant ACD subset.

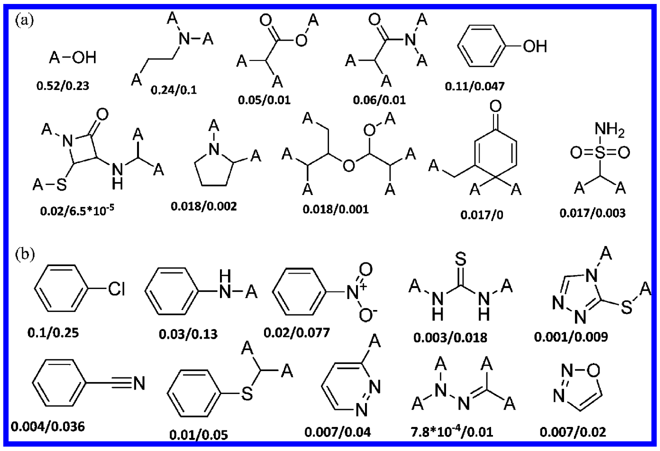

Figure 6. The top 10 (a) drug- and (b) nondrug-like molecular fragments. Each fragment is labeled with the occurrence rate in the DRUGS and ACD data sets.

property profile. The rationale for developing a model-free drug-like filter was exactly that: utilize as few assumptions as possible with respect to the perception of drugs vs nondrugs. Since the separation relies on the statistically derived evidence, i.e., fragments occurrence derived from the systematic exploration of chemical features in drugs, this empirical evidence is captured with a simple rule based DLF. When faced with a significant (almost 40%) overlap between two labels, i.e., ACD and DRUGS, with respect to chemical fragments as well as 2D molecular properties (based on Faulon's work), [31−33]kernel-based methods (SVM, neural networks) are likely to force the kernel function to adapt to noisy input data, which ultimately results in classifier models with lower external prediction accuracy, thus providing models for noise rather than signal. On the other hand a rule-based system can highlight the noise (overlap) present in both sets and provide insights on the significance of differences between the two data sets, by focusing on occurrence-based evidence. Rather than compressing the data, we eliminated most (1 402 652) of the automatically generated SMARTS, since we used only 1.13% of the total 1 418 622 potential descriptors.

Hydroxyl groups, amines, esters, amides, phenols, sulfonamides, and the conserved β-lactam are found predominantly present in drugs, as opposed to ACD. In an associative manner, these fragments combined to other such fragments contribute to the drug-like character of a chemical. Similar functional groups where discovered on several occasions to be predominant in drug molecules.[8,11,20] While we confirm these findings, it is no less important to note that these functional groups are also present in nondrug molecules, albeit with lesser occurrence rates compared to drugs. Although generally regarded as a collection of chemical reagents and as serving as the nondrugs reference set in cheminformatics papers, this version of ACD contains a wide array of chemicals that are, ultimately, added to the catalog in order to attract pharmaceutically oriented customers; thus this finding is far from surprising. We used the 2002 version of ACD as a nondrugs reference set not only for comparative purposes with earlier work (most references predate 2002) but also because newer updates of chemical catalogs have

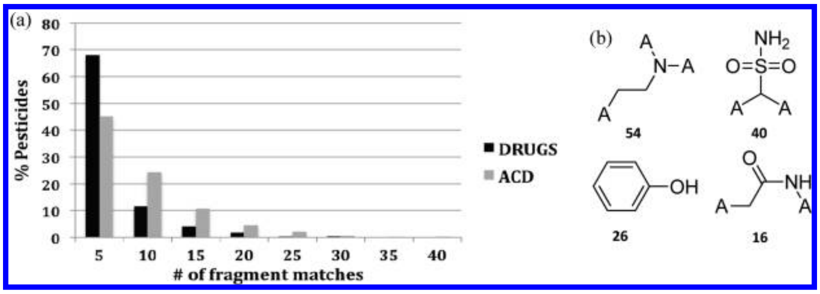

Figure 7. Distribution of DRUG/ACD-like fragments in pesticides (a) and common DRUG-like fragments (b) labeled with occurrence ratios in the pesticides data set.

deliberately shifted toward drug-like chemicals, although scaffold diversity is likely to be higher.

The inclusion of filters based on molecular fragments that are predominantly present in ACD compounds increases the sensitivity of DLF for chemical structures, where associative contributions of nondruglike fragments outweighs the contribution of drug-like fragments. Machine learning techniques, such as neural networks or kernel-based methods, perform better compared to rule-based methods when it comes to separating two label categories, i.e., drugs vs nondrugs. While the transformations applied during input data mapping to output values leads to machine learning decisions that can be understood by the experienced users, most of the drug-like classifiers are neither extensively nor convincingly validated using external data sets. Temporal validation indicates that predictive power can deteriorate in as little as four months,[39] which highlights the need for constantly updating machine learning models. This hinders in particular the utility of those drug-like classifiers that are built using third-party software because of the "black box" approach, where the model-selected criteria for discrimination are hidden from the end users. In contrast, rule-based approaches provide clear output results that can be interpreted more directly.

Pesticides Data Set. External validation is crucial in evaluating DLF performance. Having already demonstrated that DLF performs reasonably well on MDDR, a collection of molecules aimed at pharmaceutical use, we next evaluated model performance on a collection of 1482 pesticides.[40] Before filtering this list, we removed 331 chemicals that share a high similarity (Tanimoto >0.9) to chemical structures from DRUGS. From the resulting 1151 pesticides, $N = 607$ (52.74%) are DLF-compliant without using the MW ≤ 600 rule. The application of the DLF + MW (under 600) on this subset of 1151 pesticides yields 599 (52.04%) chemicals.

By examining the distribution of pesticides that contain at least one drug-like fragment (Figure 7a), we found that ~68% of pesticides contain between 1 and 7 drug-like fragments (from DRUGS); within the same range, only ~45% of pesticides match ACD-like fragments. This 22% difference in matched fragments is significant, as it indicates there are more DRUGS-related fragments than ACD-like fragments in this pesticides data set. It thus appears that the pesticides data set contains a rather significant proportion of chemicals with drug-like fragments (and implicitly 2D properties). The most commonly occurring drug-like fragments present in pesticides (Figure 7b) are also the most common fragments present in DRUGS (Figure 6a). The fragment size also plays an important role in determining the character ("label") of a chemical structure. Fragment statistics indicate that smaller DRUGS/ACD fragments (between 1 and 10 atoms) are much more likely to occur in the pesticides data set (Figure 8). Although DLF does not perform as well as one would intuitively expect, since pesticides are often regarded as nondrugs, the reason that 52% of pesticides are DLF-compliant is the same for almost 40% of ACD. Both sets of compounds share a high content of drug-like fragments that, in an additive manner, contribute to their similarity to drugs and their drug-like character. This is not surprising, since *most pesticides are designed to interact with biological targets*, which is generally how drugs work.

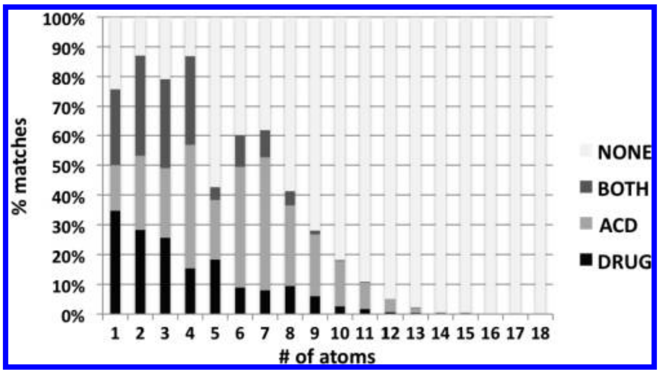

Figure 8. DRUGS/ACD fragment match distribution for most common fragment sizes in pesticides data set. Drug-like fragments are predominant at sizes 1−3 where they cover from 25−35% of pesticides, where is ACD-like fragment are more predominant at sizes 4−10.

CONCLUSIONS

A drug-likeness filter (DLF) was developed based on the occurrence ratios of molecular fragments in a collection of over 3800 drugs, compared to over 178 000 ACD chemicals. This resulted in 15 970 fragments, 11 016 from drug, and 4954 from ACD, respectively, which were encoded as SMARTS. These SMARTS contribute in an additive manner to the outcome of the filter process. After evaluating molecular weight (MW) and CLogP, we amended DLF with the "MW ≤ 600" criterion, in order to avoid large structures from being accepted. Using DLF and MW, 78.81% of DRUGS, 40.17% of ACD, and 65.64% of MDDR passed the filter. We also found that 52.04% out of 1482 pesticides are DLF-compliant. As significant subsets of the ACD and pesticide list pass DLF, we concluded that these subsets are more likely to be drug-like, as they predominantly contain fragments that occur more often in drugs. This observation highlights the danger of relying nondiscriminately on machine learning techniques that artificially separate drugs from nondrugs, in particular with respect to ACD. This is likely to negatively influence the usefulness of such classifiers, as 40% of the "negative label" (nondrugs) are similar to drugs and are more likely to have a drug-like character. By contrast, rule-based approaches do not rely on such assumptions. We use the term "model-free" to emphasize the fact that our method does not use kernel functions. Rather than fit input data using any function, we rely on occurrence ratios to derive a set of rules that is as simple as possible, serving to determine the probability of a compound to be drug-like. Naturally, any learning process relies on models, and as such, this is not a model-free system.

Developed as a simple filter, DLF is perhaps less useful in discriminating drugs from nondrugs, quite likely erring on the false positives side. DLF is however more likely to rapidly eliminate those chemicals rich in nondrug-like fragments. The DLF output is amenable to simple interpretation, and it can be easily adjusted to account for additions/modifications to the drugs/nondrugs data sets. The SMARTS-encoded fragments used by DLF are made freely available though Supporting Information and can used for library design and compound acquisition. Finally, a reliable benchmark for nondrugs is not likely to exist since it is increasingly more difficult to find large medicinal chemistry catalogs that are not biased toward existing drugs.

ACKNOWLEDGMENT

This study was supported by a NIH grant 5U54MH084690-02 (O.U. and T.I.O.) and by the University of New Mexico sabbatical leave program (T.I.O.).

Supporting Information Available: The molecular fragments encoded as SMARTS, annotated with probability values for each data set, are provided. The Java programs used for derivation of SMARTS queries and their usage on a user data set are presented. The web-based implementation of this service can be found at http://pasilla.health.unm.edu/tomcat/drug-likeness/. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Walters, W. P.; Murcko, M. A. Prediction of 'drug-likeness'. *Adv. Drug Delivery Rev.* **2002**, *54*, 255–271.

(2) 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.

(3) Ajay; Walters, W. P.; Murcko, M. A. Can We Learn To Distinguish between "Drug-like" and "Nondrug-like" Molecules? *J. Med. Chem.* **1998**, *41*, 3314–3324.

(4) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325–3329.

(5) Wang, J.; Ramnarayan, K. Toward designing drug-like libraries: A novel computational approach for prediction of drug feasibility of compounds. *J. Comb. Chem.* **1999**, *1*, 524–533.

(6) Frimurer, T. M.; Bywater, R.; Nrum, L.; Lauritsen, L. N.; Brunak, S. Improving the Odds in Discriminating "Drug-like" from "Non Drug-like" Compounds. *J. Chem. Inf. Model.* **2000**, *40*, 1315–1324.

(7) Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251–264.

(8) Wagener, M.; van Geerestein, V. J. Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features. *J. Chem. Inf. Model.* **2000**, *40*, 280–292.

(9) Xu, J.; Stevenson, J. Drug-like index: A new approach to measure drug-like compounds and their diversity. *J. Chem. Inf. Model.* **2000**, *40*, 1177–1187.

(10) Anzali, S.; Barnickel, G.; Cezanne, B.; Krug, M.; Filimonov, D.; Poroikov, V. Discriminating between drugs and nondrugs by prediction of activity spectra for substances (PASS). *J. Med. Chem.* **2001**, *44*, 2432–2437.

(11) Muegge, I.; Heald, S. L.; Brittelli, D. Simple selection criteria for drug-like chemical matter. *J. Med. Chem.* **2001**, *44*, 1841–1846.

(12) Brustle, M.; Beck, B.; Schindler, T.; King, W.; Mitchell, T.; Clark, T. Descriptors, physical properties, and drug-likeness. *J. Med. Chem.* **2002**, *45*, 3345–3355.

(13) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification. *J. Chem. Inf. Model.* **2003**, *43*, 1882–1889.

(14) Muegge, I. Selection criteria for drug-like compounds. *Med. Res. Rev.* **2003**, *23*, 302–321.

(15) Takaoka, Y.; Endo, Y.; Yamanobe, S.; Kakinuma, H.; Okubo, T.; Shimazaki, Y.; Ota, T.; Sumiya, S.; Yoshikawa, K. Development of a method for evaluating drug-likeness and ease of synthesis using a data set in which compounds are assigned scores based on chemists' intuition. *J. Chem. Inf. Model.* **2003**, *43*, 1269–1275.

(16) Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J. Chem. Inf. Model.* **2003**, *43*, 2048–2056.

(17) Muller, K.-R.; Ratsch, G.; Sonnenburg, S.; Mika, S.; Grimm, M.; Heinrich, N. Classifying 'Drug-likeness' with Kernel-Based Learning Methods. *J. Chem. Inf. Model.* **2005**, *45*, 249–253.

(18) Zheng, S. X.; Luo, X. M.; Chen, G.; Zhu, W. L.; Shen, J. H.; Chen, K. X.; Jiang, H. L. A new rapid and effective chemistry space filter in recognizing a druglike database. *J. Chem. Inf. Model.* **2005**, *45*, 856–862.

(19) Good, A. C.; Hermsmeier, M. A. Measuring CAMD Technique Performance. 2. How "Druglike" Are Drugs? Implications of Random Test Set Selection Exemplified Using Druglikeness Classification Models. *J. Chem. Inf. Model.* **2006**, *47*, 110–114.

(20) Li, Q. L.; Bender, A.; Pei, J. F.; Lai, L. H. A large descriptor set and a probabilistic kernel-based classifier significantly improve druglikeness classification. *J. Chem. Inf. Model.* **2007**, *47*, 1776–1786.

(21) Schneider, N.; Jackels, C.; Andres, C.; Hutter, M. C. Gradual in silico filtering for druglike substances. *J. Chem. Inf. Model.* **2008**, *48*, 613–628.

(22) Schierz, A.; King, R. Drugs and Drug-Like Compounds: Discriminating Approved Pharmaceuticals from Screening-Library Compounds. In *Pattern Recognition in Bioinformatics*, 1st ed.; Kadirkamanathan, V., Eds.; Springer-Verlag: Berlin, Heidelberg, Germany, 2009; pp 331−343.

(23) Rayan, A.; Marcus, D.; Goldblum, A. Predicting Oral Druglikeness by Iterative Stochastic Elimination. *J. Chem. Inf. Model.* **2010**, *50*, 437–445.

(24) Ursu, O.; Ryan, A.; Goldblum, A.; Oprea, T. I., Understanding drug-likeness. *WIREs Comp. Mol. Sci.* **2010**, submitted.

(25) Tetko, I. V.; Livingstone, D. J.; Luik, A. I. Neural network studies. 1. Comparison of overfitting and overtraining. *J. Chem. Inf. Model.* **1995**, *35*, 826–833.

(26) Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Model.* **2003**, *44*, 1–12.

(27) Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discov.* **1998**, *2*, 121–167.

(28) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naive Bayesian Classifier. *J. Chem. Inf. Model.* **2003**, *44*, 170–178.

(29) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Model.* **2004**, *44*, 1708–1718.

(30) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256–3266.

(31) Faulon, J.-L.; Visco, D. P.; Pophale, R. S. The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. *J. Chem. Inf. Model.* **2003**, *43*, 707–720.

(32) Faulon, J.-L.; Churchwell, C. J.; Visco, D. P. The Signature Molecular Descriptor. 2. Enumerating Molecules from Their Extended Valence Sequences. *J. Chem. Inf. Model.* **2003**, *43*, 721–734.

(33) Faulon, J.-L.; Collins, M. J.; Carr, R. D. The Signature Molecular Descriptor. 4. Canonizing Molecules Using Extended Valence Sequences. *J. Chem. Inf. Model.* **2004**, *44*, 427–436.

(34) JChem Base, version 5.3.1; ChemAxon: Budapest, Hungary, 2010.

(35) *SMARTS-A Language for Describing Molecular Patterns*; Daylight CIS Inc.: Aliso Viejo, CA; http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html. Accessed March 31, 2010.

(36) *CLOGP*, version 5.2; BioByte: Claremont, CA, 2009.
(37) Clark, D. E. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood-brain barrier penetration. *J. Pharm. Sci.* **1999**, *88* (8), 815−821.
(38) Clark, D. E. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. *J. Pharm. Sci.* **1999**, *88* (8), 807–814.
(39) Gavaghan, C.; Hasselgren-Arnby, C.; Blomberg, N.; Strandlund, G.; Boyer, S. Development, interpretation and temporal evaluation of a global QSAR of hERG electrophysiology screening data. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 189–206.
(40) Wood, A. Compendium of Pesticide Common Names; http://www.alanwood.net/pesticides. Accessed March 31, 2010.

CI100202P