Chemical Information Based Scaling of Molecular Descriptors: A Universal Chemical Scale for Library Design and Analysis

Brett A. Tounge,*,† Lori B. Pfahler,§ and Charles H. Reynolds*,†

Johnson & Johnson Pharmaceutical Research and Development, L.L.C., P.O. Box 776, Welsh and McKean Roads, Spring House, Pennsylvania 19477-0776, and Merck & Co., Inc., Manufacturing Division, P.O. Box 4, 770 Sumneytown Pike, West Point, Pennsylvania 19486-0004

Received January 9, 2002

Scaling is a difficult issue for any analysis of chemical properties or molecular topology when disparate descriptors are involved. To compare properties across different data sets, a common scale must be defined. Using several publicly available databases (ACD, CMC, MDDR, and NCI) as a basis, we propose to define chemically meaningful scales for a number of molecular properties and topology descriptors. These chemically derived scaling functions have several advantages. First, it is possible to define chemically relevant scales, greatly simplifying similarity and diversity analyses across data sets. Second, this approach provides a convenient method for setting descriptor boundaries that define chemically reasonable topology spaces. For example, descriptors can be scaled so that compounds with little potential for biological activity, bioavailability, or other drug-like characteristics are easily identified as outliers. We have compiled scaling values for 314 molecular descriptors. In addition the 10th and 90th percentile values for each descriptor have been calculated for use in outlier filtering.

INTRODUCTION

The use of combinatorial chemistry in drug discovery has drastically increased the number of compounds that can be synthesized. In addition the availability of software tools and relatively inexpensive, yet powerful, computer hardware allows for the enumeration of large virtual libraries. As a result, the need has arisen for techniques that can mine databases and select subsets of compounds for testing or synthesis.[1−3] These methods include diversity/similarity analysis, clustering, large-scale QSAR methods, and schemes for distinguishing drug-like and nondrug-like molecules.[4−10] One constant among these methods is the need for scaling of the molecular descriptors. Scaling is necessary to give descriptors with different absolute ranges equal opportunity to contribute to the data analysis or modeling. For example, since molecular weight values have a large absolute range compared to log *P* values, if these descriptors were used unscaled the molecular weight would tend to dominate the analysis. In most cases this problem is solved by scaling the data based on the ranges of the descriptors in the set being studied. This has the disadvantage of biasing the scaling to any peculiarities of the data set. For example, if a set has a small molecular weight range, i.e., 50, but a large log P range, i.e., 7, the molecular weight change of 50 would be given the same importance as the log P range of 7, even though we know chemically that the log P range is much more significant.

We propose a more general method of descriptor scaling that produces chemically reasonable scales based on an analysis of the property distributions of several large databases: Available Chemicals Directory (ACD), open National Cancer Institute (open NCI), MACCS-II Drug Data Report (MDDR), and Comprehensive Medicinal Chemistry (CMC). These databases encompass a wide range of structures including both drug and nondrug compounds. Using these sets, we investigated a number of topology descriptors as well as AlogP. For each descriptor we calculated the mean and variance. The use of these ranges for both chemical based universal scaling and "outlier" filtering is discussed.

METHODS

Databases. The ACD (release 2000.2 3D)[11] and the open NCI (release 2000.2 2D)[12] were used as the nondrug databases. The ACD is a collection of commercial compounds compiled from vendor catalogs. The NCI contains molecules that have been submitted to the National Cancer Institute for testing, primarily as cancer treatments. For this study we are using the open NCI database which is simply a subset of the full NCI containing compounds whose structures can be made publicly available. For these databases, all entries without structures, with atom counts greater than 400, and those that contained no carbon atoms were removed. After this filtering, the files were run through the "Wash" routine of Chemical Computing Group's MOE (version 2001.01) program to remove solvent and counterions. Of the initial 316 922 ACD compounds, 265 480 remained after filtering. For the open NCI database, 243 491 of the original 250 251 compounds remained after filtering.

The drug-like compounds for this study are represented by the MDDR (release 2000.2 3D)[11,13] and the CMC (release 2000.1 3D)[11] databases. These two sets in turn represent two classes of drug-like compounds. The MDDR is comprised of compounds that were synthesized and screened for medical

* Corresponding authors phone: (215)628-5230; fax: (215)628-4985; e-mail: btounge@prdus.jnj.com (Tounge); e-mail: creynol1@prdus.jnj.com (Reynolds).
† Johnson & Johnson Pharmaceutical Research and Development.
§ Merck & Co., Inc.

10.1021/ci025503y CCC: $22.00
Published on Web 05/10/2002

Table 1. Final Database Sizes

database	entries
ACD	253318
NCI	205842
MDDR	105095
CMC	5713
Cox2	501
protease inhibitors	33

use but have not necessarily passed clinical trials. It is compiled from patent literature, journals, meetings, and congresses. The CMC is a more focused drug database. All compounds in this database have been assigned United States Approved Names (USAN) indicating that they have likely entered at least Phase II testing. For these databases further filtering (in addition to the methods used for the nondrug databases) was done to ensure the compounds fit the drug-like criterion. All compounds without activity information and with the following activity classes were removed: adsorption promoters, aerosols, alcohol denaturants, anesthetics, antidotes, antifoams, antioxidants, antiperspirants, astringents, blood substitutes, biochemical reducing agents, buffering agents, bulking agents, chelating agents, contraceptives, cosmetics, dental, diagnostic agents/aids, dietary supplements, disinfectants, dyes, emollients, emulsifiers, food additives, insect repellents, insecticides, herbicides, imaging agents, metal complexes, minerals, pharmaceutical aids/tools, pregnancy tests, photosensitizers, plant growth regulators, prosthetic aids, propellants, radio pharmaceuticals, resins, surfactants, sunscreens, surgical aids, sweeteners, topicals, ultraviolet light absorbers, vaccines, and vitamins. This list is consistent with that used by other authors for removing compounds that are not typical pharmeceuticals.[14] Out of the initial 113 842 entries for the MDDR, 106 843 remained after the filtering. For the CMC, 5813 out of 7937 entries were kept.

In addition to the above databases, two smaller sets were studied. The first set, Cox2, is simply a subset of the MDDR. It consists of 501 entries from the MDDR that contained Cox2 in the "Activity" field. The protease inhibitor subset (PI) is part of the Supporting Information from Holloway et al.[15] It consists of 33 compounds that were designed as inhibitors of HIV protease. These two data sets were included because each is very focused and is pharmaceutically relevant, and because they are quite different from one another structurally.

Descriptors. Two commercial packages were used to generate the descriptor sets for this study. The molecular topology descriptors were generated using the Molconn-Z (version 3.51) program from eduSoft, LC. Using this package, 313 descriptors per molecule were calculated (records 1−8, 15−21, 35−36, 40−43). The AlogP values were calculated using Cerius^2 from Accelrys (AlogP98 parameter set).[16] A certain number of entries in each data set failed either the topology descriptor or AlogP98 calculation. These structures were omitted from further consideration. The final data set sizes, which include all entries for which the entire descriptor set was calculated, are summarized in Table 1. Many of the high molecular weight compounds failed one or more descriptor calculations, so they are not included in the final sets. However, the number of these compounds is small, so their absence has little effect on the scaling values. They would just appear in the boxplots (described below) as outliers.

Statistics. To avoid giving too much weight to outliers in the determination of the universal scaling factors, it was important to choose a variance measure with the least dependence on the extreme values of the data sets. For example, in the case of range scaling the following formula is used

$$z_i = \frac{x_i - \min(x)}{(\max(x) - \min(x))} \tag{1}$$

where z_i is the scaled descriptor, x_i is the unscaled descriptor, and $\min(x)$ and $\max(x)$ are the minimum and maximum values in the data set, respectively. Due to the explicit use of the maximum and minimum values of the data set, the scaling of the data can be skewed significantly by a single outlier. An alternative method is to use the mean ($\bar{x}$) and standard deviation (sd) of the data set

$$z_i = \frac{x_i - x}{\mathrm{sd}}, \quad \mathrm{sd} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}} \tag{2}$$

where n is the number of data points. While this method is less sensitive to outliers, the squared dependence on the deviation from the mean still weights the outlier values rather heavily. For this reason we have replaced the standard deviation with the Mean Absolute Deviation (mad) which is given by

$$\mathrm{mad} = \frac{1}{n}\sum |x_i - \bar{x}| \tag{3}$$

This removes the squared dependence on the deviation from the mean, so the method is much less sensitive to outliers.

All the statistics were generated using the S-plus package from Insightful Corporation. This package was also used to generate boxplots of the data. These plots are a useful visual tool for displaying the distributions of the descriptors among the data sets. They are constructed as follows. First, a dot is drawn representing the mean of the data set. Next a box (solid line) is drawn which encompass the 25th to 75th percentiles. Finally, the whiskers (dotted lines) are drawn. They are calculated as 1.5 times the spread of data in the box and are drawn either to that number or to the end of the data set. The open circles outside of the whiskers represent outliers.

RESULTS AND DISCUSSION

Descriptor Space Coverage. A summary of the distribution for a few of the key descriptors can be found in Table 2. These descriptors are representative of the general trend in that their distributions are remarkably consistent across the data sets.[17] The extent of this overlap is evident in the boxplots in Figures 1−7. With the exception of molecular weight, there is significant overlap of the 25th to 75th percentiles. The most significant differences among the sets occur in the outliers values (open circles). This is particularly true of the ACD, which contains many extreme values.

Despite the extent of their overlap, some differences can be seen among the sets. The most significant trend is the typically larger distribution and higher means for the drug-

Table 2. Sample Summary Statistics for Each Database

variable	data set	mean	mad	10th percentile	90th percentile
MW	ACD	311.96	96.58	167.16	455.38
	MDDR	442.71	122.34	272.39	642.75
	NCI	287.92	83.02	166.14	428.40
	CMC	363.46	113.10	205.28	535.06
	Cox2	381.71	46.46	314.38	449.42
	PI	574.95	33.14	524.83	621.15
Kappa Alpha 3	ACD	4.46	2.15	1.78	7.10
	MDDR	5.79	2.68	2.43	10.12
	NCI	3.89	1.79	1.58	6.76
	CMC	4.67	2.22	1.94	7.91
	Cox2	3.92	0.75	2.92	5.20
	PI	10.80	0.71	9.75	12.22
H-bond donors	ACD	1.32	1.11	0.00	3.00
	MDDR	2.15	1.51	0.00	4.00
	NCI	1.35	1.11	0.00	3.00
	CMC	1.90	1.46	0.00	4.00
	Cox2	0.72	0.61	0.00	2.00
	PI	6.12	0.49	5.00	7.00
H-bond acceptors	ACD	5.50	2.50	2.00	9.00
	MDDR	7.56	3.02	3.00	13.00
	NCI	4.90	2.14	2.00	8.00
	CMC	6.14	2.91	2.00	11.00
	Cox2	5.85	1.75	3.00	9.00
	PI	7.52	0.90	6.00	8.00
AlogP98	ACD	3.19	1.73	0.58	5.82
	MDDR	3.43	1.74	0.64	6.10
	NCI	2.62	1.57	0.16	5.12
	CMC	2.66	1.66	0.04	5.10
	Cox2	4.09	0.99	2.52	5.66
	PI	7.19	0.79	6.13	8.29
SsssN	ACD	0.44	0.74	0.00	1.88
	MDDR	1.49	1.42	0.00	4.18
	NCI	0.65	0.95	0.00	2.30
	CMC	1.23	1.39	0.00	3.84
	Cox2	0.30	0.47	0.00	1.36
	PI	0.01	0.02	0.00	0.00
SdO	ACD	15.72	11.89	0.00	35.29
	MDDR	24.30	16.59	0.00	51.45
	NCI	14.09	11.15	0.00	23.02
	CMC	18.88	14.87	0.00	38.57
	Cox2	25.52	7.33	11.56	36.33
	PI	0.00	0.00	0.00	0.00

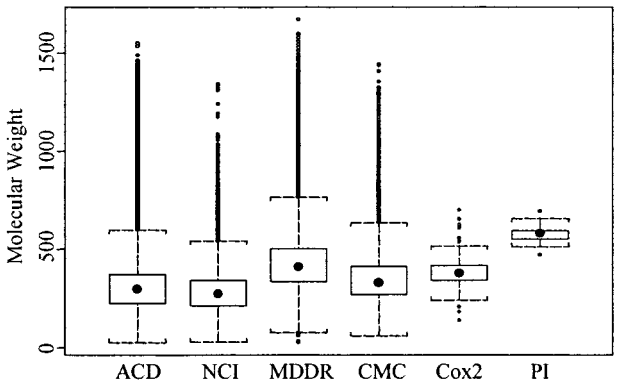

Figure 1. Boxplot comparison of molecular weight across the data sets. See the Methods section for an explanation of the boxplots. Compounds with molecular weights above 1750 do exist in the ACD, but they all failed on one or more other descriptor calculations.

like data sets. This tendency has been noted by other authors and is indicative of the greater complexity that is typical of drug-like compounds in general.[18−20] Part of the reason for this spread can be seen in the ranges defined by the Cox2 and Protease Inhibitor subsets. Each subset has a narrow,

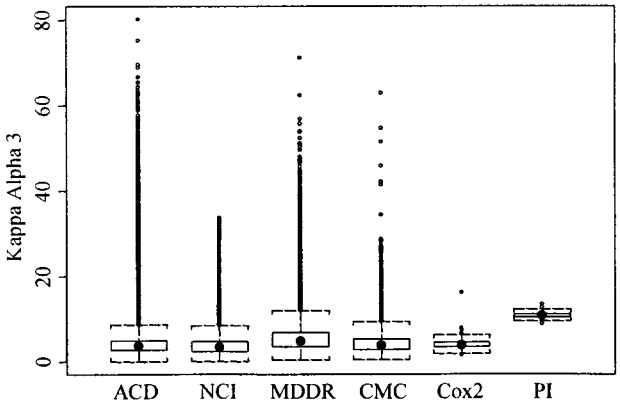

Figure 2. Boxplot comparison of Kappa Alpha 3 across the data sets. See the Methods section for an explanation of the boxplots.

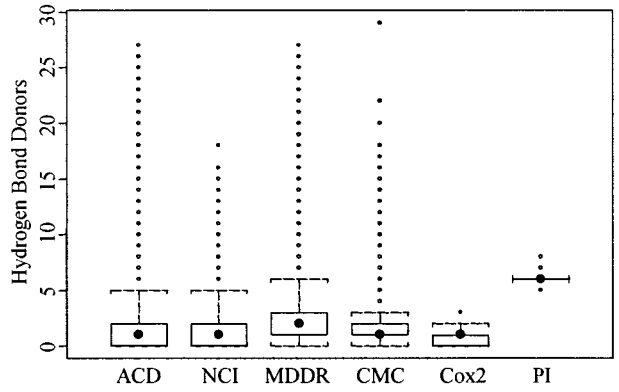

Figure 3. Boxplot comparison of the number of hydrogen bond donors across the data sets. To see the distributions more clearly, points above 30 (outliers) were removed. See the Methods section for an explanation of the boxplots.

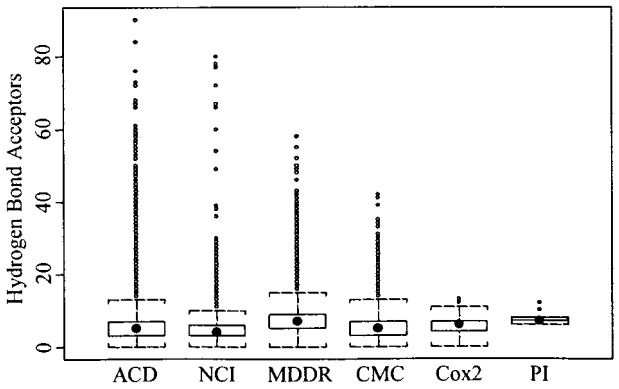

Figure 4. Boxplot comparison of the number of hydrogen bond acceptors across the data sets. See the Methods section for an explanation of the boxplots.

yet distinct, range. This results from the fact that these subsets are each designed around a specific drug target that enforces distinct physical properties on the drug molecules. Given the diversity of drug targets, it is not surprising that the drug databases as a whole tend to have the largest property distributions.

The coverage of chemical space for the descriptors used in this study is well illustrated in Table 3. Here we compare two versions of the ACD, one from 1998 and the 2000.2 version. Despite the increased number of compounds (68 172 more in the newest ACD) the statistics show very little change. This indicates that the amount of chemical space

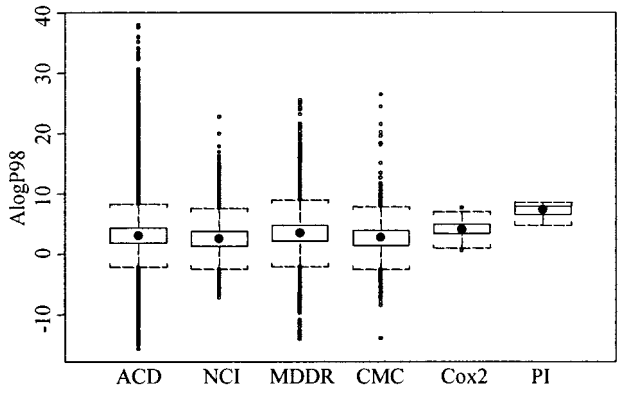

Figure 5. Boxplot comparison of AlogP98 across the data sets. See the Methods section for an explanation of the boxplots.

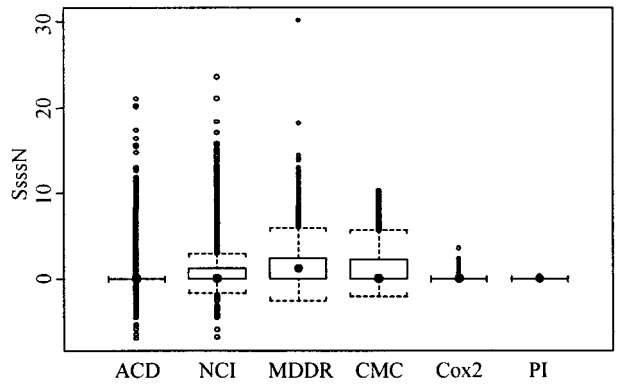

Figure 6. Boxplot comparison of the sum e-state SsssN across the data sets. See the Methods section for an explanation of the boxplots

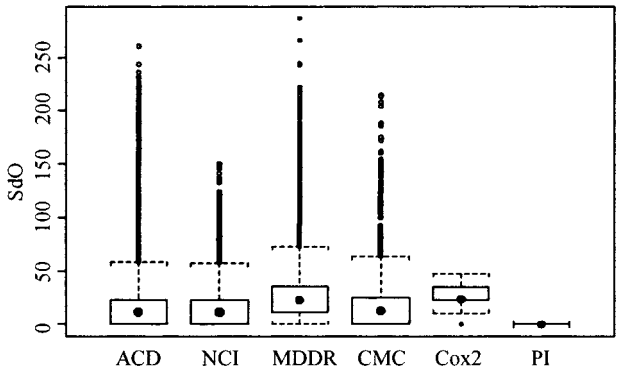

Figure 7. Boxplot comparison of the sum e-state SdO across the data sets. See the Methods section for an explanation of the boxplots

covered by the ACD is not changing significantly over time, so the scaling factors are robust with respect to the specific database or vintage of the database chosen. Given this trend, the chemically derived descriptor scales described herein should not need to be updated as more compounds are added to the respective databases since the chemical space is apparently well covered already.

Universal Chemical Based Scale. To scale a given set of data, the following formula is used

$$z_i = \frac{x_i - \bar{x}}{\text{mad}} \tag{4}$$

where the mean and mad are taken from one of the universal scaling sets. Once applied, outliers become obvious. For a Gaussian distribution, 98.4% of the data will lie within ±3 standard deviations (in this case mad) of the mean. This means that once scaled most of the data points would range from −3 to +3. Since the distributions of the descriptors used in this study tend to deviate from a Gaussian, the exact percentage of values in the −3 to +3 range will differ, but most of the scaled descriptors will still lie in this range.

Table 3. Comparison of 1998 and 2000.2 Version of the ACD

descriptor	statistic	ACD (1998) 185 146 compds	ACD (2000.2) 253 318 compds
MW	mean	314	312
	25th percentile	224	223
	75th percentile	372	373
Kappa Alpha 3	mean	4.5	4.5
	25th percentile	2.5	2.5
	75th percentile	4.9	5.0
H-bond donors	mean	1.4	1.3
	25th percentile	0	0
	75th percentile	2	2
H-bond acceptors	mean	5.4	5.5
	25th percentile	3	3
	75th percentile	7	7

The choice of which (large) database to use for scaling is not critical based on our analysis, but one might as well select a data set that is relevant to the intended application. For example, if the goal is simply to compare the distribution of properties across databases, then any of the large sets would be appropriate. However, if the goal is to compare the distribution of a data set to known drug compounds, then the CMC scaling factors would be more applicable.

A chemically meaningful absolute descriptor scale that can be applied across data sets has many advantages for QSAR and library design. First a "universal" scale greatly simplifies the task of identifying descriptors that have no chemically significant variance and that are only likely to lead to spurious correlations. For example in the case of the protease inhibitor data set, the SsssN sum electrotopological-state (e-state)[21,22] has a nonzero variance (mad = 0.0197). However, when this value is compared to the variance for the CMC SsssN e-state (mad = 1.384), it is obvious that the variance in the protease inhibitor set is actually quite small. Inclusion of variables with very small, essentially zero, variance can be a particularly significant issue in QSAR approaches where large numbers of descriptors are fed into a variable selection scheme, such as a genetic algorithm.[23] Descriptors that do not have any significant variance over the training set not only waste computer time but also can be incorporated into QSAR models because the noise in their values fortuitously correlates with the response property. This may lead to QSAR relationships that have very low predictive power outside the training set. While this might seem unlikely, the probability of this kind of coincidental correlation increases with the number of descriptors evaluated.[24]

Another obvious advantage for a chemically meaningful absolute scale is that it simplifies comparisons across disparate data sets. For example, when comparing the HIV protease inhibitors to the Cox2 inhibitors the fact that the protease set has a narrow, but higher, molecular weight range is not masked by the different scales that would result from individual scaling. The advantage of this consistent scaling is even more significant in the case of library design applications that depend on distance relationships between

Table 4. Percent of Compounds that Pass the "Rule of 5"

data set	MW ≤ 500 (%)	H-bond donors ≤ 5 (%)	H-bond acceptors ≤ 10 (%)	ALogP98 ≤ 5 (%)	pass all (%)
ACD	94	98	94	83	77
NCI	96	98	96	89	84
MDDR	74	94	83	78	59
CMC	88	95	90	89	76
Cox2	98	100	99	75	73
PI	3	15	97	3	0

Table 5. Comparison of 10th and 90th Percentiles for the ACD and CMC[a]

	ACD		CMC	
	10th percentile	90th percentile	10th percentile	90th percentile
MW (500)*	167.16	455.38	205.28	535.06
H-bond donors (5)*	0	3	0	4
H-bond acceptors (10)*	2	9	2	11
Alogp98 (5)*	0.58	5.82	0.04	5.10
number of rings	1	4	1	5
Kappa Alpha 3	1.78	7.10	1.94	7.91

[a] The asterisk denotes the "rule of 5" values.

compounds in a multidimensional descriptor space. When libraries are scaled individually they cannot be compared directly. Thus it is necessary to plan ahead and make arbitrary choices about a common scale for different libraries if the ultimate goal is to compare them in some way (e.g. pick a diverse or similar subset of a trial library relative to a reference library). If the libraries have all already been placed on a consistent chemically relevant scale, these comparisons become much more straightforward. Further, distances have more meaning in an absolute sense, i.e., it is possible to determine a priori if a distance is small or large in the scaled descriptor space.

Outlier Filtering. One of the best known applications of property distributions for outlier detection is the "rule of 5".[25,26] In this case, the upper bounds (90th percentile) of property distributions from a subset of the World Drug Index (WDI) were used to select compounds that were unlikely to be orally available. The application of these rules to the data sets used in this paper is summarized in Table 4. It is immediately obvious that the "rule of 5" is not a filter for drug-like compounds. A greater percentage of the compounds in the CMC and MDDR are filtered out than in the ACD or NCI. The "rule of 5" is simply a method for flagging compounds as being on the extreme ends of a given descriptor distribution (i.e. outlier), and by inference as being unlikely to have the desired properties, in this case oral bioavailability.

The chemically derived scale outlined in this paper can also be used as an outlier filter. A simple filter can be defined quite generally for any descriptor by taking the 10th to 90th percentile of a given descriptor distribution. Some sample values are given in Table 5. For comparison, the distribution for the "rule of 5" values is given. The cutoffs for the 90th percentile are quite close to those determined by Lipinski et al. for log P, molecular weight, hydrogen bond donors and acceptors. The application of this filter is trivial when using the "universally" scaled descriptor values. For example, any value outside of the −1.40 to 1.52 range represents a descriptor that is beyond the 10th and 90th percentiles, respectively, for molecular weight in the CMC. In this way, any relevant descriptor can be used as a filter once it has been scaled relative to a reference database.

Table 6. Variables Used in the Drug-like Filter[a]

variable	10th percentile CMC
fw	205.277
Gmin	−3.15156
ka1	10.4679
nvx	14
SHother	5.94212
SssssC	−1.78484
sumI	34.6667
x0	10.552
x1	6.68488
x2	6.07318
xch6	0.037
xp3	4.4349
xp4	3.25286
xp5	2.20734
xp6	1.29462
xv0	8.47066
xv1	4.81308
xv2	3.56682
xvch6	0.0125
dx2	1.27988
dxp3	1.19214
dxp5	0.83642

[a] Descriptors were chosen by an analysis of the lower bounds (10th percentile) of the CMC and the ACD. Descriptors that had a significantly higher lower bound in the CMC were chosen for inclusion in the filter.

Drug-like Filter. As stated previously, the majority of the differences in the property distributions between the drug-like and nondrug databases occurs in the lower bound. Of the 314 descriptors we studied, the 10th percentile of the ACD was lower in 104 cases than the CMC, whereas only 15 descriptors had a higher 90th percentile in the ACD as compared to the CMC. This same trend was noted by Oprea et al.[20] In their work, they took advantage of the disparate distributions to develop a filter for distinguishing drug-like and nondrug compounds. This type of filtering can be generally applied to any descriptor by studying the lower bounds of the CMC and ACD for the particular descriptor. For example, we defined a filter by taking the lower bound of the CMC for 22 descriptors for which the ACD had a significantly lower value than the CMC (Table 6). These cutoffs were then applied to all entries in the ACD and CMC. For the CMC 67% of the compounds passed all the rules, while only 45% of the ACD passed. While this method shows some useful discrimination between the presumably drug-rich CMC and the presumably drug-poor ACD, the percentages are far from ideal. The quality of the division between drug and nondrug compounds might be improved by optimizing the particular descriptors used, the ranges selected for the filters, or by constructing a probability based scoring function that depends on the descriptor ranges in a more sophisticated manner.[27] However, this simple exercise illustrates how the scaled descriptors might be used in developing a filter.

CONCLUSIONS

Through analysis of several large databases, we have developed chemically based and transferable scaling factors for a set of 314 molecular descriptors. These factors are based

on the mean and mean absolute deviation of the property distribution. Using these values, scaled descriptors can be defined. These scaled descriptors provide a chemically meaningful space onto which any data set can be mapped without the need for individual scaling. This allows for direct comparison of descriptors across databases and eliminates the need to define different scaling factors for each QSAR and library design application. In addition, chemically relevant scaling makes outlier values immediately obvious since almost all compounds (∼98%) should fall within the −3 to +3 range.

A comparison of the databases reveals that there is a great deal of homogeneity in the property distributions. For most descriptors, there is significant overlap of the 25th to 75th percentiles. However, there is a trend to higher values in the CMC and the MDDR. These databases, representing drug-like compounds, cover a larger chemical space due in part to the greater complexity that results from compounds designed specifically to be drugs. This illustrates two interesting points. First, almost any of the standard large structural databases can reasonably be used to create a transferable chemically relevant (universal) scale for molecular descriptors. Second, despite common misconceptions, the drug databases (CMC and MDDR) are actually more diverse on average than the ACD. The ACD only appears more diverse when looking at absolute extremes because it contains more structures that are truly outliers.

ACKNOWLEDGMENT

Brett Tounge thanks the Johnson & Johnson Corporate Office of Science and Technology for support of his postdoctoral fellowship.

Supporting Information Available: Space delimited ASCII file and Excel 2000 spreadsheet with summary statistics for all 314 descriptors for the ACD and CMC. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Kubinyi, H. Combinatorial and computational approaches in structure-based drug design. *Curr. Opin. Drug Discovery Dev.* **1998**, *1*, 16−27.
(2) Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity. *Perspect. Drug Discovery Des.* **1998**, *9*, 339−353.
(3) Willett, P. Chemoinformatics − similarity and diversity in chemical libraries. *Curr. Opin. Drug Discovery Dev.* **2000**, *11*, 85−88.
(4) Van Drie, J. H.; Lajiness, M. S. Approaches to virtual library design. *Drug Discovery Today* **1998**, *3*, 274−283.
(5) Mitchell, T.; Showell, G. A. Design strategies for building drug-like chemical libraries. *Curr. Opin. Drug Discovery Dev.* **2001**, *4*, 314−318.
(6) Lewis, R. A.; Pickett, S. D.; Clark, D. E. Computer-aided molecular diversity analysis and combinatorial library design. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Ed.; VCH Publishers: New York, 2000; Vol. 16, pp 1−51.
(7) Gorse, D.; Lahana, R. Functional diversity of compound libraries. *Curr. Opin. Chem. Biol.* **2000**, *4*, 287−294.
(8) Sadowski, J. Optimization of chemical libraries by neural networks. *Curr. Opin. Chem. Biol.* **2000**, *4*, 280−282.
(9) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(10) 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.
(11) MDL Information Systems, Inc., http://www.mdli.com.
(12) National Cancer Institute, http://cactus.nci.nih.gov/ncidb2/download.html.
(13) Prous Science, http://www.prous.com.
(14) Bemis, G. W.; Murcko, M. A. Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(15) Holloway, M. K.; Wai, J. M.; Halgren, T. A.; Fitzgerald, P. M. D.; Vacca, J. P.; Dorsey, B. D.; Levin, R. B.; Thompson, W. J.; Chen, L. J.; et al. A priori prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site. *J. Med. Chem.* **1995**, *38*, 305−317.
(16) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J. Phys. Chem. A* **1998**, *102*, 3762−3772.
(17) Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. Molecular Diversity in Chemical Databases: Comparison of Medicinal Chemistry Knowledge Bases and Databases of Commercially Available Compounds. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750−763.
(18) Menard, P. R.; Mason, J. S.; Morize, I.; Bauerschmidt, S. Chemistry Space Metrics in Diversity Analysis, Library Design, and Compound Selection. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1204−1213.
(19) Voigt, J. H.; Bienfait, B.; Wang, S.; Nicklaus, M. C. Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 702−712.
(20) Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251−264.
(21) Hall, L. H.; Kier, L. B. The electrotopological state: An atomic index for QSAR. *Quant. Structure−Activity Relat.* **1991**, *10*, 43−48.
(22) Hall, L. H.; Kier, L. B. The electrotopological state: Structure information at the atomic level for molecular graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76−83.
(23) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.
(24) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure−Activity Relationships. *J. Med. Chem.* **1979**, *22*, 2.
(25) 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.* **2001**, *46*, 3−26.
(26) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2001**, *44*, 235−249.
(27) Arup, K. G.; Viswandahan, 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.

CI025503Y