Assessment of Molecular Similarity from the Analysis of Randomly Generated Structural Fragment Populations

José Batista, Jeffrey W. Godden, and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, Germany

Received April 5, 2006

A novel method termed MolBlaster is introduced for the evaluation of molecular similarity relationships on the basis of randomly generated fragment populations. Our motivation has been to develop a similarity method that does not depend on the use of predefined structural or property descriptors. Fragment profiles of molecules are generated by random deletion of bonds in connectivity tables and quantitatively compared using entropy-based metrics. In test calculations, MolBlaster accurately reproduced a structural key-based similarity ranking of druglike molecules.

INTRODUCTION

The majority of contemporary molecular similarity methods utilize structural or property descriptors for the generation of appropriate chemical reference spaces.[1] Exceptions include, for example, the use of simplified molecular graph representations as queries for similarity searching.[2] The selection of descriptors that are most suitable for specific applications such as a compound class-directed clustering or virtual screening of databases often presents a difficult task[1,3] and might be influenced by subjective decisions or require machine learning.[3] We set out to investigate the design of a descriptor-independent method to assess molecular similarity by focusing on the chemical information of small molecules. An interesting study along those lines has previously been presented by Graham and colleagues who generated "tape recordings" of synthetic molecules from atom-bond-atom units extracted from molecular graphs by random walks.[4] This approach is distantly related to a systematic exploration of connectivity pathways through a molecule for the design of 2D fingerprints[5] or to the generation of atom pair descriptors.[6] Rather than considering topological molecular features, we have focused on the generation of molecular fragment populations. A number of well-defined computational fragmentation schemes have been devised including hierarchical fragmentation of molecules for the analysis of core structures in drugs[7] or retrosynthetic fragmentation of compounds for de novo ligand design.[8] Furthermore, the generation of dictionaries of structural key-type descriptors that are important tools in pharmaceutical research[9] involved knowledge-based fragment design.

However, departing from hierarchical, retrosynthetic, or other defined fragmentation strategies, we have asked the question whether randomly generated molecular fragment populations would encode sufficient information to serve as "descriptor-free" signatures of compounds for similarity evaluation. For this purpose, we have subjected diverse molecules to extensive random fragmentation and monitored the resulting fragment populations in histograms. These fragment profiles have been quantitatively analyzed and compared using entropic measures. Profile comparisons accurately accounted for molecular similarity relationships among various druglike compounds. The development of the MolBlaster approach and its evaluation are reported herein.

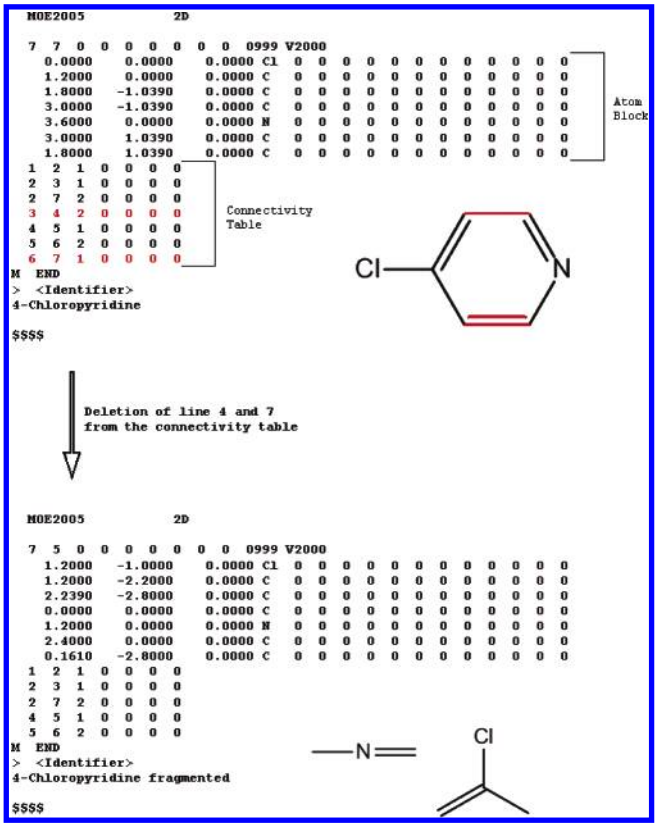

Figure 1. Exemplary fragmentation step. The red lines (four and seven) in the upper connectivity table were randomly selected. Corresponding bonds in 4-chloropyridine are also colored red. Deletion of these lines results in a new connectivity table encoding two molecular fragments.

* Corresponding author phone: +49-228-2699-306; fax: +49-228-2699-341; e-mail: bajorath@bit.uni-bonn.de.

10.1021/ci0601261 CCC: $33.50
Published on Web 07/20/2006

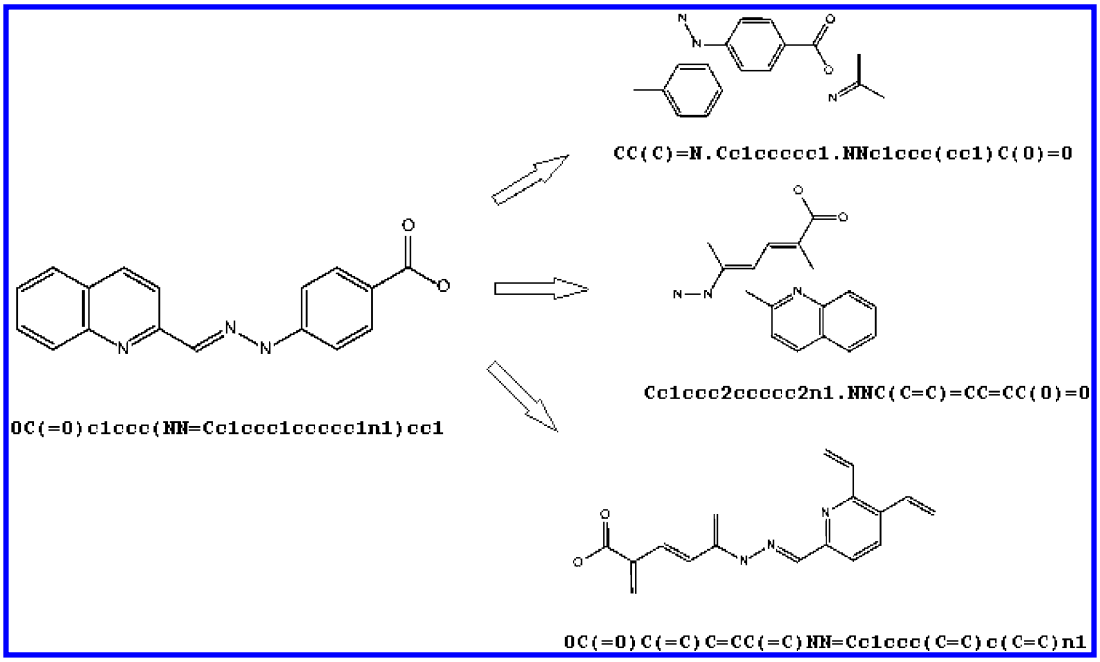

Figure 2. Fragment representations. Shown are sets of fragments of increasing size with corresponding SMILES string representations produced in different deletion steps.

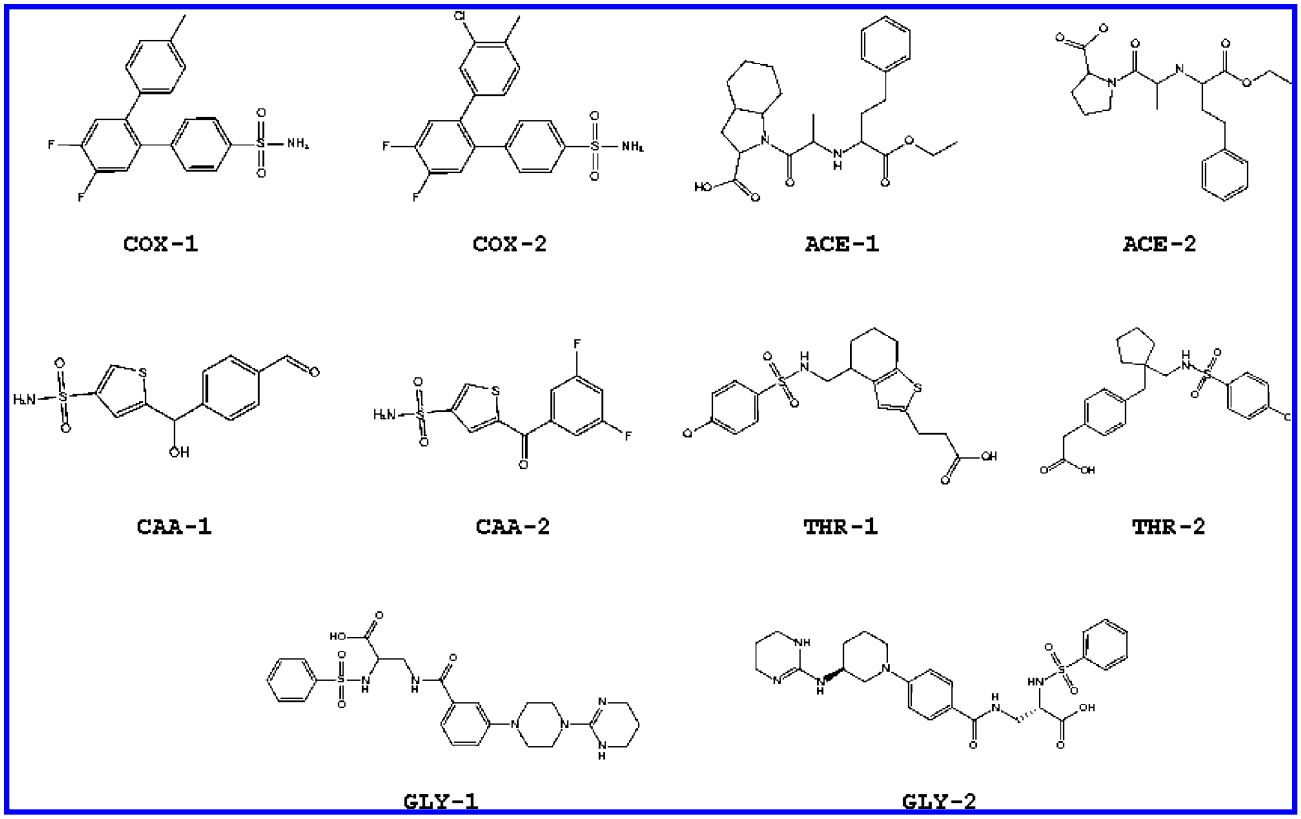

Figure 3. Test compounds. MDDR activity class abbreviations: COX, cyclooxygenase-2 inhibitors; CAA, carbonic anhydrase inhibitors; GLY, glycoprotein IIb-IIIa inhibitors; ACE, acetylcholine esterase inhibitors; THR, thromboxane antagonists.

METHODS AND CALCULATIONS

Molecular fragmentation was facilitated by random deletion of rows from connectivity tables of hydrogen-suppressed 2D graph representations. This process corresponds to randomly breaking bonds in a molecule and is illustrated in Figure 1. Connectivity tables were calculated with the Molecular Operating Environment (MOE),[10] and rows were selected for deletion with a random number generator. If a number fell outside the range of lines in the connectivity table, no deletion was carried out. From reduced connectivity tables, SMILES[11] strings of molecular fragments were exported for fragment sampling, as shown in Figure 2. Fragments consisting of only one or two atoms were omitted from further analysis.

The results of MolBlaster calculations largely depend on two parameters, the maximum number of permitted bond deletions per step (iteration) and the number of fragment-producing iterations. The number of deletions is responsible

for the fragmentation degree and determines the average fragment length and size distribution. The number of iterations controls the total number of diverse fragments that are generated and the frequency of occurrence of each fragment. In test calculations, we have systematically varied these two parameters in order to identify ranges preferred for molecular comparisons, as discussed below.

Fragment populations of a molecule can be conveniently displayed in histograms. Recording different fragment populations in histograms with constant binning schemes makes it possible to quantify and compare their information content using Shannon entropy[12] (SE) calculations adapted for chemoinformatics applications.[13,14] Briefly, Shannon entropy is defined as

$$\mathrm{SE} = -\sum_i p_i \log_2 p_i$$

Here p_i is the probability of a data point falling as a count c within a specific data range i (histogram bin) and calculated as $p_i = c_i/\sum c_i$.

The SE value is used as a quantitative measure of the information content of a data distribution in a histogram format. In our case, a data point is a molecular fragment. Differential Shannon entropy (DSE)[14] is an extension of this concept and takes differences in variability and also value ranges of distributions into account. It is defined as

$$\mathrm{DSE} = \mathrm{SE}_{\mathrm{AB}} - \left(\frac{\mathrm{SE}_{\mathrm{A}} + \mathrm{SE}_{\mathrm{B}}}{2}\right)$$

SE_A and SE_B are the SE values for two different data distributions (in our case, fragment histograms of two molecules, A and B), and SE_{AB} is the Shannon entropy calculated for the combined data set. Thus, a nonzero DSE value represents an increase or decrease in data variability due to synergies in information content between the individual data distributions. In other words, the larger the absolute DSE value is, the more different are the individual data distributions. For our analysis of fragment histograms we have calculated scaled SE (sSE) and DSE (sDSE) values that are normalized by the number of histogram bins (and thus bin number-independent). Furthermore, to emphasize differences between DSE values smaller than one (as typically produced by DSE calculations), we have calculated reciprocal DSE values as

$$\mathrm{rDSE} = 1/|\mathrm{sDSE}|; \text{ for } \mathrm{sDSE} \neq 0$$

An sDSE value of zero means that the compared distributions are identical.

To evaluate MolBlaster calculations for the analysis of molecular similarity relationships, we have selected five pairs of molecules with similar activity from the Molecular Drug Data Report (MDDR).[15] The structures of these active compounds are shown in Figure 3. The molecules were subjected to systematic pairwise comparisons by calculation of the Tanimoto coefficient (Tc)[16] using a fingerprint consisting of a set of 166 publicly available structural keys.[17] Table 1 reports the resulting similarity ranking and confirms the presence of a spectrum of compound relationships with decreasing molecular similarity that span almost the entire Tc range. For MolBlaster calculations, Tc similarity served as the reference. For our analysis, we required a set of molecules with gradually decreasing pairwise similarity, and, therefore, the selected compounds provided a meaningful test set.

Table 1. Tc-Based Similarity Ranking of Test Molecule Pairs[a]

rank	molecule pair		Tc	
1	GLY-1	GLY-2	0.99	T 1
2	COX-1	COX-2	0.98	T 2
3	ACE−1	ACE−2	0.92	T 3
4	THR-1	THR-2	0.91	T 4
5	CAA-1	CAA-2	0.84	T 5
6	CAA-2	COX-1	0.79	T 6
7	CAA-2	COX-2	0.78	T 7
8	CAA-1	COX-1	0.65	T 8
9	CAA-1	COX-2	0.64	T 9
10	CAA-1	THR-1	0.64	T 10
11	CAA-2	THR-1	0.61	
12	GLY-1	THR-2	0.61	
13	GLY-2	THR-2	0.60	
14	GLY-1	THR-1	0.60	
15	GLY-2	THR-1	0.60	
16	CAA-1	THR-2	0.58	
17	COX-2	THR-2	0.55	
18	CAA-2	THR-2	0.55	
19	COX-2	THR-1	0.54	M 19
20	COX-1	THR-2	0.54	M 20
21	COX-1	THR-1	0.53	M 21
22	ACE−2	GLY-1	0.47	M 22
23	ACE−2	GLY-2	0.47	M 23
24	ACE−1	GLY-1	0.45	M 24
25	ACE−1	GLY-2	0.45	M 25
26	CAA-1	GLY-1	0.45	M 26
27	CAA-1	GLY-2	0.44	M 27
28	ACE−1	THR-1	0.41	M 28
29	ACE−2	THR-1	0.40	
30	COX-1	GLY-1	0.38	
31	CAA-2	GLY-1	0.38	
32	ACE−1	THR-2	0.38	
33	ACE−2	THR-2	0.38	
34	COX-2	GLY-1	0.38	
35	COX-1	GLY-2	0.38	
36	CAA-2	GLY-2	0.38	B 36
37	COX-2	GLY-2	0.38	B 37
38	ACE−1	CAA-1	0.27	B 38
39	ACE−2	CAA-1	0.27	B 39
40	ACE−1	CAA-2	0.20	B 40
41	ACE−2	CAA-2	0.20	B 41
42	ACE−1	COX-1	0.17	B 42
43	ACE−1	COX-2	0.16	B 43
44	ACE−2	COX-1	0.15	B 44
45	ACE−2	COX-2	0.15	B 45

[a] All possible pairwise comparisons were carried out for the molecules shown in Figure 3, and molecule pairs were ranked according to decreasing Tc values. "T", "M", and "B" designate the top 10 (most similar), midrange, and bottom (least similar) pairs, respectively. These 30 molecule pairs were labeled to permit an easy graphical comparison of this similarity ranking with the results of MolBlaster calculations.

RESULTS AND DISCUSSION

Prototypic Fragment Populations. To illustrate the features of fragment populations and their histogram representations (profiles), we have calculated fragment populations for aliphatic molecules, ring structures, and compounds combining different moieties. Examples are shown in Figure 4. In each of these cases, a maximum of five bonds were deleted per step. For *n*-octadecane, all fragments having the same length are identical, leading to a simple and undifferentiated population histogram, and small fragments with four to six atoms dominate the fragment population. A more

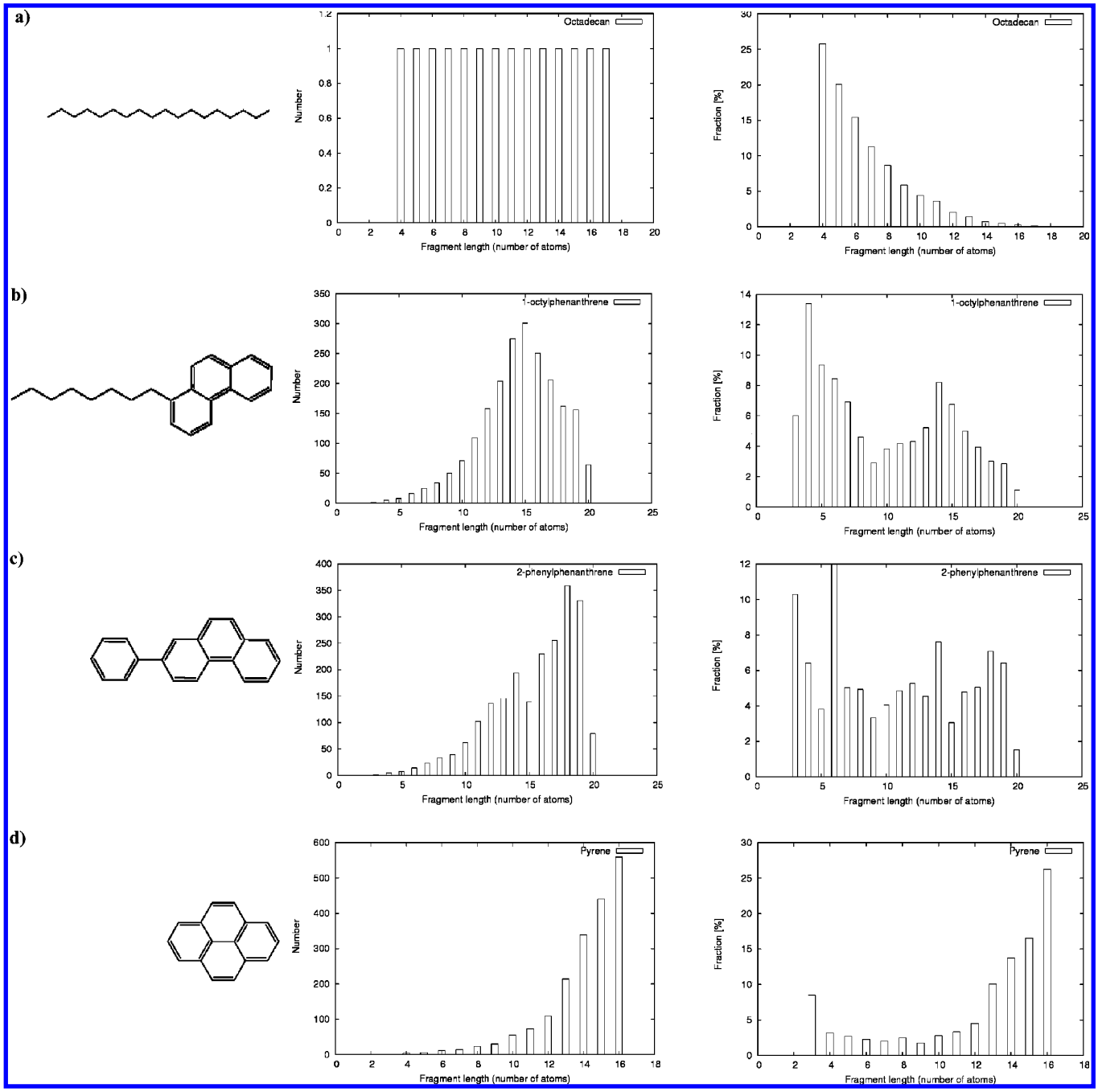

Figure 4. Fragment populations. Shown are histogram representations of fragment populations produced for four molecules when permitting five bond deletions per step in calculations with 3000 iterations: (a), *n*-octadecane, (b), 1-octylphenanthrene, (c) 2-phenylphenanthrene, (d) pyrene. The graphs on the left reflect the diversity of fragments of equal length. Here "Number" reports the count of unique (structurally distinct) fragments of equal size produced during the fragmentation process. The graphs on the right show the variation of fragment length within the fragment population. "Fraction" gives the percentage of fragments having a specific length of the total fragment population. In these calculations, multiple copies of the same fragments are taken into account.

informative distribution is obtained for 1-octylphenanthrene. Here different fragments of equal size are produced, and there is considerable diversity for small to mid-size fragments. This compound has higher order connectivity and is thus much more resistant against decomposition into smaller fragments than *n*-octadecane. Accordingly, the fraction of large fragments substantially increases. The analogue 2-phenylphenanthrene contains an additional ring instead of the aliphatic substituent, and the increased ring content leads to the generation of larger fragments. Furthermore, the fragment distribution is more balanced than for the aliphatic phenanthrene derivative. The peak fraction of fragments contains six atoms, which can be rationalized by considering that only a single bond deletion is required to produce an isolated benzene ring as a fragment. Pyrene is a more complex and symmetrical molecule, and its fragment profile is dominated by the occurrence of many copies of relatively few large fragments. In this case, mid-size fragments are the least frequently detected, but 10−15% of the total fragments are small (consisting of three to four atoms), although these small fragments comprise only less than 0.5% of the unique fragments.

Taken together, these examples illustrate that compounds of different chemical complexity produce different fragment

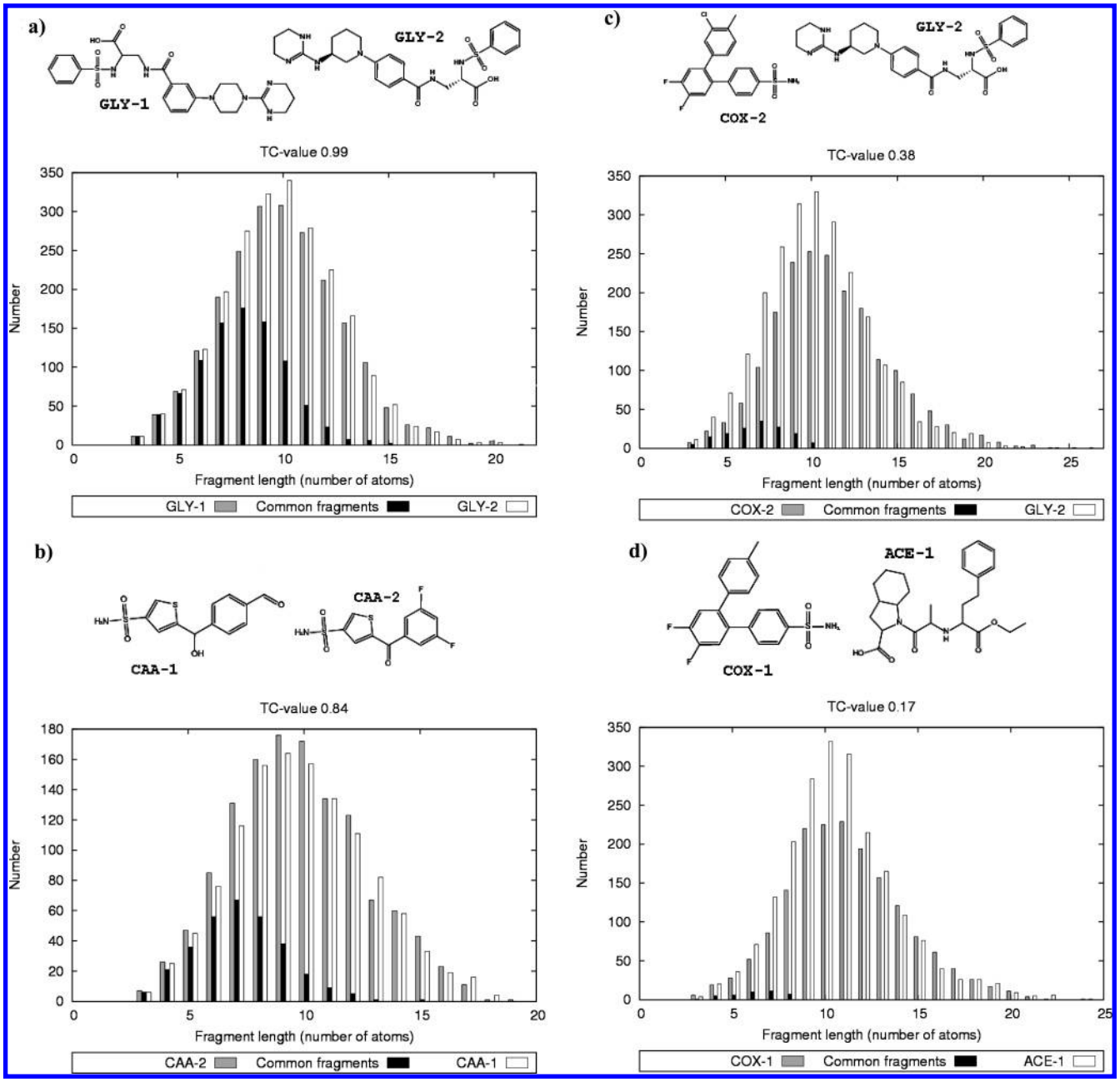

Figure 5. Comparison of fragment distributions. Fragment populations of pairs of test molecules, with decreasing Tc similarity from (a) to (d), are represented in a single histogram for each pair. "Number" reports the number of unique fragments of a given length. Subpopulations of fragments shared by both molecules ("common fragments") are shown in black.

populations when subjected to MolBlaster calculations. Moreover, comparison of the histograms indicates that even closely related molecules such as the phenanthrene analogues produce different fragment distributions. These findings suggest that fragment profiles are sensitive to characteristic molecular information and capable of differentiating between molecules having similar core structures.

Molecular Comparisons. In the next step, we compared fragment populations of pairs of test molecules. Examples of pairwise comparisons of molecules with decreasing Tc similarity are given in Figure 5. These fragment populations were produced permitting 17 deletions per step for 5000 iterations, and only unique fragments were taken into account. At this fragmentation level, population histograms display useful and intuitive fragment information. Histogram comparison revealed some general trends. With decreasing structural similarity between molecules in a pair, the number of shared unique fragments decreases. Thus, these fragment subpopulations mirror structural similarity, at least qualitatively. Furthermore, with decreasing Tc similarity, differences in the numbers of equally sized fragments occur, which changes the shape of the distributions.

Entropy-Based Similarity Metrics. The next question has been how to quantitatively assess differences in fragment distributions as a measure of molecular similarity. There are several possibilities. For example, one could focus on varying amount of common fragments, as illustrated in Figure 5. However, simply taking the number or fraction of common fragments as a measure of similarity would not be sufficiently accurate because their length distribution differs for each molecule pair and the molecules differ in size. Therefore, we have investigated sSE and rDSE calculations. The SE

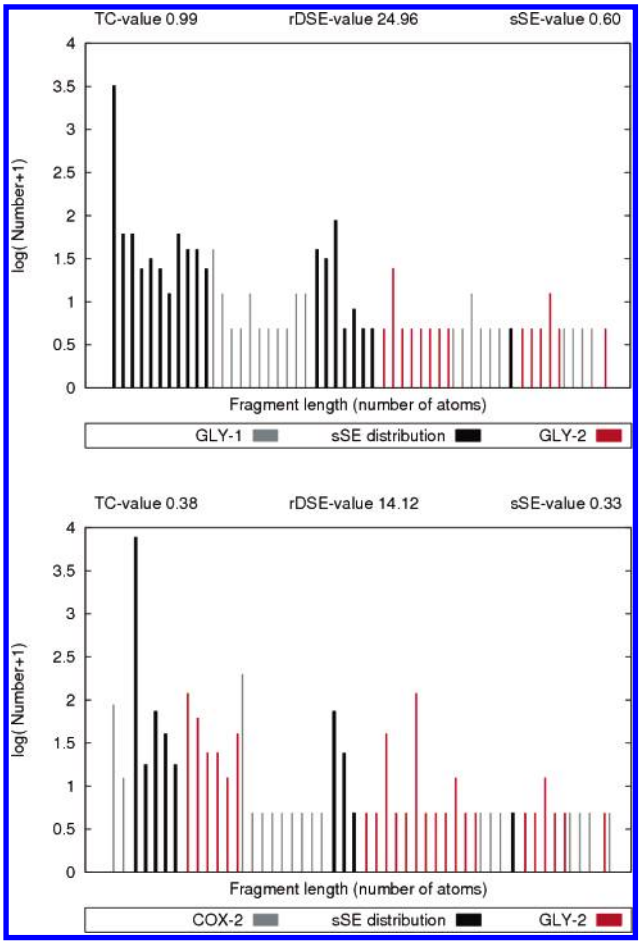

Figure 6. Calculation of entropy values. The figure illustrates how sSE and rDSE values are calculated to quantitatively compare fragment profiles. The molecule pairs at the top and bottom correspond to (a) and (c) in Figure 5, respectively. Fragments are ordered according to increasing length. Average numbers of fragments common to both distributions are shown in black. Common fragments are taken as a separate distribution for the calculation of sSE values. For the calculation of rDSE values, the individual fragment distributions are compared. The "log(Number+1)" representation, where "Number" refers to the count of each unique fragment, was used here for clarity (so that single peaks do not dominate the histogram representation).

concept from information theory[12] has been introduced in the chemoinformatics field a few years ago. In parallel with our original adaptation of the SE concept for descriptor and database profiling,[13] SE calculations were also applied in diversity design.[18] Subsequent investigations have applied the SE formalism to describe the information content of organic compounds,[19] select descriptors and features,[20] or analyze electron distributions on molecular surfaces.[21] SE calculations are designed to quantify the information content of histogram-formatted data distributions and DSE calculations to determine differences between them, even if they are subtle.[14] It follows that SE-DSE metrics should also be suitable to quantify similarities and differences between individual fragment populations.

Similarity Calculations. In Figure 6, we illustrate how sSE and rDSE values are calculated for fragment profiles. For comparison of two molecules, we generate a histogram representation where each unique fragment is assigned to a bin and its frequency of occurrence recorded. Fragments are ordered according to increasing length. The region of common fragments is considered a single and separate distribution for which the sSE value is calculated. The rDSE value is calculated for the comparison of the individual distributions. Calculation of sSE and rDSE values account for similarities and differences between fragment populations. In general, an increase in fragment overlap correlates with a distribution having a higher sSE value. In addition, broader distributions of common fragments correspond to an increase in fragment diversity and also result in higher sSE values. In information-theoretic terms, this means that larger and more widely distributed sets of common fragments have higher information content than smaller ones. These findings are consistent with our qualitative (and intuitive) observations that the more similar molecules are the greater their fragment overlap becomes. Furthermore, as we also observe, decreasing structural similarity causes differences between fragment profiles, in particular, in the relative distributions of similarly sized fragments. Our calculations indicate that rDSE values are a sensitive measure to quantify global differences between fragment profiles. For increasingly similar compounds, rDSE values become larger (as DSE values approach zero).

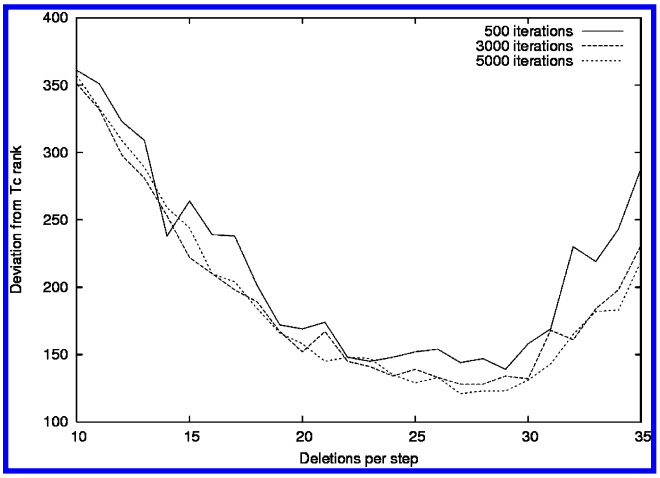

Figure 7. Determination of preferred fragmentation levels. Systematic variations of permitted deletions per step are shown for calculations of different length. In this example, rDSE values were used for fragment-based similarity ranking. The sum of deviations from MACCS Tc ranking was calculated as described in the text.

Preferred Fragmentation Levels. To investigate which fragmentation levels were most suitable for similarity analysis, we systematically varied the number of deletions and iterations for pairwise comparison of our test compounds and calculated deviations from Tc ranking. Runs of varying length (with 100−5000 iterations) were carried out while permitting between five and 50 deletions per step. Compounds were ranked based on sSE or rDSE values, and deviations from the Tc ranking were calculated as the sum of deviations in rank positions over all compounds. Representative results are reported in Figure 7. In these calculations, sSE- and rDSE-based ranking produced very similar results. Notable fluctuations in relative compound rankings were only observed during the initial ∼2000 iterations; then only minor differences could be detected. Thus, many copies of unique fragments were not required to differentiate between molecular similarity relationships. However, the fragmentation level had significant influence on the quality of similarity rankings. When permitting between 10 and 20 deletions per step, the correspondence between Tc- and

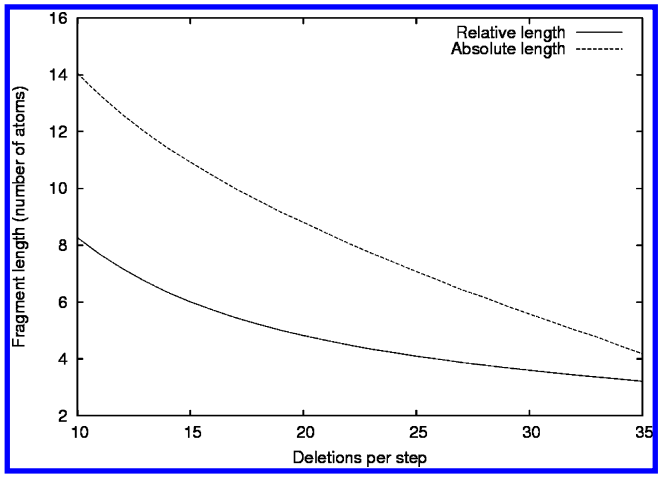

Figure 8. Average fragment length. The average length of all fragments was calculated over all molecules as a function of the number of deletions. "Absolute length" reports the average and "Relative length" the average length of fragments weighted by their frequency of occurrence. Calculations were run for 5000 iterations.

fragment-based similarity rankings improved in a near linear fashion, and the smallest deviations occurred within the range of 27−29 deletions per step. Going beyond 30 deletions, deviations from Tc rankings slightly increased again (probably because the fragments became on average too small). Figure 8 reports average fragment lengths as a function of the number of deletions. At preferred fragmentation levels, fragments contained on average about six atoms (or four when fragments are weighted by their frequency). Thus, extensive fragmentation was required for effective similarity ranking.

Similarity Ranking. A key question in our study has been whether the MolBlaster approach can detect molecular similarity relationships and distinguish similar compounds from others. Therefore, we have used Tc similarity-based molecule pair rankings as a reference for our calculations. Table 2 shows the fragment-based ranking of our test compounds for 28 deletions per step and using rDSE as similarity metric. The ranking closely reproduces the Tc-based ranking in Table 1 including the top, midrange, and bottom molecule pairs. For the top molecule pairs, covering a MACCS Tc range from 0.99 to 0.64, both rankings are nearly identical. Differences in single rank positions mostly occur within the Tc from 0.20 to 0.40 where many Tc values are similar or identical. Within this range, rDSE is a more sensitive similarity metric, as it further distinguishes between pairs of molecules having the same Tc value. For ligand-based virtual screening[3] and scaffold hopping,[22] a similarity method must be capable of recognizing gradually decreasing similarity relationships, including remote ones.[3] The overall close correspondence between Tc- and rDSE-based compound pair rankings suggests that MolBlaster calculations should have considerable potential for such applications.

Related Concepts. In addition to other studies designed to systematically capture the information content of organic compounds,[4] the MolBlaster approach bears some resemblance to the analysis of mass fingerprint spectra of peptides[23] or small molecules[24] (although its underlying fragmentation scheme does not produce ionized molecular fragments). Similar to experimental mass profiles the MolBlaster profiles generate a fragmentation signature of a molecule. However, in fragment profiles, only the frequencies of fragments of a given length are recorded and not mass/charge distributions of fragments. Moreover, the entropy-based comparison of fragment profiles presented herein is distinct from database search, data mining, or pattern recognition techniques applied to compare mass fingerprint spectra.[25,26]

Table 2. rDSE-Based Similarity Ranking of Test Molecule Pairs[a]

rank	molecule pair		rDSE	
1	GLY-1	GLY-2	212.55	T 1
2	COX-1	COX-2	185.17	T 2
3	ACE−1	ACE−2	160.23	T 3
4	THR-1	THR-2	96.79	T 4
5	COX-1	CAA-2	72.19	T 6
6	COX-1	CAA-1	56.60	T 8
7	CAA-2	CAA-1	51.26	T 5
8	COX-2	CAA-2	50.52	T 7
9	COX-2	CAA-1	41.13	T 9
10	COX-2	THR-2	39.44	
11	COX-2	THR-1	36.77	M 19
12	CAA-1	THR-1	36.27	T 10
13	COX-1	THR-2	35.86	M 20
14	COX-1	THR-1	34.00	M 21
15	CAA-1	THR-2	33.51	
16	CAA-2	THR-1	29.81	
17	GLY-2	THR-2	28.45	
18	GLY-2	THR-1	27.99	
19	GLY-1	THR-1	26.66	
20	CAA-2	THR-2	26.52	
21	GLY-1	THR-2	26.50	
22	ACE−1	THR-2	25.30	
23	GLY-2	ACE−1	25.11	M 25
24	GLY-2	ACE−2	24.49	M 23
25	GLY-1	ACE−2	22.85	M 22
26	ACE−2	THR-2	22.82	
27	GLY-1	ACE−1	22.46	M 24
28	CAA-1	GLY-1	21.87	M 26
29	CAA-1	GLY-2	21.46	M 27
30	COX-1	GLY-1	21.21	
31	COX-1	GLY-2	21.20	
32	COX-2	GLY-1	20.93	
33	COX-2	GLY-2	20.88	B 37
34	CAA-2	GLY-1	20.54	
35	ACE−1	THR-1	20.43	M 28
36	CAA-2	GLY-2	20.35	B 36
37	ACE−2	THR-1	19.12	
38	COX-1	ACE−2	19.11	B 44
39	COX-1	ACE−1	17.88	B 42
40	COX-2	ACE−2	17.51	B 45
41	COX-2	ACE−1	16.75	B 43
42	CAA-1	ACE−2	16.46	B 39
43	CAA-1	ACE−1	15.78	B 38
44	CAA-2	ACE−2	14.67	B 41
45	CAA-2	ACE−1	14.42	B 40

[a] The rDSE-based ranking was calculated for 28 possible deletions per step and 5000 iterations. Molecular pairs are designated according to Table 1.

CONCLUSIONS

We have introduced a similarity method that does not depend on the use of conventional molecular descriptors or abstract molecular representations such as reduced graphs. Random fragment populations were found to encode sufficient chemical information to mirror molecular similarity relationships. For the analysis and comparison of molecular fragment profiles, entropy-based measures adapted from information theory provided sensitive similarity metrics. The application of MolBlaster calculations in virtual screening involves the comparison of precomputed fragment profiles

of database compounds with those of query compounds, which is computationally straightforward.

REFERENCES AND NOTES

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

(2) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity searching using reduced graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338−345.

(3) Bajorath, J. Integration of virtual and high-throughput screening. *Nature Rev. Drug Discovery* **2002**, *1*, 882−894.

(4) Graham, D. J.; Malarkey, C.; Schulmerich, M. V. Information content in organic molecules: quantification and statistical structure via Brownian processing. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1601−1611.

(5) James, C. A.; Weininger, D. *Daylight theory manual*; Daylight Chemical Information Systems, Inc.: Irvine, CA, 1995.

(6) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure−activity studies: definitions and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(7) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(8) Schneider, G.; Lee, M. L.; Stahl, M.; Schneider, P. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 487−494.

(9) Merlot, C.; Domine, D.; Cleva, C.; Church, D. J. Chemical substructures in drug discovery. *Drug Discovery Today* **2003**, *8*, 594−602.

(10) Molecular Operating Environment (MOE), Chemical Computing Group Inc., 1255 University Street, Montreal, Quebec, Canada H3B 3X3, 2005.

(11) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(12) Shannon, C. E. A mathematical theory of communication. *Bell Syst. Technol. J.* **1948**, *27*, 379−423.

(13) 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.

(14) 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.

(15) Molecular Drug Data Report (MDDR). MDL Information Systems Inc., San Leandro, CA, 2005.

(16) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(17) MACCS structural keys. MDL Information Systems Inc., San Leandro, CA, 2002.

(18) Maggiora, G. M.; Shanmugasundaram, V. Similarity-based Shannon-like diversity measure. Abstract no. 119, 219th American Chemical Society National Meeting. Division of Computers in Chemistry, 2000.

(19) Graham, D. J.; Schacht, D. Base information content in organic molecular formulae. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 942−946.

(20) Wegner, J. K.; Zell, A. Prediction of aqueous solubility and partition coefficient optimized by a genetic algorithm based descriptor selection method. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077−1084.

(21) Clark, T. Shannon entropy as a local surface property. Abstract no. 267, 229th American Chemical Society National Meeting. Division of Computers in Chemistry, 2005.

(22) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold hopping" by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 2894−2896.

(23) Aebersold, R.; Goodlett, D. R. Mass spectrometry in proteomics. *Chem. Rev.* **2001**, *101*, 269−295.

(24) Stuke, M. Ultrafast fingerprint detection of organometallic compounds by laser multiphoton ionization mass spectrometry. *Appl. Phys. Lett.* **1984**, *45*, 1175−1177.

(25) Quadroni, J. P.; Carafoli, E.; Gonnet, G. Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.* **1993**, *195*, 58−64.

(26) Hilario, M.; Kalousis, A.; Pellegrini, C.; Müller, M. Processing and classification of protein mass spectra. *Mass Spec. Rev.* **2006**, *25*, 406−449.

CI0601261