Chemical Descriptors with Distinct Levels of Information Content and Varying Sensitivity to Differences between Selected Compound Databases Identified by SE-DSE Analysis

Jeffrey W. Godden[†] and Jürgen Bajorath*[,#,†]

Albany Molecular Research, Inc., Bothell Research Center (AMRI-BRC), 18804 North Creek Parkway, Bothell, Washington 98011, and AMRI-BRC and Department of Biological Structure, University of Washington, Seattle, Washington 98195

Received July 4, 2001

Analysis of the variability of molecular descriptors in large compound databases has recently been carried out using both the Shannon entropy (SE) and differential Shannon entropy (DSE) concepts that reduce descriptor distributions to their information content (SE analysis) and detect intrinsic differences between descriptor settings in compound databases (DSE analysis). Here it is shown that a combination of SE and DSE calculations, termed SE-DSE analysis, makes it possible to identify molecular descriptors most sensitive to systematic differences in databases consisting of synthetic, drug-like, and natural molecules. Descriptors with consistently high information content are detected, and database-specific differences are quantified. Different sets of only very few descriptors were found to be most responsive to principal differences between synthetic, natural, and drug-like molecules. Descriptors with DSE values furthest away from zero are likely to best distinguish between compounds with different characteristics. SE-DSE analysis also reveals that a number of descriptors are not sensitive to compound class-specific features, despite their complexity and consistently high information content.

INTRODUCTION

A wide variety of chemical descriptors are used to quantitatively describe molecular structures and properties for many applications in computational chemistry and chemoinformatics.[1,2] The choice of descriptors appropriate for specific computational tasks, for example, similarity searching, compound classification, or database comparison, is often far from being obvious. Thus, it would be beneficial to be able to compare intrinsic characteristics of such descriptors, for example, how effectively they capture molecular information or how sensitive they are to chemical differences. Such insights are thought to provide a basis for the identification of compound class-specific descriptors or of those descriptors most likely to detect and quantify molecular diversity or similarity.

However, comparing the variability of chemical descriptors in molecular data sets and their sensitivity to specific chemical features is hindered by the fact that very many different types of descriptors exist, having different units and value ranges, which capture a wide spectrum of chemical and structural properties.[1,3] Thus, a direct comparison of molecular descriptor settings and their variability is often similar to the problem of comparing "apples and oranges". Nevertheless, if reduced to some essential features, for example, carbohydrate or vitamin content, even "apples and oranges" can be compared, albeit indirectly. Similarly, different molecular descriptors and their values could be compared, regardless of their nature, if it were possible to reduce them to common characteristics or quantities. However, a solution to the problem is less obvious in this case, and it required considerable time and effort until we recognized that introduction of measures such as "information content" or "relative information content" would open the door to systematic and large-scale comparison of diverse molecular descriptors.

A formalism to calculate and compare the information content of data distributions has in fact been introduced, however, not in chemistry but digital communication theory. This concept, an entropic formulation termed Shannon entropy,[4] allows us to reduce any data distribution to its information content that can be represented in histograms with a defined binning scheme. We have therefore adapted[5,6] and modified[6,7] this concept to study and compare the information captured by various descriptors in different compound databases. While SE calculations permit the quantification of information content of single descriptors, the differential Shannon entropy approach, an extension of this concept, takes both differences in the variability and value range distributions of descriptors into account.[7] This in turn makes it possible to quantitatively compare the information content of selected descriptors in different molecular datasets, even if differences in the distribution of their values are subtle and difficult to detect.

In initial studies,[6,7] we were able to demonstrate that SE and DSE calculations can identify descriptors that are sensitive to intrinsic differences in compound collections, for example, natural products and synthetic molecules.[6]

* Corresponding author phone: (425)424-7297; fax: (425)424-7299; e-mail: jurgen.bajorath@albmolecular.com.

[†] Albany Molecular Research, Inc., Bothell Research Center (AMRI-BRC).

[#] University of Washington.

10.1021/ci0103065 CCC: $22.00
Published on Web 11/28/2001

Furthermore, we could show, with the introduction of the DSE concept, that a number of descriptors yielding the largest DSE values in database comparison were chemically intuitive. For example, we found that descriptors capturing nitrogen or halogen content were among those with largest DSE values when comparing natural product and synthetic compound databases.[7] These findings suggested to us that entropy-based analysis of descriptor information content is capable of identifying descriptors that are most sensitive to the intrinsic chemical features that characterize different compound sets, and we have therefore extended our investigation of this approach.

To provide a conceptually novel basis for descriptor selection, we have now gone a step further and closely coupled SE and DSE analysis. To facilitate effective combination of these approaches, some extensions of the methodology were required, as reported herein, including calculation of histogram bin-independent SE and DSE values and the determination of generally applicable SE and DSE threshold values for descriptor comparison and classification. Applying the SE-DSE concept, we have then continued our systematic descriptor analysis by pairwise comparison of approximately 150 descriptors in four representative molecular databases, consisting of synthetic compounds, natural or drug-like molecules, or known drugs. For each comparison, we have determined descriptors with consistently high information content and significant differences in value distribution and have identified those descriptors that are most responsive to distinct chemical features in the compound classes studied here.

METHODS

SE-DSE Concept. Shannon entropy is defined and calculated as

$$SE = -\sum p_i \log_2 p_i \quad (1)$$

where '*p*' is the sample probability of a data point to fall as a count '*c*' within a specific data range '*i*'. '*p*' therefore is obtained as

$$p_i = c_i / \sum c_i \quad (2)$$

The logarithm to the base two represents a scale factor which permits SE to be considered as a metric of information content, entirely like answering the question of how many different signals can be carried within a fixed framework of bits. It can be rationalized as a "binary" detector of counts (i.e., does the count appear in a given data interval or not?). The fixed bit framework is established by consistently using a number of histogram bins to represent each data range, here 100 bins. The major benefit of these manipulations is that descriptors with very different distributions and ranges are rendered comparable. Captured by the SE metric, descriptor variability may vary from zero for a single valued descriptor to a maximum of the logarithm to the base two of the number of bins utilized. In the case of 100 bins, this maximum value is ~6.64. While the choice of 100 bins is convenient for this study, it may also be necessary to analyze smaller compound collections where 100 bins would be excessive. It therefore becomes useful to establish a bin independent SE value, scaled SE or sSE, that can be generally compared, regardless of the number of bins used for histogram representation of datasets. sSE is obtained by dividing an observed SE value by the maximum possible SE value for the number of bins used (SE divided by the logarithm to the base two of the number of bins).

$$sSE = SE/\log_2(\text{bins}) \quad (3)$$

Differential Shannon entropy (DSE) extends our adoption of the SE concept and provides a means of comparing the SE variability of two compound populations. DSE is defined as

$$DSE = SE_{AB} - (SE_A + SE_B)/2 \quad (4)$$

where "SE_{AB}" is the SE calculated for the combination of two compound databases A and B under consideration. Therefore, DSE is a measure of the complementarity of two compound collections with regard to the descriptor under analysis. Therefore, a "high" DSE value would reflect a descriptor that detects a significant difference between the two compound sets (thus, the variance of the data set changes significantly from the one in each collection when the two datasets are combined). By contrast, a "low" DSE value indicates that a descriptor monitors the two compound sets like a single population and that there is little or no difference between these databases with respect to the descriptor in question.

Analogous to sSE, there is a bin-independent form of DSE

$$sDSE = DSE/\log_2(\text{bins}) \quad (5)$$

While SE calculations reduce data or descriptor distribution to an entropy value and information content, calculation of DSE introduces a value-range dependence of this information content. Thus, combining SE and DSE calculations as a two step SE-DSE analysis allows to first select descriptors having consistently high information content (high SE values) and then detect those descriptors that have significant differences in their value range distributions in two databases under comparison (high DSE values).

Databases. Four compound databases were used in this study: the Available Chemicals Database (ACD)[8] currently with 199 419 entries, Chapman and Hall (C&H)[9] containing 116 364 selected entries, the Comprehensive Medicinal Chemistry database (CMC)[10] with 7576 entries, and the drug subset of Synthline (SYNTH)[11] consisting of 4124 entries. ACD spans a large array of synthetic compounds including many commonly used reagents, C&H is a repository of natural products, CMC is a collection of molecules with drug-like properties commonly used in medicinal and lead optimization chemistry, and the Synthline subset used here consists of established drugs or late stage drug candidates. In all cases, the database entries were filtered to exclude single ions, hydrates, and separate other noncovalent complexes. Thus, only single molecules were used for our analysis.

Descriptors. A total of 143 single numerical descriptors[7] were evaluated in our study. These descriptors account for a number of diverse properties including physicochemical and bulk parameters, atom and bond counts, chemical composition, and surface, topological, or shape descriptors.[1−3] Our only requirement for selection of descriptors was that

Table 1. Definitions of Descriptors Discussed in This Study

name	description
SlogP_VSA0	ca. van der Waals atomic surface with logP < -0.4[14]
SlogP_VSA1	ca. van der Waals atomic surface with $-0.4 <$ logP ≤ -0.2[14]
SlogP_VSA2	ca. van der Waals atomic surface with $-0.2 <$ logP ≤ 0.0[14]
SlogP_VSA6	ca. van der Waals atomic surface with $0.2 <$ logP ≤ 0.25[14]
SlogP_VSA7	ca. van der Waals atomic surface with $0.25 <$ logP ≤ 0.3[14]
SMR_VSA3	ca. van der Waals atomic surface with molar refractivity is $0.35 < R_i \leq 0.39$[14,15]
SMR_VSA5	ca. van der Waals atomic surface with molar refractivity is $0.44 < R_i \leq 0.485$[14,15]
SMR_VSA6	ca. van der Waals atomic surface such with molar refractivity is $0.485 < R_i \leq 0.56$[14,15]
a_ICM	entropy of the element distribution in the molecule calculated from its atomic composition[12]
a_base	number of basic atoms
a_nC	number of carbon atoms
a_nN	number of nitrogen atoms
a_nF	number of fluorine atoms
a_nCl	number of chlorine atoms.
a_nBr	number of bromine atoms
a_nI	number of iodine atoms
a_nS	number of sulfur atoms
a_nP	number of phosphate atoms
a_hyd	number of hydrophobic atoms
b_rotR	fraction of rotatable bonds
b_1rotR	fraction of rotatable single bonds
b_heavy	number of bonds between heavy atoms
b_single	number of single bonds
b_triple	number of triple bones
balabanJ	Balaban's topological connectivity index[16]
chi0_C	sum of the inverse square root of heavy atoms bonded to each atom
chi0v_C	sum of the inverse square root of a valence electron function
chi1_C	sum of the inverse square root of cross terms of a valence electron function
weinerPol	half the sum of all the distance matrix entries with a value of 3[17]
weinerPath	half the sum of all distance matrix entries[17]
vsa_base	ca. van der Waals surface area of basic atoms
vsa_don	ca. van der Waals surface area of hydrogen bond donors (not counting those that are also acceptors)
vsa_hyd	ca. van der Waals surface area of hydrophobic atoms
vsa_pol	ca. van der Waals surface area of H-bond donors (counting OH)
PEOE_VSA-4	van der Waals surface area where atomic partial charge $-0.25 \leq q < -0.2$[14,18]
PEOE_VSA-3	van der Waals surface area where atomic partial charge $-0.20 \leq q < -0.15$[14,18]
PEOE_VSA-2	van der Waals surface area where atomic partial charge $-0.15 \leq q < 0 \times 10$[14,18]
PEOE_VSA+0	van der Waals surface area where atomic partial charge $0.00 \leq q < 0.05$[14,18]
PEOE_VSA+1	van der Waals surface area where atomic partial charge $0.05 \leq q < 0 \times 10$[14,18]
PEOE_VSA+6	van der Waals surface area where atomic partial charge is greater than 0.3[14,18]
PEOE_VSA_HYD	polar van der Waals surface area of hydrophobic atoms
PEOE_VSA_FHYD	fractional polar van der Waals surface area of hydrophobic atoms
PEOE_VSA_FPOL	fractional polar van der Waals surface area where absolute value of the atomic partial charge is >0.2[14,18]
PEOE_VSA_FNEG	fractional negative van der Waals surface area
PEOE_VSAFPOS	fractional positive van der Waals surface area
PEOE_VSA_FPPOS	fractional positive van der Waals surface area such that partial charge is greater than 0.2 divided by the total surface area
PEOE_VSA_FPNEG	fractional negative van der Waals surface area such that partial charge is less than -0.2 divided by the total surface area
PEOE_RPC+	the largest positive atomic partial charge divided by the positive sum
VAdjEq	function of a logarithm to the base two of adjacency map
VAdjMa	one plus the logarithm to the base two of the number of heavy-heavy bonds
VDistEq	sum of log base two distance matrix entries minus a function of distance matrix entries of a common value
VDistMa	sum of log base two distance matrix entries minus a function of shortest path lengths
zagreb	sum of squares of the number of heavy atoms bonded to each atom

they could be calculated from two-dimensional representations of molecules. This constraint was applied so that we would not bias the analysis by generation of hypothetical three-dimensional conformations of molecules in databases. Also excluded from the analysis were "binary" descriptors such as structural keys[12] (detecting the presence or absence of fragments in a molecule) that cannot be subjected to SE-DSE analysis because these descriptors have no information content that can be captured by SE analysis. Descriptors discussed in the Results and Discussion section are listed and defined in Table 1.

Calculations. The compound databases were implemented and manipulated with MOE,[13] version 2001.01. Descriptor values were also calculated with MOE. All SE-DSE calculations were carried out with Perl programs written by the authors.

RESULTS AND DISCUSSION

Determination of SE Threshold Values. Since we are interested in the identification of descriptors with distinctly different levels of information content, an important question is what SE values we should generally consider to be "high" or "low"? Can we develop a rationale for defining appropriate SE threshold values? To answer this question, we have calculated SE values for the 143 descriptors for the four databases studied (containing a total of approximately 328 000 diverse molecules). Of 572 possible SE values, 77 were zero. An SE value of zero means that a descriptor accounts for properties not present in database compounds or, alternatively, that all compound have the same descriptor value. Thus, a total of 495 (nonzero) SE values were subjected to graphical analysis. The resulting SE distribution

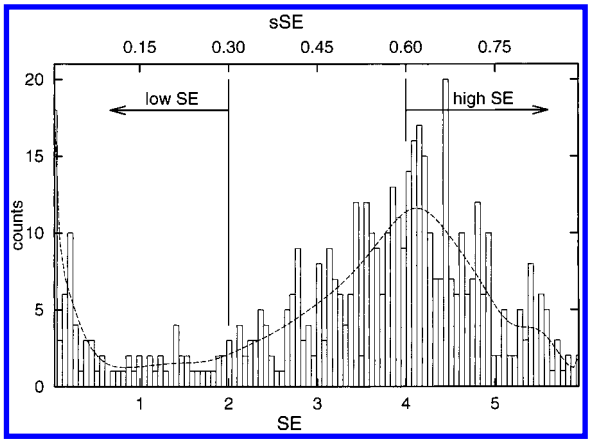

Figure 1. Histogram of 495 nonzero SE values spanning four compound databases. The bimodal nature of the distribution is emphasized with a dotted spline curve. The bottom abscissa is labeled with SE bin values for 100 bins. The top abscissa reports (bin-independent) sSE values, as defined in the Methods section. The sSE threshold values of 0.30 for "low" sSE and 0.60 for "high" SE are indicated.

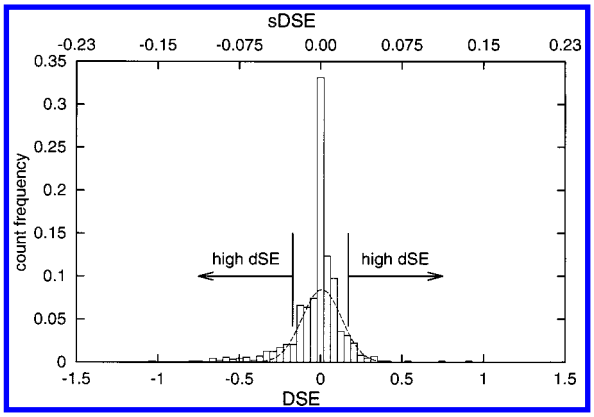

Figure 2. Histogram of 858 DSE values for 143 descriptors in all pairwise comparison of four compound databases. The top abscissa indicates the bin independent sDSE values. "High" DSE values (that can be positive or negative) are defined via the indicated one sigma limits of the fitted normal curve at approximately 0.171 DSE or 0.026 sDSE.

is shown in Figure 1. As can be seen, the distribution is essentially bimodal, with a number of descriptor having low SE values and a Gaussian-like curve toward high SE values. On the basis of these findings, we define an sSE value of 0.3 as the threshold for "low SE" and an sSE value of 0.6 as a threshold for "high SE", as indicated in Figure 1. sSE values between 0.3 and 0.6 are considered intermediate. These definitions are only based on the observed shape of the overall SE value distribution. Had we, for example, arbitrarily defined an sSE threshold value of 0.75 for "high SE", a number of descriptors would have been eliminated from the "high SE" category for each database comparison.

DSE Threshold Values. Similarly, DSE threshold values were determined. We have calculated DSE values for all 143 descriptors and pairwise comparison of the four databases, yielding a total of possible 858 values. The resulting DSE value distribution is shown in Figure 2. This distribution is distinct from the original SE distribution and displays a Gaussian tendency but does not conform exactly to a normal curve. Since the DSE calculation employs an averaging operation, this distribution results from deviations of mean values, and, therefore, the central limit theorem applies, consistent with the presence of a Gaussian-like distribution. DSE values can of course be negative or positive, considering the way they are calculated. In either case, the absolute value of DSE reflects the magnitude of differences in descriptor variability and value range distribution. Model fitting of a normal curve to the data of the distribution shown in Figure 2 results in a sigma value of 0.171 DSE units or 0.026 sDSE units. DSE values outside of this one sigma limit or standard deviation are defined as "high DSE" and values inside one sigma as "low DSE". Using this threshold value, the analysis revealed a total of 147 "high DSE" descriptor comparisons over all six database pairs, and these comparisons involve and represent 59 distinct molecular descriptors.

SE-DSE Analysis. On the basis of SE and DSE calculations and the corresponding threshold values, database comparisons of descriptor variability can essentially be divided into four "SE-DSE" categories: "high-high", "low-high", "high-low", and "low-low". For clarity, we exclude the descriptors with intermediate SE values from DSE calculations. For each compound database, descriptors with "high" and "low" SE values are first determined. For comparison and calculation of DSE values, a descriptor must belong to the same (either "high" or "low") category in both databases. Examples of histograms of descriptor distributions representing this classification scheme are shown in Figure 3. The histograms also illustrate the striking differences between descriptor distributions having high and low SE values, regardless of DSE. As one would expect, high information content correlates with the presence of distributions that can be graphically well represented and compared. For the purpose of our analysis, we are mainly interested in descriptors belonging to the "high-high" category because these descriptors have consistently high information content in diverse compound collections yet significantly different value distributions. Thus, these information-rich descriptors are most sensitive to systematic differences between synthetic, natural, and drug-like molecules and can be regarded as the best estimators of complementarity or diversity among those databases.

Descriptors with High SE and DSE Values. For each of the six pairwise database comparisons, the most variable descriptors are reported in Table 2. Only a total of 11 of the 143 descriptors evaluated herein fulfill our selection criteria. For each comparison, at least one descriptor has been identified, and the numbers of "high-high" descriptors vary between one and five, depending on the databases compared. Most descriptors sensitive to systematic differences (four or five) are observed for comparison of either synthetic or natural molecules with drug-like molecules or known drugs. The database comparisons reveal partially overlapping yet distinct descriptor sets. Most of the descriptors identified have complex designs (see also Table 1), a rather unexpected exception being "b_single", a descriptor simply counting the number of single bonds in a molecule. Interestingly, this descriptor responds to differences between drug-like molecules (CMC) and known drugs (SYNTH), which are principally similar databases. Another noteworthy observation is that only one molecular descriptor with high information content and significant DSE value was identified for the comparison of synthetic molecules and natural products, which have systematic chemical differences.[6,19] Molecular descriptors to distinguish these compound classes can be

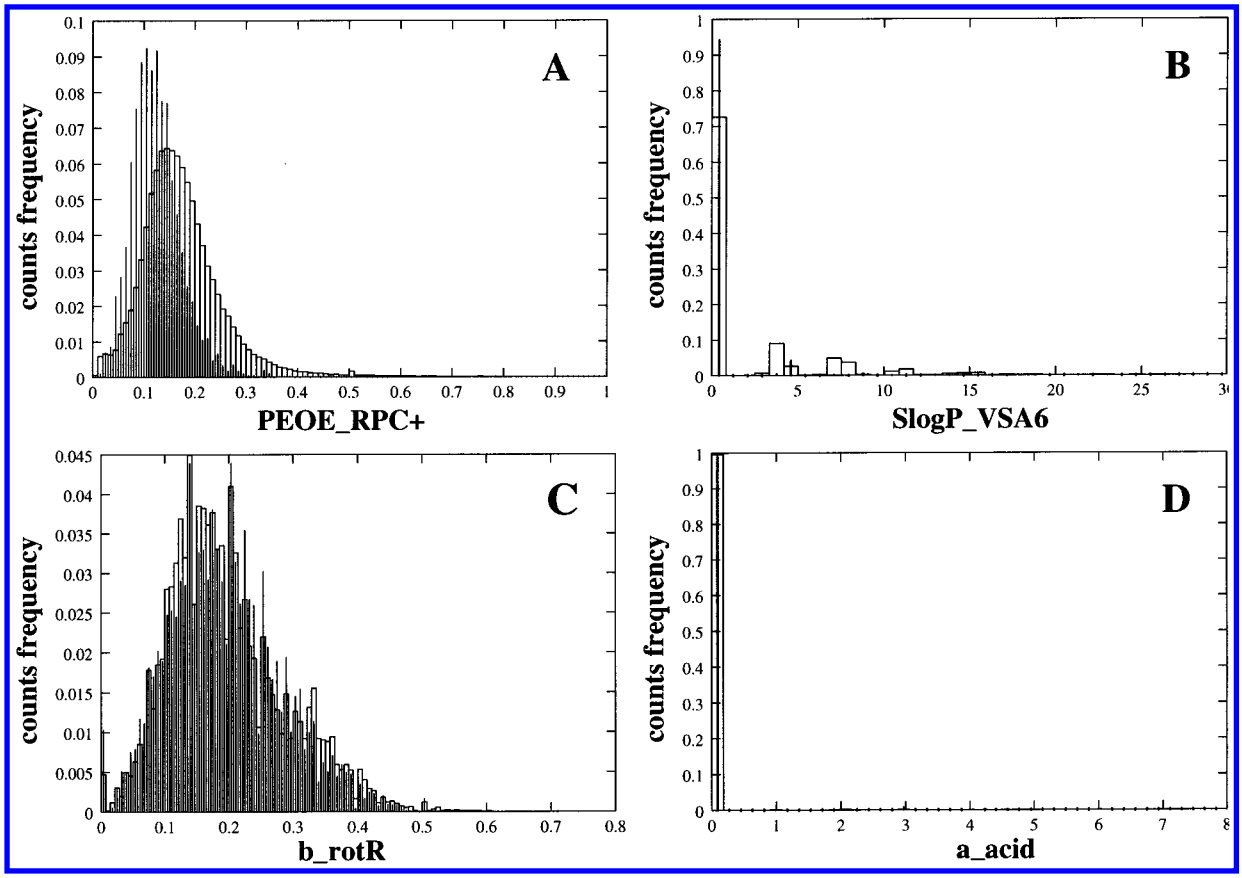

Figure 3. Representative histograms of descriptor distributions providing examples of four SE-DSE categories. Four different descriptors are shown. In each case, descriptor values are reported on the horizontal axis and their frequency of database occurrence on the vertical axis. Examples are (a) "high-high", descriptor "PEOE_RPC+" for the ACD (bars) versus SYNTH (lines) comparison, (b) "low-high", "SlogP_VSA6" for ACD (bars) versus C&H (lines), (c) "high-low", "b_rotR" for C&H (bars) versus CMC (lines), and (d) "low-low", "a_acid" for ACD (bars) versus C&H (lines). What appears to be a single valued bin in (d) is an artifact of a linear scale, since there are relatively small (~1:1000) yet occupied bins at positions 1.0, 2.0, 3.0....

Table 2. Descriptors Belonging to the "High-High" SE-DSE Category

database pair	descriptor	sDSE	sSE 1	sSE 2
ACD C&H	a_ICM	0.042	0.808	0.704
ACD CMC	PEOE_VSA+0	−0.027	0.634	0.697
ACD SYNTH	PEOE_RPC+	0.050	0.739	0.638
	PEOE_VSA+0	−0.032	0.634	0.701
	PEOE_VSA_FPOL	0.032	0.853	0.811
	a_ICM	0.033	0.808	0.741
	balabanJ	0.033	0.694	0.626
C&H CMC	SMR_VSA3	0.062	0.670	0.534
	SMR_VSA6	−0.050	0.582	0.703
	SlogP_VSA2	0.030	0.644	0.576
	weinerPol	0.035	0.727	0.655
C&H SYNTH	SMR_VSA6	−0.056	0.617	0.742
	SlogP_VSA7	−0.059	0.539	0.670
	b_single	0.029	0.689	0.629
	balabanJ	0.029	0.735	0.676
	weinerPol	0.036	0.727	0.653
CMC SYNTH	PEOE_VSA_FPOL	0.113	0.829	0.811
	b_single	0.029	0.652	0.549

identified but, as shown below, most of them have low SE values and thus limited information content. In general, molecular descriptors have varying complexity. For example, relatively simple descriptors may account for the number of specific atoms or bonds in a molecule, whereas more complex descriptor often combine information provided by two or more other descriptors. Recurrent among complex descriptors belonging to the "high-high" category is a new class of descriptors recently introduced by Labute.[14] These descriptors (designated *_VSA*) map diverse atomic properties on molecular surface areas approximated from 2D representations of molecules (and are thus best rationalized as implicit 3D descriptors).

Descriptors with Low SE and High DSE Values. For comparison, Table 3 lists descriptors belonging to the "low-high" category. These descriptors have limited variability within the databases but account for distinct differences between them. Table 3 shows that considerably more descriptors can be identified that belong to this than the "high-high" category, albeit only for five of six comparisons (except CMC and SYNTH, as further discussed below). About half of these "low-high" descriptors are simple atom (halogen or sulfur) count descriptors. For example, in this case, seven descriptors are responsive to systematic differences between synthetic compounds (ACD) and natural products (CH), two of which simply detect halogen atoms that rarely occur in natural molecules. A descriptor accounting for nitrogen content, another known principal difference between synthetic and natural molecules,[19] does not occur in this limited list because it falls into the intermediate SE value range (according to Figure 1) in both databases. However, its DSE value for the ACD-C&H comparison is high, thus reflecting general differences in molecular composition. Taken together, these findings make an important point. Compounds with systematic chemical differences can be distinguished by finding descriptors that detect only a single yet highly class-specific feature. In an extreme case,

Table 3. Descriptors Belonging to the "Low-High" SE-DSE Category

database pair	descriptor	sDSE	sSE 1	sSE 2
ACD C&H	PEOE_VSA-3	0.038	0.218	0.024
	PEOE_VSA-4	0.039	0.275	0.071
	SlogP_VSA6	0.041	0.259	0.062
	a_nCl	0.033	0.181	0.024
	a_nF	0.029	0.144	0.002
	a_nS	0.030	0.166	0.035
	vsa_don	0.035	0.290	0.116
ACD CMC	PEOE_VSA-3	0.050	0.218	0.111
	SlogP_VSA6	0.038	0.257	0.178
	a_nF	0.029	0.144	0.084
ACD SYNTH	PEOE_VSA-3	0.030	0.218	0.157
	a_base	−0.038	0.011	0.090
	a_nCl	0.030	0.181	0.117
	vsa_base	−0.045	0.006	0.102
C&H CMC	PEOE_VSA+6	−0.069	0.069	0.232
	PEOE_VSA-2	−0.071	0.068	0.236
	PEOE_VSA-3	−0.050	0.038	0.154
	SlogP_VSA6	−0.048	0.062	0.179
	a_nCl	−0.044	0.024	0.131
	a_nF	−0.033	0.002	0.084
	a_nS	−0.047	0.035	0.149
	weinerPath	0.027	0.227	0.166
C&H SYNTH	PEOE_VSA+6	−0.098	0.081	0.295
	PEOE_VSA-2	−0.051	0.069	0.182
	PEOE_VSA-3	−0.084	0.032	0.217
	SlogP_VSA6	−0.065	0.062	0.206
	a_base	−0.029	0.027	0.090
	a_nCl	−0.042	0.024	0.117
	a_nF	−0.057	0.002	0.129
	a_nS	−0.060	0.035	0.169
	vsa_base	−0.044	0.005	0.102

these would be, for example, binary descriptors detecting the presence or absence of this one molecular feature or property. The problem with applying such descriptors for database analysis and comparison is, however, that they have little, if any, information content within each database. Thus, they are difficult, if not impossible, to apply if differences between databases are subtle. This is demonstrated, for example, by the fact that no descriptor of the "low-high" category can be identified to respond to differences between CMC and SYNTH compounds (i.e., this combination is absent in Table 3), which are drug-like molecules or drugs and thus very similar. By contrast, DSE calculations using descriptors with high information content can readily detect differences between database distributions of chemical features, even if the differences are subtle.

Descriptors with High SE But Low DSE Values. Another important question has been whether descriptors with high information content always have the tendency to detect differences between the compound classes studied here. SE-DSE analysis showed that this is clearly not the case. Many information-rich descriptors do not measurably respond to relevant chemical differences. In fact, as shown in Table 4a, 24 descriptors with generally high SE values have consistently low DSE values in all six database comparisons. This descriptor set also includes a number of the previously mentioned surface descriptors,[14] suggesting that they are diverse with regard to the compound-class specific information they capture. These findings demonstrate that only a relatively small subset of descriptors with high information content has the general ability to detect systematic chemical differences in large databases. Since they are information-rich, a combination of "high-low" descriptors

Table 4

descriptor	highest sDSE (lowest 0.00)
a. Descriptors with Consistently "High" SE Values in All Databases and "Low" DSE Values for All Database Comparisons	
PEOE_VSA_FPNEG	0.003
PEOE_VSA_NEG	0.006
PEOE_VSA_FHYD	0.012
SMR_VSA5	0.012
a_hyd	0.012
chi1_C	0.012
b_rotR	0.014
a_nC	0.015
b_1rotR	0.015
PEOE_VSA_HYD	0.017
VAdjEq	0.017
chi0_C	0.017
chi0v_C	0.018
PEOE_VSA_FPPOS	0.020
VDistEq	0.020
VDistMa	0.020
b_heavy	0.020
chi1v_C	0.020
vsa_hyd	0.020
VAdjMa	0.021
PEOE_VSA+1	0.023
zagreb	0.024
PEOE_VSA_FNEG	0.026
PEOE_VSA_FPOS	0.026
b. Descriptors with Consistently "Low" SE and "Low" DSE Values	
a_nBr	0.012
a_nP	0.012
a_nI	0.014
b_triple	0.014

should nevertheless be useful for a number of computational applications, for example, the assessment of molecular diversity.[20] In addition, we have also identified a small set of descriptors that have a generally limited database variability and little, if any, discriminatory power for the compound collection analyzed. For comparison, these "low-low" descriptors are shown in Table 4b. Among these is a descriptor accounting for the number of triple bonds in a molecule. Thus, the distribution of triple bonds, which are relatively rare, does not show database differences detectable by SE-DSE analysis, and, thus, the effort to calculate such descriptors may be avoided.

Alternative Entropic Formulations. In information theory, other entropy-based concepts have been introduced[21,22] including, for example, the Kullback-Leibler function[22,23] that measures the similarity between a statistical model and a true data distribution. However, this function is not suitable for the comparison of descriptor distributions reported herein, for two reasons. The Kullback-Leibler function is asymmetric with regard to the compared distributions, and, thus, its value depends on which of the two data distributions is considered the "true" one. By contrast, the DSE formalism quantifies the difference in information content of descriptor distributions without the need to assign a reference distribution and produces the same value for either pairwise comparison. Furthermore, the Kullback-Leibler function is not defined if the model distribution has zero probability in a data interval, which we frequently observe in descriptor analysis.

Conclusions. In this study, we have introduced SE-DSE calculations to systematically study and compare the distributions of molecular descriptors in selected compound

databases. Our findings suggest that SE-DSE analysis is a robust and sensitive method to quantify differences in such database distributions of numerical descriptors, regardless of the chemical characteristics they capture. Attractive features of the approach are its conceptual simplicity and computational efficiency. Thus, the method can effectively evaluate very large numbers of descriptors and molecules. Although our current analysis has only considered descriptors that can be calculated from two-dimensional representations of molecules, SE-DSE analysis can be readily applied to study database statistics of three-dimensional descriptors. Descriptors with significant information content and large DSE values are most likely to differentiate compounds within specific collections and, at the same time, account for features that distinguish them from other classes of compound. However, only about 8% of the relatively large number of descriptors evaluated in this study fulfill these requirements (i.e., belong to the "high-high" SE-DSE category). By contrast, many information-rich descriptors are not responsive to database-specific differences. Regardless of observed DSE values, the intrinsic information content of molecular descriptors approximately correlates with the complexity of their design. Among the most complex descriptors are those that combine information provided by two or more other descriptors, for example, calculated molecular surface area and atomic partial charges. However, as we have demonstrated, only a fraction of these descriptors show significant DSE values in database comparisons, and, as also shown, not all descriptors belonging to the "high-high" SE-DSE category are particularly complex. These findings illustrate the complementary nature of SE and DSE calculations in descriptor analysis. Without these calculations, it would have been difficult, if not impossible, to predict which of the many descriptors studied here capture the most information and are sensitive to compound class-specific differences.

REFERENCES AND NOTES

(1) Xue, L.; Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Combin. Chem. High Throughput Screen.* **2000**, *3*, 363−372.

(2) Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233−245.

(3) Livingstone, D. J. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.

(4) Shannon, C. E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, IL, U.SA., 1963.

(5) Godden, J. W.; Stahura, F. L.; Bajorath, J. Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 796−800.

(6) Stahura, F. L.; Godden, J. W.; Xue, L.; Bajorath, J. Distinguishing between natural products and synthetic molecules by Shannon descriptor entropy analysis and binary QSAR calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1245−1252.

(7) Godden, J. W.; Bajorath, J. Differential Shannon entropy as a sensitive measure of differences in database variability of molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1060−1066.

(8) Available Chemicals Directory, MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577.

(9) Chapman & Hall Dictionary of Natural Products, CD-ROM version 1999; CRC Press LLC: Boca Raton, FL, USA.

(10) Comprehensive Medicinal Chemistry Database, version 99.1; MDL Information Systems, Inc.: San Leandro, CA 94577.

(11) Synthline Drug Database on STN International, taken from a compendium of a Prous Science Journal, Drugs of the Future (comprehensive drug monographs), since 1984; Prous Science: Barcelona, Spain.

(12) McGregor, M. J.; Pallai, P. V. Clustering of large databases of compounds: Using MDL "keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.

(13) MOE (Molecular Operating Environment); Chemical Computing Group Inc.: Montreal, Quebec, Canada.

(14) Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model.* **2000**, *18*, 464−477.

(15) Wildman, S. A.; Crippen, G. M. Prediction of physiochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.

(16) Balaban, A. T. Five new topological indices for the branching of tree-like graphs. *Theor. Chim. Acta* **1979**, *53*, 355−375.

(17) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

(18) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(19) Henkel, T.; Brunne, R. M.; Müller, H.; Reichel, F.; Statistical investigation into the structural complementarity of natural products and synthetic compounds. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 643−647.

(20) Martin, Y. C. Diverse viewpoints on computational aspects of molecular diversity. *J. Comb. Chem.* **2001**, *3*, 1−20.

(21) Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **2000**, *85*, 461−464.

(22) Cover, T. M.; Joy, A. T. *Elements of Information Theory*; John Wiley and Sons: New York, NY, U.S.A., 1991.

(23) Kullback, S. *Information Theory and Statistics*; Dover Publications: Mineola, NY, U.S.A., 1997.

CI0103065