Differential Shannon Entropy as a Sensitive Measure of Differences in Database Variability of Molecular Descriptors

Jeffrey W. Godden† and Jürgen Bajorath*,‡

New Chemical Entities, Inc., 18804 North Creek Parkway, Bothell, Washington 98011, and New Chemical Entities and Department of Biological Structure, University of Washington, Seattle, Washington 98195

Received February 13, 2001

A method termed Differential Shannon Entropy (DSE) is introduced to compare differences in information content and variance of molecular descriptors between compound databases. The analysis is based on histograms recording the individual and grouped distributions of molecular descriptors and calculation of Shannon entropy (SE), a formalism originally applied to digital communication. We have recently shown that SE values reflect the nonparametric variability of descriptor settings. Now the analysis has been advanced to assess differences in information content of 143 molecular descriptors in databases containing synthetic compounds, natural products, or drug-like molecules. The DSE metric captures the degree to which descriptor distributions complement or duplicate information contained in molecular databases. In our analysis, we observe significant differences for a number of descriptors and rank them according to their associated DSE values. Using DSE calculations, relative information content of different types of descriptors can be quantified, even if differences are subtle.

INTRODUCTION

Molecular descriptors typically account for physicochemical properties, molecular topology or connectivity, conformational parameters, or structural fragments. Descriptor-based representations project molecules into abstract chemical space(s) to capture and compare their structural features and properties. However designed, such representations are predominantly employed to analyze molecular similarity, describe the diversity of libraries, cluster or partition molecules, or study structure−activity relationships and drug-like properties.[1−11] The performance of descriptor sets for specific tasks such as, for example, representative subset selection or diversity analysis has been evaluated in a number of case studies.[3−8] When encoded as binary molecular fingerprints, combinations of molecular descriptors are often used for similarity searching and database mining.[11−16] Hundreds of descriptors have been reported over time, often for very different purposes, which account in different ways for diverse molecular features or properties. Thus, the molecular information captured by these descriptors is in general very difficult to compare, and, consequently, descriptors for specific applications are frequently selected on the basis of chemical intuition, rather than systematic analysis.

We intended to analyze the distribution of molecular descriptors in large compound data sets to provide a basis for rational descriptor selection for specific applications. However, this effort was complicated by the fact that distributions of many descriptors cannot be subjected to conventional statistical analysis because their units and value ranges differ, and their distributions may significantly depart from a normal curve. For example, "logP" presents a continuum of values, whereas a descriptor counting the number of "aromatic bonds" in a molecule typically adopts a narrow range of discrete values. To design a metric that could reduce molecular descriptions to their information content, we have focused on the Shannon Entropy (SE)[17] concept. SE provides a connection between entropy and information content and was originally applied in digital communication technology to determine the amount of data that could be transmitted given a range of frequencies.[17]

Since SE calculations can in principle reduce very different kinds of data to their information content, we have adapted the SE concept to analyze the variability of molecular descriptors in compound databases.[18] In initial studies, we have been able to calculate and, to some extent, compare the variability of different descriptors[18] and shown that it is possible to select some compound class-specific descriptor sets that can be used to distinguish between molecules from different sources.[19] However, as further discussed below, we have also determined that calculation of SE values alone may not be sufficient to select descriptors with significant discriminatory power.[19] We have therefore developed an extension of the approach, called Differential Shannon Entropy (DSE), which takes into account both differences in the variability and value range distributions of compared descriptors.

METHODS

Shannon entropy is defined as[17]

$$\mathrm{SE} = -\sum p_i \log_2 p_i \qquad (1)$$

In this formulation, p is the probability of a data point or "count" c to adopt a value within a specific data interval i. Thus, p is calculated as

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

* Corresponding author phone: (425)424-7297; fax: (425)424-7299; e-mail: jbajorath@nce-mail.com. Mailing address: 18804 North Creek Pkwy, Bothell, WA 98011-8012.

† New Chemical Entities, Inc.

‡ New Chemical Entities and University of Washington.

10.1021/ci0102867 CCC: $20.00 © 2001 American Chemical Society
Published on Web 06/06/2001

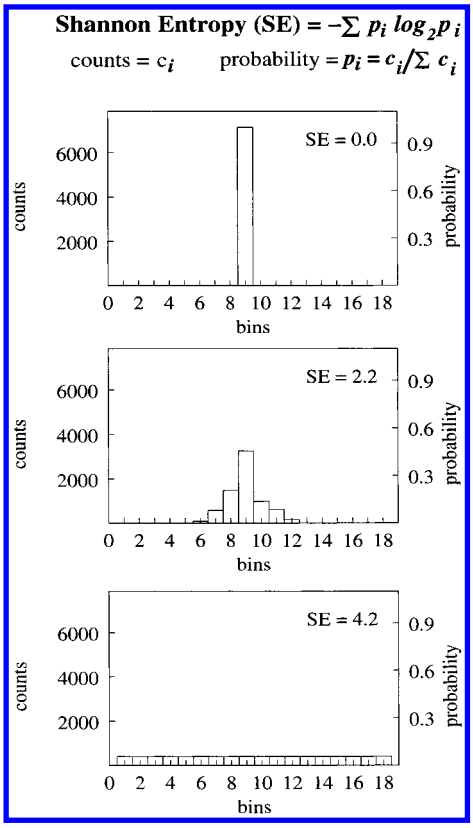


Figure 1. Shannon entropy calculations. For example histograms with 18 bins SE can range from a minimum of 0 (top) to a maximum of approximately (bottom) 4.2 (logarithm to the base 2 of 18). The middle graph shows a data distribution more typically observed for molecular descriptor settings in compound databases. As illustrated in the graph at the bottom, maximum SE is achieved when the same probability of data counts is observed for each interval. For all subsequent histograms generated and analyzed in this study the number of bins is consistently set to 100, which permits a maximum SE value of 6.64.

Equation 1 converts probability to Shannon entropy and contains a logarithm to the base 2. This scale factor permits SE to be considered as a metric of digital information content. For our purposes, a key feature of this concept is that probabilities and, in turn, SE values can be calculated and compared for any set of data as long as the data representation is uniform, regardless of units and value ranges. This is the case when data sets are represented in histograms with data ranges consistently divided into the same number of bins. As long as the number of data intervals between the minimum and maximum value is held constant, SE values are independent of the size of the interval. Thus, if reduced to their information content, very different types of molecular descriptors can be compared, regardless of whether they adopt discrete values or a continuum of values. For comparison of descriptor variability between different compound databases, the absolute number of bins in a histogram is not critical as long as it is consistent in all calculations. Of course, choosing very small or exceedingly large bin numbers will not produce reasonable SE values for comparison. If a descriptor only accounts for a relatively small number of discrete values (for example, between zero and 10 rotatable bonds in a molecule), many bins in a histogram consisting of 100 bins will not be occupied. Figure 1 shows how SE values are calculated from histogram representations of hypothetical distributions of discrete or continuous descriptor values. It also illustrates that SE values may encompass a range of values from zero to a maximum value, which corresponds to the logarithm to the base 2 of the number of histogram bins. Thus, maximum SE values depend on the number of bins selected for data representation. It follows that the information content of different descriptor distributions can only be directly compared if the applied binning scheme is consistent. An SE value of zero, corresponding to a single bin probability of one, indicates that a descriptor adopts only one value and has thus no information content with respect to the data set. As shown in Figure 1, the maximum SE value will be observed if all possible descriptor

Table 1. Definition of Selected Molecular Descriptors

name	description
SlogP_VSA1	approx. vdW atomic surface with $-0.4 < \log P \leq -0.2$[24]
SlogP_VSA2	approx. vdW atomic surface with $-0.2 < \log P \leq 0.0$
SlogP_VSA4	approx. vdW atomic surface with $0.1 < \log P \leq 0.15$
SlogP_VSA6	approx. vdW atomic surface with $0.2 < \log P \leq 0.25$
SlogP_VSA8	approx. vdW atomic surface with $0.3 < \log P \leq 0.4$
SMR_VSA3	approx. vdW atomic surface where molar refractivity is $0.35 < R_i \leq 0.39$[24, 25]
SMR_VSA4	approx. vdW atomic surface where molar refractivity is $0.39 < R_i \leq 0.44$
a_aro	number of aromatic atoms
a_ICM	entropy of the element distribution in the molecule
a_nN	number of nitrogen atoms
a_nO	number of oxygen atoms
a_nF	number of fluorine atoms
a_nCl	number of chlorine atoms
b_ar	number of aromatic bonds
density	molecular mass density; MW divided by an approx. vdW volume
balabanJ	Balaban's connectivity topological index[26]
weinerPol	half the sum of all the distance matrix entries with a value of 3[27]
a_acc	number of hydrogen bond acceptor atoms (counting OH)
vsa_don	approx. vdW surface area of H-bond donors (not counting OH)
vsa_pol	approx. vdW surface area of H-bond donors (counting OH)
PEOE_VSA+4	vdW surface area where atomic partial charge $0.2 \leq q < 0.25$
PEOE_VSA+2	vdW surface area where atomic partial charge $0.1 \leq q < 0.15$
PEOE_VSA-3	vdW surface area where atomic partial charge $-0.2 \leq q < -0.15$
PEOE_VSA-4	vdW surface area where atomic partial charge $-0.25 \leq q < -0.2$
PEOE_VSA-5	vdW surface area where atomic partial charge $-0.3 \leq q < -0.25$
PEOE_VSA-6	vdW surface area where atomic partial charge less than -0.3
PEOE_RPC+	the largest positive atomic partial charge divided by the positive sum

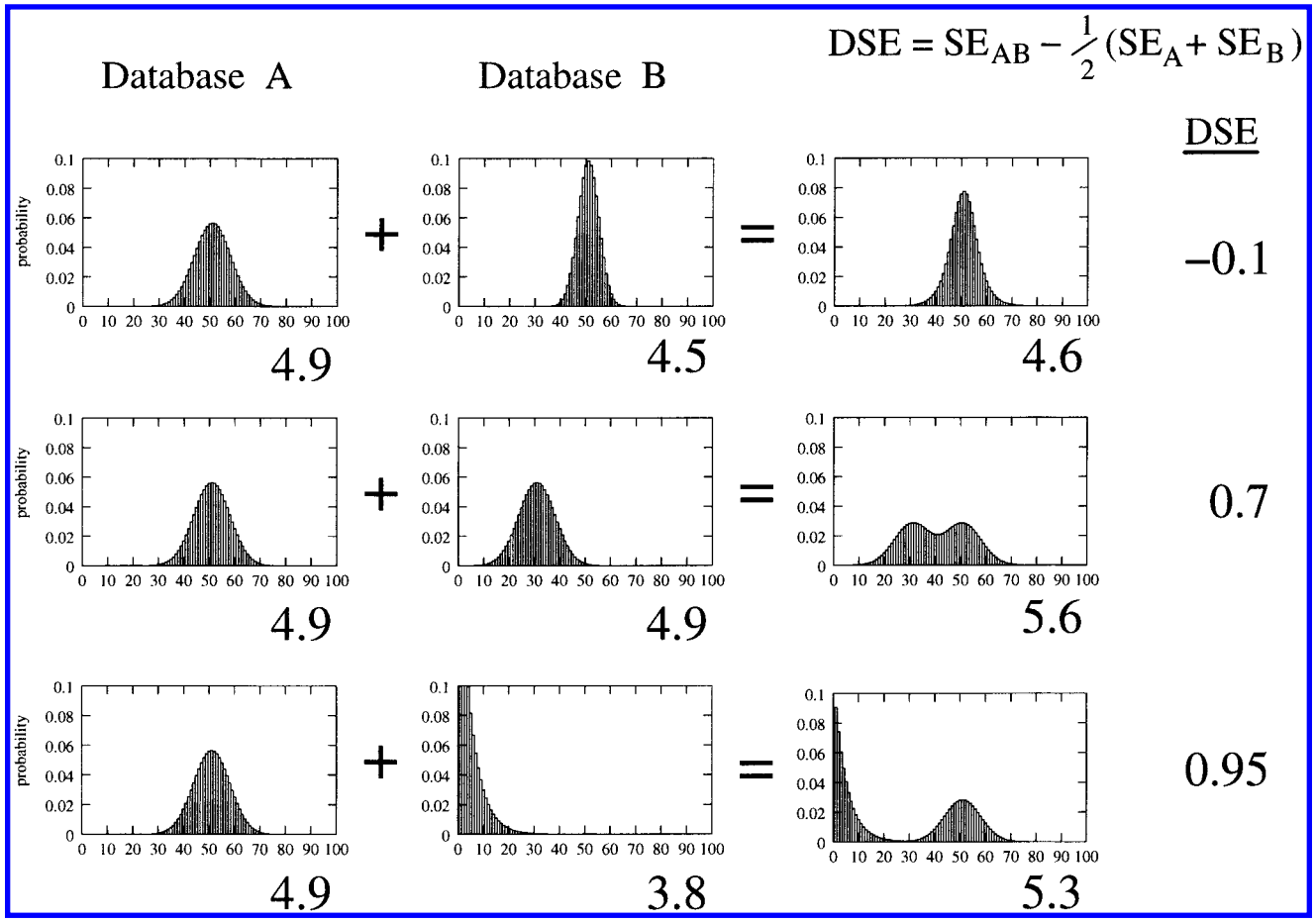


Figure 2. DSE calculations. Hypothetical descriptor distributions in two compound databases A and B that may combine to increase or decrease the Shannon entropy of the resulting probability histogram. DSE is the difference between the renormalized histogram of both databases binned together and the average of the their independent histograms. If the envelope of the combined distribution (graphs on the right) shows no increased data variability, a negative DSE value may be observed (top). If data distributions are distinct, positive DSE values are observed, even if differences are subtle.

values are equally probable, which corresponds to maximum information content.

SE values capture both the intrinsic variability of molecular descriptors and their extrinsic variability with respect to a given data set (i.e., a collection of molecules). Since we are particularly interested in the identification of descriptors that are sensitive to compound class-specific features, we need to compare the variability of descriptors in different compound databases. As will be rationalized in this study, a simple calculation of differences in SE values of molecular descriptors is insufficient to achieve these ends. We therefore introduce a new metric, differential Shannon entropy (DSE), which extends the SE concept specifically for comparative analysis of different molecular descriptors. Differential Shannon entropy is defined as

$$\mathrm{DSE} = \mathrm{SE}_{AB} - (\mathrm{SE}_A + \mathrm{SE}_B)/2 \qquad (3)$$

"SE_{AB}" refers to the Shannon entropy calculated from a single histogram reflecting the distribution of the entire population of compounds from both databases. The terms "SE_A" and "SE_B" represent the SE values of each of the two databases when considered individually. DSE can be rationalized as an increase or decrease in descriptor variability due to complementary features or synergy in information content of the compared databases. For comparison, we also discuss "entropic separation", as introduced previously.[19]

Entropic separation is defined as

$$\mathrm{ES} = |M_A - M_B|/((\mathrm{SE}_A + \mathrm{SE}_B)/4) \qquad (4)$$

This equation accounts for the absolute value of the number of histogram bins separating the modes (M) of two descriptor distributions (the bin number where highest counts are observed) divided by half of the average SE values of these distributions. This statistical metric was defined to account for the difference between two descriptor populations without the need to assume the presence of a parametric model.

Analysis of descriptor distributions and SE and DSE calculations were performed with Perl programs written by the authors. DSE analysis was carried out for a total of 143 molecular descriptors available in the Molecular Operating Environment (MOE, version 2000.02)[20] and three different compound databases, the Available Chemicals Directory (ACD),[21] the Comprehensive Medicinal Chemistry (CMC),[22] and Chapman & Hall (C&H)[23] databases. ACD contains many organic compounds often used as reagents, CMC consists of molecules with drug-like properties, and C&H is a compendium of natural products. A total of 199 420 ACD, 116 364 C&H, and 7580 CMC compounds were used. A comparison of descriptor SE values between ACD and the publicly available database of the National Cancer Institute (NCI) has been reported in our previous publication.[18] Descriptors analyzed here include bulk properties,

physicochemical parameters, atom and bond counts, and topology, surface, and shape descriptors. Values were calculated only for single compounds, excluding noncovalent complexes. Table 1 lists and explains descriptors discussed in this study. Histograms of descriptor distributions were uniformly generated using 100 data intervals, which produced graphically meaningful data dispersion. For 100 bins used here, the theoretical maximum SE value for even data dispersion over all bins would be approximately 6.64.

RESULTS AND DISCUSSION

The DSE Concept. The underlying idea of DSE calculations is that combinations of descriptor distributions in histograms with consistent binning scheme are not the sum of single distributions. Combining distributions requires renormalization of the data over a constant number of intervals and thus generates a new envelope for data representation. As illustrated in Figure 2, DSE is the difference between the renormalized histogram of both distributions and the average of the their independent histograms. If the envelope of the combined distributions shows increased spread or variability, a positive DSE is observed. In contrast, if the envelope shows no increased variability, a negative DSE value may be observed. Thus, even subtle differences in distributions and their value ranges can be quantified.

Application of DSE Calculations. DSE analysis makes it possible to identify descriptors that are sensitive to systematic differences in properties and/or composition of molecules belonging to different classes or databases. Such descriptors must have significant database variability and also differences in the distribution of their values.[19] These descriptors capture compound class-specific features and can be used to effectively distinguish between compounds with different characteristics.[19] Therefore, the identification of suitable descriptors depends on the assessment of two factors, their variability and value range distributions. While descriptor variability directly correlates with calculated SE values,[18] differences in value ranges are more difficult to quantify. This is illustrated in Figure 3 for hypothetical descriptor distributions. In both cases, the descriptor has significant variability and also differences in its value ranges. Thus, it displays the features discussed above. In the first case shown, the value ranges are quite distinct. Accordingly, both entropic separation and DSE can serve as measure of distribution difference. However, in the second case, the distributions overlap more closely, their modes coincide, and entropic separation becomes zero. By contrast, DSE quantifies differences in variability and distribution of values in both cases.

DSE Analysis of Molecular Descriptors. We have calculated and compared SE and DSE values for a total of 143 descriptors in three major compound databases containing synthetic compounds (ACD), drug-like molecules (CMC), and natural products (C&H) to identify those descriptors that are most sensitive to systematic differences in these compound collections. Pairwise comparisons between these three databases were carried out. Descriptor distributions resulting in largest DSE values are reported in Figure 4. As can be seen, DSE values can be readily calculated and compared for very different types of descriptors, regardless of whether their distributions are continuous or discrete. Table 2 lists

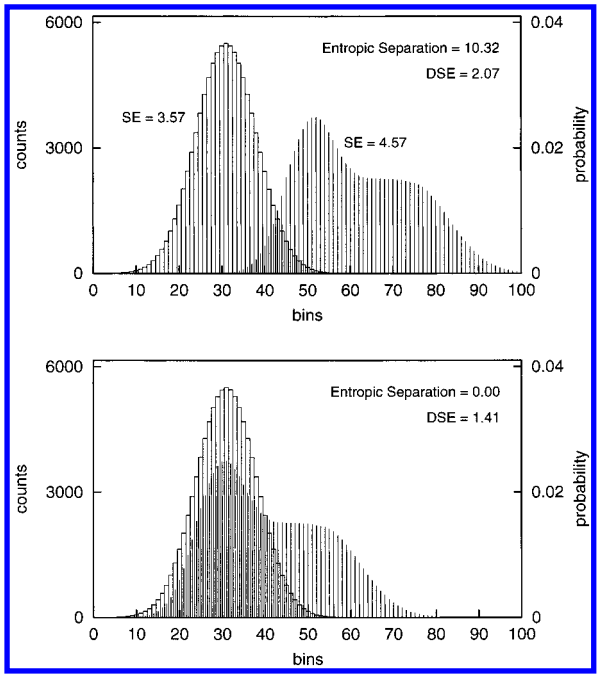
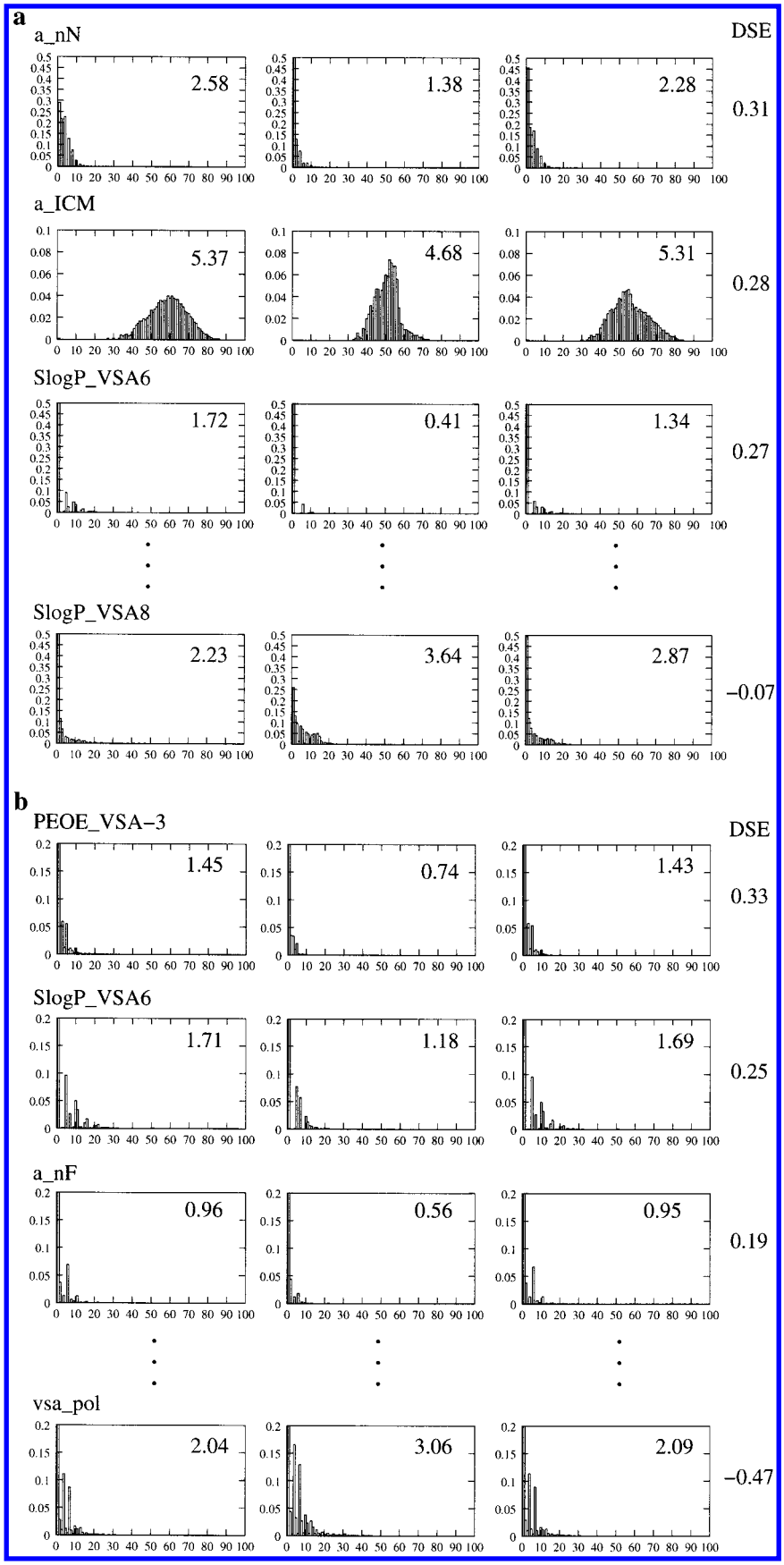


Figure 3. Entropic separation versus DSE. Shown are histograms of theoretical descriptor distributions in two compound databases. In both cases, the distributions are equivalent and have significant variability, as reflected by SE calculation, but their value ranges differ. In the top graph, the peak (or mode) bins are well separated, and calculation of entropic separation and DSE values indicates significant differences in value ranges. Such differences are never taken into account when only differences between absolute SE values are calculated. The bottom graph shows a situation where the distributions are overlapping yet distinct. In this case, entropic separation can no longer be used as a measure of differences in value range distribution. By contrast, DSE values detect the distribution difference and quantify the greater overlap of the distributions compared to the first case.

the top 10 descriptors with largest DSE values for each database comparison. These descriptors have the most complementary values, which results in the greatest augmented variability. Some chemically intuitive trends can be observed. For example, on average, DSE values are smaller for ACD/CMC comparison than for comparison of these databases with natural products (C&H), thus indicating that many natural products are chemically more distinct from synthetic compounds than drug-like molecules (mostly produced by medicinal chemistry). Similarly, a descriptor accounting for nitrogen atoms displays the largest DSE value for the ACD/C&H comparison, which reflects the prevalence of amide chemistry in synthetic compounds compared to naturally occurring molecules. Also evident are differences in aromatic character of synthetic and natural molecules. On the other hand, halogen content is a distinguishing factor for many ACD and CMC compounds. Table 2 also shows that several recently introduced descriptors (*_VSA*)[24] repeatedly occur in the top 10 list of all three comparisons. These descriptors map different chemical properties on deduced surface representations of molecules,[24] and our analysis suggests that these descriptors are among the ones most sensitive to differences in chemical information content.

CONCLUSIONS

The Shannon entropy concept reduces data distributions to their information content and, in our implementation,


a
a_nN
DSE
2.58
1.38
2.28
0.31
a_ICM
5.37
4.68
5.31
0.28
SlogP_VSA6
1.72
0.41
1.34
0.27
SlogP_VSA8
2.23
3.64
2.87
−0.07
b
PEOE_VSA−3
DSE
1.45
0.74
1.43
0.33
SlogP_VSA6
1.71
1.18
1.69
0.25
a_nF
0.96
0.56
0.95
0.19
vsa_pol
2.04
3.06
2.09
−0.47

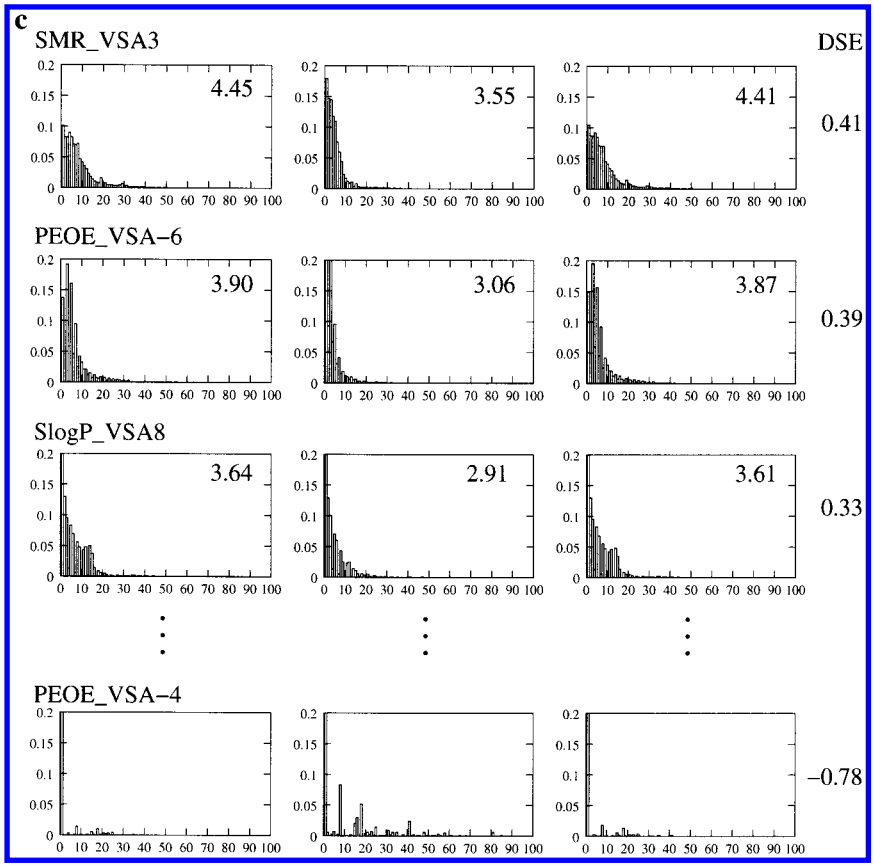


Figure 4. DSE analysis of descriptors and different compound databases. For pairwise database comparison, descriptors whose distributions produce the three largest DSE values are shown followed by the one yielding the lowest. In each case, a total of 143 descriptors was analyzed. Histograms are reported according to Figure 2, with separate databases in the two columns on the left and in the middle and combined distributions on the right. Resulting DSE values are reported: (a) ACD versus C&H, (b) ACD versus CMC, and (c) C&H versus CMC.

Table 2. Descriptors with Largest DSE Values in Database Comparison

ACD/C&H		ACD/CMC		C&H/CMC	
a_nN	0.31	PEOE_VSA-3	0.33	SMR_VSA3	0.41
a_ICM	0.28	SlogP_VSA6	0.25	PEOE_VSA-6	0.39
SlogP_VSA6	0.27	a_nF	0.19	SlogP_VSA8	0.33
PEOE_VSA-4	0.26	PEOE_VSA+2	0.19	weinerPol	0.23
SMR_VSA4	0.25	density	0.16	SlogP_VSA4	0.22
PEOE_VSA-3	0.25	a_nCl	0.15	vsa_pol	0.21
vsa_don	0.23	a_ICM	0.12	a_nO	0.21
a_aro	0.23	SMR_VSA4	0.11	SlogP_VSA2	0.20
SlogP_VSA1	0.23	PEOE_RPC+	0.11	a_acc	0.19
b_ar	0.22	balabanJ	0.09	PEOE_VSA+4	0.19

provides a basis for the evaluation of database variability of molecular descriptors with different units and value ranges (that could not be directly compared). The DSE formalism presented here extends the SE concept by quantitatively taking differences in value range distributions into account. It introduces a value range-dependence of the information content that is, as shown here, an important aspect in descriptor evaluation and selection. We have demonstrated that DSE analysis is capable of identifying descriptors with significant differences in major compound databases. Following the extended SE approach, descriptors most sensitive to systematic differences in chemical features between compound collections are those having large SE and DSE values. However, DSE calculations can also detect subtle differences in descriptor distributions that would otherwise be difficult to recognize and quantify.

ACKNOWLEDGMENT

The authors thank Ling Xue for help in the assembly and calculation of molecular descriptors.

REFERENCES AND NOTES

(1) Brown, R. D. Descriptors for diversity analysis. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 31−49.

(2) Bures, M. G.; Martin, Y. C. Computational methods in molecular diversity and combinatorial chemistry. *Curr. Opin. Chem. Biol.* **1998**, *2*, 376−380.

(3) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D molecular descriptors relevant to ligand−receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731−740.

(4) 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.

(5) Matter, H. Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.

(6) Matter, H.; Pötter, T. Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211−1225.

(7) Xue, L.; Godden, J.; Gao, H.; Bajorath, J. Identification of a preferred set of molecular descriptors for compound classification based on principal component analysis. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 669−704.

(8) Rusinko, A., III.; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of large structure/biological activity data sets using recursive partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.

(9) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.

(10) Ajay; Walters, P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.
(11) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical similarity using geometric pair descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128−136.
(12) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(13) James, C. A.; Weininger, D. *Daylight fingerprints. Daylight theory manual*; Daylight Chemical Information Systems, Inc.: Irvine, CA, 1995.
(14) Xue, L.; Godden, J.; Bajorath, J. Database searching for compounds with similar biological activity using short binary bit string representations of molecules. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 881−886.
(15) Xue, L.; Godden, J.; Bajorath, J. Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1227−1234.
(16) Mason, J. S.; Cheney, D. L. Library design and virtual screening using multiple point pharmacophore fingerprints. *Pac. Symp. Biocomput.* **2000**, *5*, 576−587.
(17) Shannon, C. E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, U.S., 1963.
(18) 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.
(19) 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.
(20) MOE (Molecular Operating Environment); Chemical Computing Group Inc.: 1255 University Street, Montreal, Quebec, Canada, H3B 3X3.
(21) *Available Chemicals Directory*; MDL Information Systems, Inc.: 14600 Catalina Street, San Leandro, CA 94577.
(22) *Comprehensive Medicinal Chemistry Database, version 99.1*; MDL Information Systems, Inc.; 14600 Catalina Street, San Leandro, CA 94577.
(23) *Chapman & Hall Dictionary of Natural Products, CD-ROM version 1999*; CRC Press LLC: 2000 NW Corporate Blvd., Boca Raton, FL 33431, USA.
(24) Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model.* **2000**, *18*, 464−477.
(25) Wildman, S. A.; Crippen, G. M. Prediction of physiochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(26) Balaban, A. T. Five new topological indices for the branching of tree-like graphs. *Theor. Chim. Acta* **1979**, *53*, 355−375.
(27) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

CI0102867