A Simple Approach for Indexing the Oral Druglikeness of a Compound: Discriminating Druglike Compounds from Nondruglike Ones

Dhrubajyoti Biswas, Sujata Roy, and Srikanta Sen*

Chembiotek Research International, Biology Lab, Tower-B, Bengal Intelligent Park Ltd., Block EP & GP, Sector-V, Salt Lake, Calcutta-700091, India

Received October 19, 2005

A knowledge-based simple score has been developed for indexing the oral druglikeness of compounds based on the concept that oral druglikeness should be independent of the drug targets and, thus, are closely related to the global absorption, distribution, metabolism, and excretion related properties. We have considered several simple molecular descriptors as the key determinants of druglikeness. The patterns of the distributions of these molecular descriptors for a set of drug molecules have been extracted using a nonlinear neural network method. We assumed direct correlations of these patterns to the expectation values that a given compound may behave like a drug. On the basis of this assumption, we have defined a simple druglike index or score (DLS) combining the contributions coming from the descriptors considered. This index scales the druglikeness of a compound in the range 0.0−1.0, 1.0 being the highest druglikeness. The index applied for a drug data set, a mixed data set, and three different bioactive databases produced expected features and indicated that even the marketed drugs have druglike scores varying over a considerable range. A total of 73.3% of the drugs considered showed DLS > 0.5, while it is only 44.7% for the HIC-Up compounds (unbiased ligand database). For the ChemBank, Asinex-Gold collection, and NCI databases 61.2%, 76.0%, and 79.1% of the compounds have DLS > 0.5.

INTRODUCTION

Modern drug discovery philosophy is largely based on the high-throughput screening (HTS) of large compound libraries either taken from the existing compound databases or obtained by chemical synthesis by combinatorial chemistry techniques. However, considering the huge size of the chemical space involved, such an approach appears to be unnecessarily expensive and time-consuming if we are not guided by some rules to select the right kind (druglike) of subset of the library. According to a detailed examination of the transition of the compounds from HTS hits to the market, it appears that the probability of reaching the market is one in a million.[1] Therefore, choosing the lead structures with appropriate qualities has become an important focus in preclinical drug research. Thus, primarily, a screening library should be designed such that the screening hit rate is improved compared to that of a random library.

Drug molecules generally act on specific targets at the cellular level and, upon binding to the receptors, exert a desirable alteration of the cellular activities that is regarded as the pharmaceutical effects. Before interacting with the receptor, the drug molecule must travel from the entry point through the body to reach the site of drug action, and after executing its activity, it should steadily be eliminated from the body. The entire process can be broadly described in terms of a few subprocesses such as absorption, distribution, metabolism, excretion, and toxicity (ADMET). One of the major objectives of the present work is to take care of the relevant ADMET issues in order to quantify the possibility that an orally taken compound will reach the target location. Among the ADMET issues, absorption (A), distribution (D), and excretion (E) are relevant to the journey of the compound from the application point to the site of action and also to its elimination from the system. These three issues are also of similar nature and are, thus, dependent on similar descriptors. The other two issues, metabolism (M) and toxicity (T), are of a quite different nature and should depend on the chemical reactivity and electronic features and so forth. The issues M and T should be considered separately in the next step.

A molecule having characteristics similar to those of the known bioactive compounds is referred to as "druglike". In general, the quality of a drug is determined by two factors, (i) the affinity of the compounds toward a specific target molecule and (ii) the ADMET-related properties. The first factor determines the specificity of the compound toward a desired target, and the second factor determines its overall (independent of the specific target) interactions with the biosystem. Thus, for developing a "target-independent" general type screening library, it is desirable to prescreen druglike compounds, and in silico methods can be quite valuable in this respect.

Efforts have been made to define "rules" to discard compounds in databases that are not druglike.[2−13] These include methods using the characteristics of known drug compounds in terms of molecular descriptors such as molecular weight, log *P*, the number of hydrogen-bond donors, the number of hydrogen-bond acceptors, the number of rotatable bonds, the number of rigid bonds, the number of rings in a molecule, and so forth and neural network or other statistical techniques. Lipinski and co-workers[3] have described perhaps the most widely used set of counting rules

* Corresponding author e-mail: srikanta@chembiotek.com.

10.1021/ci050459i CCC: $33.50
Published on Web 01/21/2006

to filter out compounds likely to show poor absorption properties. Simple counting methods based on functional group filters have also been used in exploring the chemical space of known drugs. The presence of some functional groups (like amine, amide, alcohol, sulfone, sulfonamide, etc.) known to have H-bonding capabilities are generally observed in known drugs and are, thus, considered as pharmacophoric points. The number of such pharmacophoric points present in a compound is used as a measure of its druglikeness. The problem with this approach is that the number of such chemical groups covering the entire druglike space may be quite large, and also, a larger number of pharmacophore points in a compound may also have non-druglike effects.

Lipinski's criteria are useful for separating druglike compounds from nondruglike ones, and given two druglike compounds, they cannot quantitatively predict which one is more druglike. A druglike index like the present one can quantify the druglikeness of a compound. Compared to most of the other methods[2−13] suggesting druglike indices, our approach differs in regard to two major issues. The first issue indicates that, while others use many kinds of structural and other descriptors, we use descriptors of a global nature only, and none of these descriptors are directly associated with any specific structural components but have direct physical correlation to the absorption, distribution, and excretion aspects of ADME issues. The second issue is related to the fact that, in other works, bin or histogram methods were used to describe the distribution pattern, and thus, they suffer from the fine structural features that are mostly artifacts due to the over- or under-representation of the data set in that range. In contrast, we have extracted a global pattern of the distribution data avoiding any fine structure and, thus, have minimized the dependence of the results on the specific training data set used.

Here, we have presented a simple method to quantify the druglikeness of a given compound through a druglike "score" (DLS) or "index". The approach is knowledge-based and relies on the concept that the drug compounds, in general, follow a specific distribution pattern for each of a set of molecular descriptors. We have used a neural network method to derive a relationship describing the global distribution pattern for each relevant descriptor. We then combined those relations to compute a "score" correlating to the fitness of the compound as a drug. In the present work, we have chosen descriptors to take care of the three ADMET issues, A, D, and E, in defining the druglikeness score or index. The present score thus represents the druglikeness of a compound on the basis of these three issues. We have applied this method to the assessment of several data sets consisting of only known drugs and other bioactive compounds, and the results were found to be quite encouraging.

METHODS

The present approach is based on the fact that we have data for the distribution of each property (like mw, log *P*, etc.) over a range (bin) of values of the respective property, and it only represents the cumulative data over the range. So, the variation of the property over the range of each bin is missing. Here, we have extracted a simple mathematical expression for the global distribution pattern for each descriptor and have used that global pattern to generate a local pattern even within the bin range depending on the global profile of the property. In recognizing the hidden global pattern, we assumed that there was no fine structure; that is, the dependence is smooth. We further assume that very few drugs are available for compounds having a property with a value that corresponds to a low fraction in the respective distribution curve, because that value of the property makes a compound less druglike. Similarly, a compound with a value close to the peak of the distribution indicates that many drugs are available with this value and, hence, makes a compound more druglike. Thus, the distribution pattern is directly correlated to the expectation value that a compound with a specific value of the respective descriptor will behave as a druglike one.

Choice of the Descriptors. It may be emphasized at the beginning that, in representing a compound, the use of a large number of descriptors may lead to overdescription of the system. Therefore, in our case, we have tried to be limited in diverse and at least partially independent descriptors that can be easily computed from the 3D structure of a given compound. To select the descriptors set, we paid more attention to their physicochemical roles in determining their overall ADME-related properties as we are looking for target-independent druglike features. The factors in Lipinski's rule of five and a few more seem to be a good starting point. Thus, our initial descriptor set consists of molecular weight (mw), the octanol−water partitioning coefficient (log $P_{o/w}$), the polar surface area (psa), the number of H-bond donors (nhbd), the number of H-bond acceptors (nhba), and the number of rotatable bonds (nrb). From physical considerations, it appears that the present set of descriptors captures all the essential features of A, D, and E issues. For example, mw takes care of the size dependence of the compound while log *P* takes care of the balance between the hydrophobicity and hydrophilicity of the compound. The polar surface area along with log *P* takes care of the aqueous solubility. The number of H-bond acceptors and donors also takes care of the potential of forming H bonds. The number of rotatable bonds represents the flexibility and is partially correlated to the size of the compound. All these aspects are important for absorption, distribution, and excretion and seem to cover the essential aspects from a physical point of view. Moreover, each of these quantities exhibits very well-defined distribution patterns for a large collection of oral drugs.[14−17]

Computation of the Chosen Descriptors. For training purposes, we used the data on the distribution of different descriptors directly available from the published work of Vieth et al.[17] For the test compound sets, we used our in-house-developed computer codes to compute the descriptors directly from their coordinates. All these descriptors have been computed directly from the Protein Data Bank (PDB) code of the respective compound using the code we have generated. The molecular weights were calculated by adding the atomic weights of the atomic composition of the compound; nhbd, nhba, and nrb were directly counted from the PDB on the basis of the atom type, and the descriptor psa was calculated by a Monte Carlo method. In computing psa, all the N, O, and polar H atoms have been considered. Finally, the log $P_{o/w}$ values for the compounds of the test data sets were calculated using our own model developed on the basis of a neural network.

Training by Neural Network Technique. The distribution data of each descriptor for the drug compounds were trained by a nonlinear neural network with a (1−2−1) architecture using a single hidden layer.[18,19] The tanh function was used as the activation function.[18] We assumed that the distribution patterns smoothly depended on the descriptors without any fine structure.

Let X represent any of the descriptors in the descriptor set (mw, log P, psa, nhbd, nhba, and nrb) and f_X represent the fraction of compounds, in percent of the drug population in the drug database, having descriptor X equal to a given value. Now, for use in training by a nonlinear neural network, we have preprocessed the data by using the following relations

$$X' = (X - \bar{X})/\sigma_X \tag{1a}$$

$$f_{X'} = (f_X - \bar{f_X})/\sigma_{f_X} \tag{1b}$$

where X' and $f_{X'}$ represent the respective quantities for the processed or reduced values. $\bar{X}$ and $\bar{f_X}$ represent the mean values of the respective quantities over the entire data set, and σ_X and σ_{fx} respectively represent the root-mean-square deviations of them. Preprocessing of the data was required to ensure that the input data would not be concentrated in the asymptotic regions of the activation function. Training was continued until the mean squared error (MSE) converged. MSE is defined as

$$\mathrm{MSE} = \frac{1}{N}\sum_{\{X'\}}(f^c_{X'} - f^t_{X'})^2 \tag{2}$$

where N is the number of data points in the data set, $f^c_{X'}$ is the computed value, and $f^t_{X'}$ is the target value of the fraction. In the training, a learning rate of 0.005 was used.

The neural network used here represented a relation connecting $f_{X'}$ to X as

$$f_{X'} = \omega_1 \tanh(\omega_{a1}X' + \omega_{a2}) + \omega_2 \tanh(\omega_{b1}X' + \omega_{b2}) + \omega_3 \tag{3}$$

Training provides the values of the weights (ω_1, ω_2, ω_3, ω_{a1}, ω_{a2}, ω_{b1}, and ω_{b2}) corresponding to the converged value of the MSE. As this relation is obtained for the preprocessed data, we need to revert back to the preprocessing method to get the relationship applicable to the original data set, and this is done by the relation

$$f_X = f_{X'}\sigma_{fX} + \bar{f_X} \tag{4}$$

Each training curve has a maximum ($f_X^{\max}$) in its distribution pattern. To make the scale the same for all the different f_X values, we need to normalize each of them by the respective $f_X^{\max}$.

DEFINITON OF THE DRUGLIKE SCORE (DLS)

The f_X value actually indicates the expectation that a randomly picked compound from the drug set will have a value of the respective quantity close to X. Here, we interpret f_X as the measure of the expectation that the compound will behave like a drug. Note that it is not a probability from a mathematical point of view but is consistent with the nature of the distribution patterns. We have generalized this correlation by considering f_X as a measure of the expectation that the compound with X will be a drug.

The DLS may then be defined as

$$\mathrm{DLS} = \frac{1}{N_X}\sum_{\{X\}}\frac{f_X}{f_X^{\max}} \tag{5}$$

where N_X is the number of descriptors (X) used.

Thus, the definition of DLS indicates that its value is bounded in the range 0−1 and implies that the most druglike compound will have a value of 1.0. For discriminating druglike compounds from the nondruglike ones, one can define a cutoff value of DLS, $\mathrm{DLS}_{\mathrm{cut}}$, such that a compound with $\mathrm{DLS} < \mathrm{DLS}_{\mathrm{cut}}$ should not be regarded as a drug.

RESULTS

Figure 1 represents the training of the input data sets obtained by a nonlinear neural network using the tanh(x) activation function. In all cases, the training was converged to values MSE < 0.05. It is clearly seen that, in each case, the fitted curve, according to the training, represents the global nature of the distribution pattern quite well as is expected from the respective MSE values. As we assumed the distribution patterns to be smooth without any fine structure that would appear artificially due to a lack of adequate data in that range, we neglected one or two data points in some cases to remove fine structures in the respective distribution pattern. In the training of the distribution pattern of each descriptor, we recorded the final values of the weights and the quantities used for preprocessing the input data. These data were used to compute the DLS value of a given compound through the relations 1−5.

To assess the ability of DLS to describe the drugs in a proper way, we computed the DLS values for 100 known marketed drugs. However, seven of them were excluded from the data set as these could not be modeled quantum chemically due to their sizes, and finally, 93 compounds were considered (Table 1). The nature of the distribution of the DLS scores for known drug compounds is demonstrated in Figure 2. As expected, it is clearly seen that the peak appeared toward the highest possible value (1.0) of DLS. The spread of the distribution curve also indicated that even all the marketed drugs are not of very similar quality in terms of their druglikeness. This is also consistent with the facts that the required doses of different drugs are quite diverse and, in some cases, formulation is required to improve its applicability. It is further noticeable that some of the drugs have produced considerably poor DLS values. Two of them are found to be not for oral application. The low scores for the other drugs may be due to fact that a metabolite of the original compound is active[20] or due to the inherent statistical uncertainty of the method. However, the overall prediction quality indicates the reasonability of the present score scheme of druglikeliness. Table 1 summarizes the DLS values for the 93 known drugs considered here for validation of the score.

As a next step of validation of the present method, one needs to demonstrate that, for a set of known nondrugs, DLS produces low scores. Unfortunately we have no access to

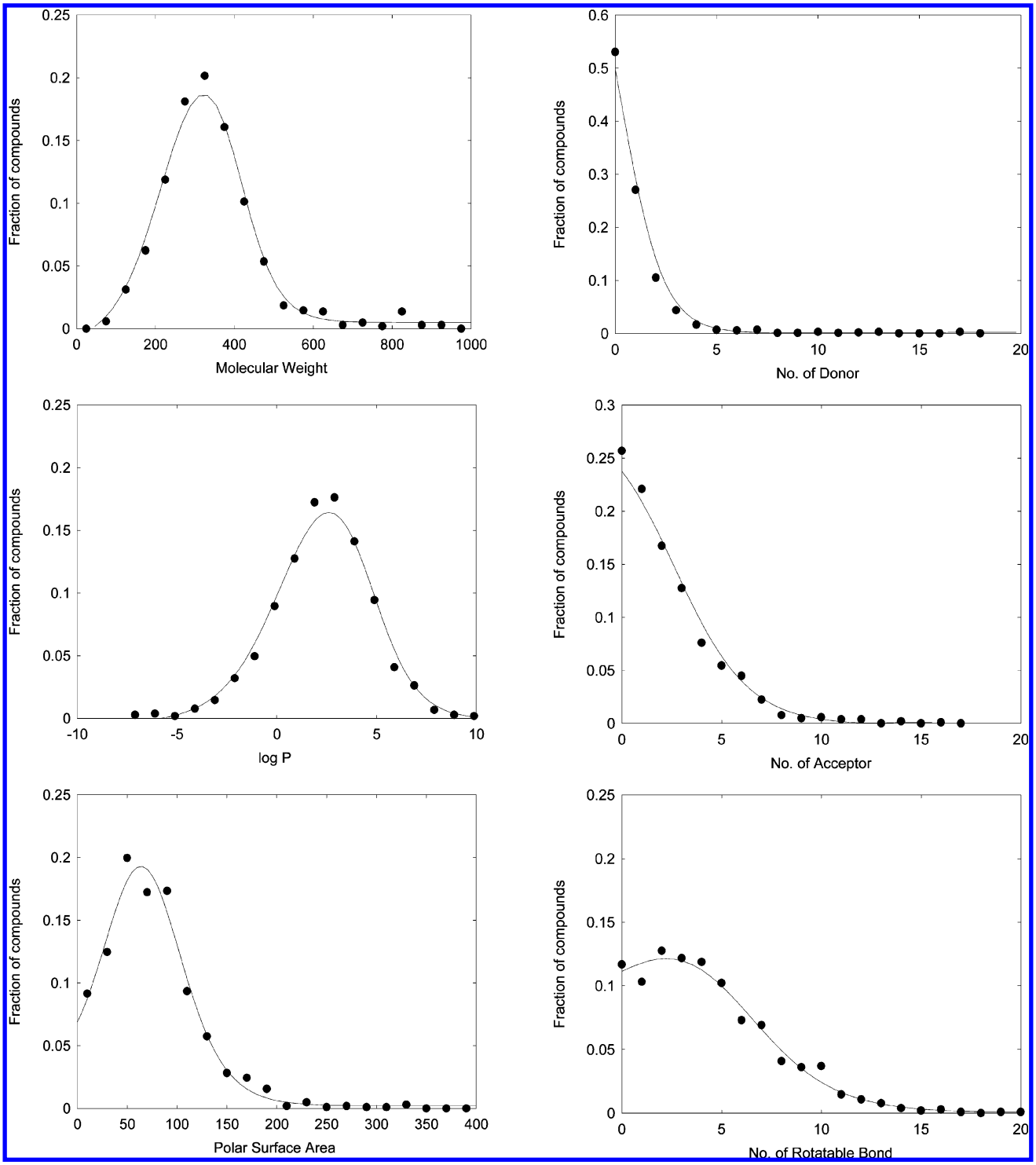

Figure 1. Plots of the distribution patterns of molecular weight (mw; Dalton), log *P*, polar surface area (psa; Å^2), number of H-bond donors, number of H-bond acceptors, and number of rotatable bonds in the drug database (Vieth et al. *J. Med. Chem.* **2004**, *47*, 224−232).[17] The symbols represent the bin-based values obtained from the database analysis, and the solid lines represent the training curve describing the global patterns.

any nondruglike database. So, we considered the HIC-Up database that contains all the ligands bound to their protein counterparts in the protein−ligand complexes for which the crystal structures have been solved.[21] As there is no preassigned property condition, these compounds can be drugs or druglike and nondruglike. For comparison purposes, we randomly selected 94 compounds from this database to keep the sizes of the compound data sets similar and computed their DLS values. Figure 2 compares the distribution of the DLS values for this data set and the drug data set. The distribution patterns appeared quite different, and it is clearly seen that, for HIC-Up compounds, the distribution is rather flat compared to that of the drug data set. The computed DLS values for the 94 HIC-Up compounds are presented in Table 2.

Figure 2 also indicates that 0.5 is a reasonable choice for DLS_{cut}, and compounds having DLS < DLS_{cut} are expected to have little potential as drugs. With this cutoff value, the

Table 1. Names and Computed DLS Values of 93 Known Oral Drugs

Sl number	compound name	DLS	Sl number	compound name	DLS
1	streptomycin[a] (injection)	0.01	48	trinitroglycerin	0.39
2	abacavir sulfate	0.55	49	tuberin	0.67
3	N-4-hydroxyphenyl-acetamide	0.65	50	valium	0.85
4	acetylsalicylic acid	0.69	51	xanthocillin	0.70
5	aciclovir	0.39	52	zanamivir	0.20
6	albendazole	0.63	53	dexamethasone	0.55
7	albuterol	0.51	54	dinitrochlorobenzene	0.54
8	allopurinol	0.53	55	dronabinol	0.65
9	alprazolam	0.87	56	erythromycin	0.16
10	altace/remipril	0.48	57	fiacitabine	0.46
11	amoxicillin	0.36	58	fialuridine	0.49
12	atenolol	0.51	59	fluconazole	0.68
13	azathioprin	0.46	60	ganciclovir	0.27
14	benzocaine	0.69	61	guaifenesin	0.58
15	caffine	0.73	62	ifosfamide	0.77
16	cefaclor	0.51	63	isotretinoin	0.64
17	cimetadine	0.44	64	lamivudine	0.58
18	clozapine	0.77	65	methadone	0.69
19	codeine	0.77	66	ampicillin	0.46
20	dextromethophan	0.81	67	methylprednisolone	0.57
21	dicoumarol	0.70	68	nandrolone decanoate	0.58
22	dimetapp	0.75	69	nystatin[a] (topical)	0.04
23	diphenhydramine	0.70	70	zalcitabine	0.54
24	droperidol	0.70	71	valacyclovir	0.31
25	enalapril	0.62	72	stavudine	0.57
26	fluroxene	0.73	73	saquinavir	0.18
27	haloperidol	0.67	74	prednisone	0.65
28	hydrochlorothiazide	0.45	75	amitriptyline	0.75
29	ibuprofen	0.67	76	neflinavir	0.32
30	indinavir	0.28	77	levofloxacin	0.71
31	mercaptopurine	0.66	78	oxazepam	0.76
32	molindone	0.76	79	valproic acid	0.62
33	naproxen	0.74	80	sorivudine	0.41
34	nifedipine	0.57	81	aspirin	0.68
35	penicillin G	0.62	82	barbiturate	0.73
36	pentaerythritol nitrate	0.37	83	nicotine	0.76
37	piracetam	0.65	84	THC	0.65
38	procainamide	0.60	85	amino salicylic acid	0.55
39	promethazine	0.76	86	ciprofloxacin	0.70
40	prontosil	0.44	87	didanosine	0.55
41	prozac	0.67	88	imipramine	0.76
42	quinine	0.72	89	clotrimazole	0.79
43	ranitidine	0.50	90	nevirapine	0.77
44	reserpine	0.35	91	glutamic acid	0.37
45	sulfanilamide	0.57	92	simvastine	0.55
46	tenormin	0.50	93	megestrol	0.72
47	terfenadine	0.40			

[a] Nonoral (intravenous or topical application) drugs.

drug data set was found to contain 73.3% druglike compounds, while for the HIC-Up data set, it was only 44.7%. The 26.7% (a total of 24 drugs in the present case) of drugs appearing to be nondruglike on the basis of the present druglike scoring scheme were identified, and when examined further, 2 out of 24 were found to be drugs for nonoral (injection or topical) application (footnotes of Table 1). Moreover, 6 of the residual 22 drugs with DLS $<$ DLS_{cut} were found to have very high molecular weights (Table 1). For a further assessment of the reasonability of DLS, a few known drugs for nonoral application were considered. The computed DLS values were found to be considerably low and are thus consistent with their nature of administration (Table 3).

Finally, we tried to check the performance of DLS for the predefined druglike databases such as ChemBank,[22] the Asinex-Gold collection,[23] and the NCI database.[24] The

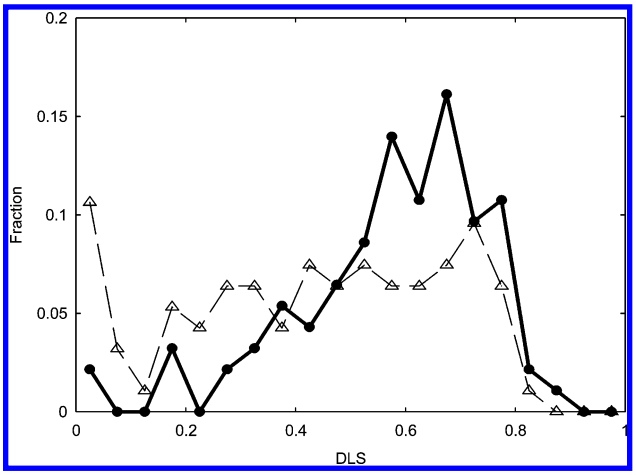

Figure 2. Comparison of the distribution patterns of the fraction of drugs vs the druglike score (DLS) for a library of known drugs (——, ●) and an equal number of compounds selected randomly from the HIC-Up database (— —, △).

Table 2. A Few Known Nonoral Drugs and Their Computed DLS Values Indicating Their Poor Quality as an Oral Drug

Sl number	compound name	DLS
1	cytarabine (injection)	0.34
2	daunorubicin (injection)	0.17
3	amphotericin B (injection)	0.33
4	vinblastine (injection)	0.20
5	ceftriaxone (injection)	0.10
6	taxol (injection)	0.18
7	vidarabine (ophthalmic ointment)	0.35
8	zanamivir (oral inhalation)	0.20

distribution of the compounds in terms of the computed DLS values for 4292 compounds in the ChemBank database, 7965 representative compounds of the Asinex-Gold collection, and also 1000 compounds from the NCI database clearly indicated (Figure 3) that each of the databases contains a good fraction of druglike compounds. However, the peak height and its positions further pointed out that the Asinex and NCI databases contain more high-quality druglike compounds compared to the ChemBank database. The use of $DLS_{cut} = 0.5$ indicates that ChemBank contains 61.2% druglike compounds while Asinex-Gold contains 76.0% druglike compounds and the NCI database contains 79.1% druglike compounds. Considering the fact that both Asinex and ChemBank are generated to be preferentially druglike, while HIC-Up contains significant amount of nondrugs, the predicted values appear quite realistic.

The comparison of our results with the published works on druglikeness was not very straightforward as, in most cases, people used separate special known databases for drug compounds and nondrug compounds to which we have no access. However, the results on a 93 known drugs set and an equal number of HIC-Up compounds showed that our method exhibited better discrimination compared to some published work[4] where the databases WDI and SPRESI were used as a drug database and a nondrug database, respectively, and a large overlap in the distributions of these two databases against their druglike index was observed.

In another work,[5] it was observed that a large fraction of the drug molecules (WDI) showed a score of 1.0, indicating that many drugs are perfect drugs. In contrast, in our case,

Table 3. HIC-Up Codes and Computed DLS Values for 94 Randomly Selected Compounds from the HIC-Up Database

Sl number	HIC-UP ID	DLS	Sl number	HIC-UP ID	DLS
1	14W	0.57	48	OES	0.64
2	1PE	0.50	49	NON	0.65
3	3BT	0.51	50	MST	0.55
4	4PN	0.70	51	NTS	0.63
5	ACZ	0.59	52	QNC	0.78
6	AR4	0.41	53	NAQ	0.03
7	BDD	0.62	54	NHO	0.01
8	BP3	0.74	55	2CP	0.02
9	C5X	0.46	56	AEB	0.06
10	CCB	0.76	57	CNE	0.21
11	CDG	0.49	58	CXT	0.17
12	CNS	0.53	59	CYC	0.21
13	DAT	0.27	60	DSX	0.32
14	TMD	0.53	61	EDE	0.01
15	DEL	0.23	62	EPX	0.14
16	DHM	0.58	63	GBX	0.17
17	DP1	0.20	64	HMG	0.01
18	DQH	0.72	65	KEU	0.10
19	FFC	0.29	66	NMY	0.01
20	FMB	0.36	67	NOV	0.16
21	FRA	0.42	68	PFG	0.08
22	G2S	0.26	69	PTY	0.17
23	GNR	0.60	70	R13	0.68
24	RCA	0.37	71	SAQ	0.51
25	PIC	0.79	72	TMS	0.01
26	MIL	0.74	73	VDN	0.41
27	TCC	0.75	74	XMB	0.29
28	NDD	0.82	75	C4X	0.46
29	OMN	0.78	76	CDD	0.43
30	SDS	0.70	77	CSF	0.01
31	TDG	0.21	78	CMB	0.52
32	MUP	0.71	79	DEO	0.47
33	REY	0.32	80	DOF	0.32
34	MUT	0.27	81	HAL	0.31
35	IM2	0.38	82	HPI	0.28
36	LY1	0.74	83	LAM	0.02
37	DBS	0.38	84	LG2	0.73
38	TBI	0.55	85	LS3	0.65
39	IBM	0.67	86	MDL	0.34
40	IPB	0.68	87	MTT	0.01
41	PBN	0.69	88	MUG	0.41
42	POO	0.68	89	THK	0.54
43	THS	0.47	90	TQ3	0.73
44	MCY	0.43	91	TR1	0.40
45	IMC	0.35	92	TTB	0.71
46	LEO	0.76	93	YPA	0.61
47	M1C	0.46	94	ZTW	0.76

we did not get anything above 0.9, and the fraction we obtained in the range above 0.8 is considerably small. This clearly indicated that our results are more realistic compared to their results. However, to make a more comprehensive comparison with these works, we need to use the same databases, which we could not do as a result of a lack of access to them.

To check if DLS can preferentially enrich a library sorted according to the DLS values of the compounds, we mixed the 93 drugs and the 94 compounds from the HIC-Up database together and sorted the compounds of the library according to the descending order of the computed DLS values. It is quite clear from Figure 4 that reasonable enrichment was achieved. In the best case, when the top 50% could contain only the drug compounds (100%), we obtained 63.2% drugs. The remaining 36.8% of the compounds came from the HIC-Up compounds, and there is no guarantee that all these HIC-Up compounds are really nondruglike.

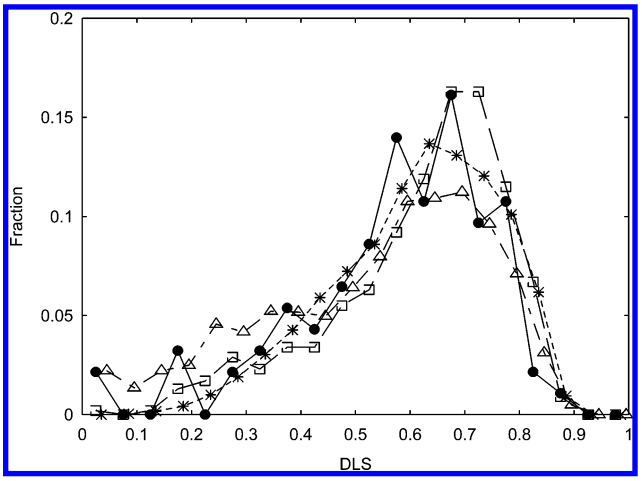

Figure 3. Comparison of the distribution pattern of the fraction of compounds vs the druglike score (DLS) for 93 drugs (——, ●), 4292 compounds from the ChemBank database (— — —, △), 7965 compounds from the Asinex-Gold database (- - - - - -, ∗), and 1000 compound from the NCI database (— —, □). In each database, a considerable to large fraction of druglike compounds are identified.

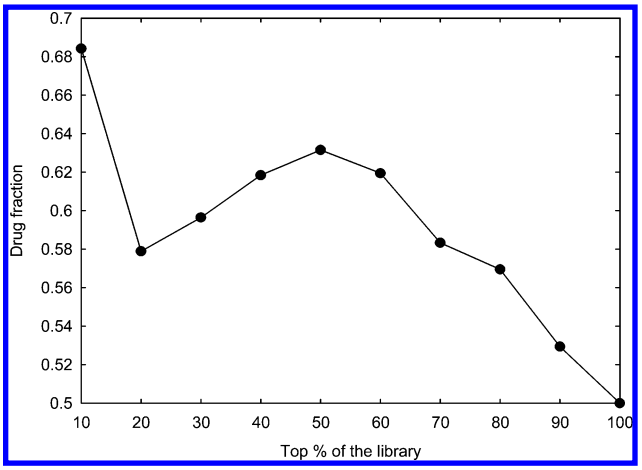

Figure 4. Demonstration of the content of known drugs in the top *x*% of the list sorted according to their DLS values for a mixture of 93 known drugs and 94 randomly collected compounds from the HIC-Up database.

Figure 5 demonstrates the distribution of the DLS values computed for the different variants around a core compound generated by adding two fragments each time from a library. It is seen that, even if the original compound has a low DLS score (0.72), still a few of them may have significantly high DLS values. Similarly, the converse is also true; a core compound having a high DLS value may have variants with significantly low DLS values. This may be quite helpful in selecting core compounds and the fragments to be used in generating druglike compound in chemical synthesis by combinatorial chemistry.

DISCUSSIONS

The present work suggested a DLS for indexing a set of compounds in order to prepare a potential druglike compound library. DLS is based on the properties of existing drugs, and the marketed drugs must have passed many development and safety barriers before reaching the market. DLS provides a global measure for druglike compounds, and hence, two compounds having similar DLS values may have quite

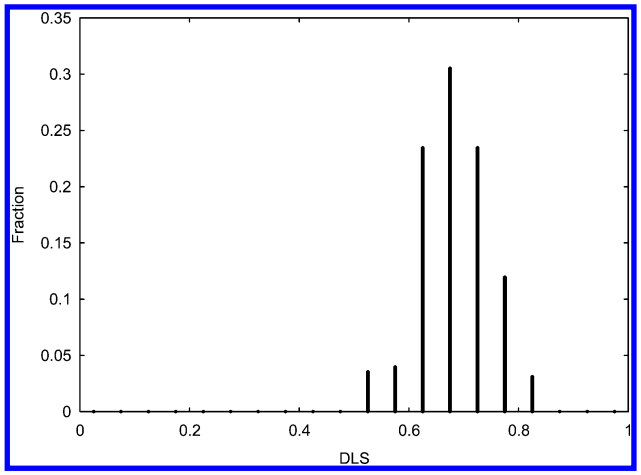

Figure 5. Distribution of variants generated around a given core by adding druglike fragments to it against the computed DLS values. The original core compound has a DLS score of 0.72. The use of 15 fragments has generated variants that are mostly druglike with DLS values varying in the range 0.5−0.85.

different topological scaffolds and even different ADME properties but within the acceptable limits.

We have considered only a small descriptor set chosen on the basis of physicochemical reasoning and have achieved good results. We have tried to be limited in diverse and at least partially independent descriptors that can be easily computed from the 3D structure of a given compound and are free from overdescribing the system.

In nonlinear neural network methods, overfitting appears because of the consideration of many nonlinear units in the hidden layer. This is because a significantly large number of hidden units can generate any kind of complex pattern, and that is at the root of its excellent performance in pattern fitting. In extracting the global pattern of the distribution curve, we have used only a single descriptor as input in each case and have used only two nonlinear units in the hidden layer. It may be pointed out that having two units in the hidden layer is minimal as consideration of only one being too restricted is not useful. This clearly indicates that the possibility of overfitting is least in the present case.

In the present work, the scoring function is defined in a way so that it takes into account the individual contributions coming from all the descriptors. It could be optimized by assigning weights to the individual contributions provided some experimental or other index representing the druglikeness was available that could be used as the target value. However, due to the lack of the existence of such an index, optimization of the score was not possible, and we have used unit weight factors for each contribution.

However, statistical methods rely on "collective features", and thus, the correlations drawn from such an analysis are more valid collectively for the data set than the individual members of the data set. Thus, despite the apparent success of any statistical method, the predictions based on statistical analysis should be interpreted carefully.

We modeled a small number (93) of known drugs and 94 HIC-Up compounds individually and used them to test the discriminating capability of our method. As we have no access to any drug or nondrug database (like ACD), we could not examine the discriminating power of this score in further detail.

The method is computationally fast enough. The actual computation for 1000 compounds, on average, takes only 4 min where the input is merely the3D coordinates of the compounds.

CONCLUSIONS

The DLS suggested in the present manuscript is a knowledge-based measure of the likeliness that a compound behaves like a drug and appears to be a good index for expressing the druglike nature of a given compound in a quantitative fashion where the issues of absorption, distribution, and excretion are taken care of through appropriate molecular descriptors.

The DLS value indicates the fact that the drugs on the market are not of the same quality. However, on the basis of the results obtained for a randomly selected drug set as a representative of the market drugs, it appears that most of the drugs are of very high quality, even though they are not of the highest quality.

The qualitative features examined for the compounds of a known drug set and several known druglike compound databases indicated features that are expected.

This score can also be used to discriminate between druglike compounds and nondruglike ones.

Generated variants around a given core compound may have significant diversity in their respective DLS values depending on the druglike qualities of the fragments used to generate them. DLS, thus, can be used to design compounds with adequate druglikeness.

REFERENCES AND NOTES

(1) DiMasi, J. A. Success rates for new drugs entering clinical testing in the United States. *Clin. Pharmacol. Ther.* **1995**, *58*, 1−14.
(2) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(3) 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.
(4) 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.
(5) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(6) Ajay; Walters, W. P.; Murcko, M. Can We Learn To Distinguish between "Drug-like" and "Nondrug-like" Molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.
(7) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem.* **1999**, *1*, 55−68.
(8) Bemis, G. W.; Murcko, M. A. Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42*, 5095−5099.
(9) Chen, X.; Rusinko, A., III; Tropsha, A.; Young, S. S. Automated Pharmacophore Identification for Large Chemical Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 887−896.
(10) Oprea T. I. Property distribution of drug-related chemical databases. *J. Comput.-Aided. Mol. Des.* **2000**, *14*, 251−264.
(11) Xu, J.; Stevension, J. Drug-like Index: A New Approach To Measure Drug-like Compounds and Their Diversity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177−1187.
(12) 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.
(13) Muegge, I.; Heald, S. L.; Brittelli, D. Simple Selection Criteria for Drug-like Chemical Matter. *J. Med. Chem.* **2001**, *44*, 1841−1846.
(14) Oprea, T. I. Current trends in lead discovery: are we looking for the appropriate properties? *J. Comput.-Aided Mol. Des.* **2002** *16*, 325−334.

(15) 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.

(16) Leeson, P. D.; Davis, A. M. Time-Related Differences in the Physical Property Profiles of Oral Drugs. *J. Med. Chem.* **2004**, *47*, 6338−6348.

(17) Vieth, M.; Siegel, M. G.; Higgs, R. E.; Watson, I. A.; Robertson, D. H.; Savin, K. A.; Durst, G. L.; Hipskind, P. Characteristic Physical Properties and Structural Fragments of Marketed Oral Drugs. *J. Med. Chem.* **2004**, *47*, 224−232.

(18) Fausett, L. *Fundamentals of Neural Networks*; Prentice Hall: New York, 1994.

(19) Introduction to Neural Networks. http://www.idsia.ch/NNcourse/intro.html (accessed Aug 2005).

(20) Barthe, L.; Woodley, J.; Houin, G. *Fundam. Clin. Pharmacol.* **1999**, *13*, 154.

(21) HIC-Up welcome page. http://xray.bmc.uu.se/hicup/welcome.html (accessed Aug 2005).

(22) ChemBank Search. http://chembank.broad.harvard.edu/compounds/ (accessed Aug 2005).

(23) ASINEX − Intelligent Chemistry. http://www.asinex.com/prod/download.html (accessed Aug 2005).

(24) NCI Database Download Page. http://cactus.nci.nih.gov/ncidb2/download.html (accessed Aug 2005).

CI050459I