Chemical Database Mining through Entropy-Based Molecular Similarity Assessment of Randomly Generated Structural Fragment Populations

José Batista and Jürgen Bajorath*

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

Received August 28, 2006

We describe a novel approach to search for active compounds that is based on the generation of random molecular fragment populations. As a similarity-based methodology, fragment profiling does not depend on the use of predefined descriptors of molecular structure and properties and the design of chemical space representations. To adapt the generation and comparison of random fragment populations for large-scale compound screening, we compare different fragmentation schemes, introduce the concept of compound class-specific fragment frequencies, and develop a novel entropic similarity metric for compound ranking. The approach has been extensively tested on 15 different compound activity classes with varying degrees of intraclass structural diversity and produced promising results in these calculations, comparable to similarity searching using fingerprints. A key feature of fragment profile searching is that the calculation of compound class-specific proportional Shannon entropy of random fragment distributions enables the identification of database molecules that share a significant number of signature substructures with known active compounds.

INTRODUCTION

The majority of similarity-based methods for molecular comparisons, compound classification, or database searching utilize molecular descriptors for the definition of chemical references spaces.[1,2] Currently, more than 5000 standard descriptors are available for space design,[3] and finding descriptor combinations that are suitable for specific chemoinformatics applications is often a challenging task, and much emphasis has been put on trying to rationalize this process.[4−7] For this and other reasons, we have previously investigated the design of a similarity method that does not depend on the use of predefined structural or property descriptors and introduced the MolBlaster approach that evaluates molecular similarity on the basis of randomly generated molecular fragment populations.[8] As a similarity measure, the information content of fragment profiles of different molecules was quantitatively compared using one of our adaptations of the Shannon entropy concept[9] termed differential Shannon entropy (DSE).[10] In this study, we could demonstrate that fragment profile comparisons accurately accounted for different levels of similarity between various druglike compounds and were able to closely reproduce a structural similarity-based ranking of decreasingly similar molecule pairs.[8] Thus, our key findings have been that randomly generated fragment populations indeed represented a molecular signature and that their quantitative comparison could serve as a measure of structural similarity. Fragment profiles differ from molecular fingerprints[6] because they do not organize predefined or catalogued descriptors in a specific manner and are based on the randomization of structural information, rather than canonical bit string representations.

The results of this initial proof-of-principle study encouraged us to go a step further and investigate the question as to whether fragment profiling could also be developed into a methodology to search databases for active compounds. This presents a number of different challenges. First and foremost, we now need to go beyond the detection of structural resemblance and evaluate structure−activity relationships. Moreover, in database mining, we need to explore such relationships on a large scale[6] and face the problem that the vast majority of database molecules must be correctly recognized as being dissimilar to active probes. Therefore, it needed to be investigated whether fragment profile searching could be sufficiently specific to correctly detect compounds having similar activity and discard others or whether random fragment sampling would create too much background noise for such applications. Finally, sufficient computational efficiency must be achieved to be able to generate and process fragment profiles for many database molecules.

Here, we report the adaptation of random molecular fragmentation and fragment profile comparisons for virtual screening, which was facilitated through the implementation of a modified fragmentation scheme, the calculation of compound class-specific fragment frequencies, and the application of a newly derived entropic similarity metric. The feasibility of this approach and its predictive value was confirmed in test calculations on a variety of compound classes. In these calculations, the performance of the new methodology was comparable to that of similarity searching using state-of-the-art 2D fingerprints and multiple active reference structures. As a descriptor-independent methodology, fragment profile searching complements currently available virtual screening tools.

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

10.1021/ci600377m CCC: $37.00
Published on Web 11/23/2006

METHODOLOGY

Fragment Profile-Based Molecular Comparisons. The basic idea of the fragment profiling approach is that randomly generated molecular fragment populations encode sufficient chemical information to evaluate the degree of similarity of molecules under comparison. The generation of fragment profiles is somewhat reminiscent of the analysis of mass fingerprint spectra,[11] although our computational fragmentation procedures do not generate ionized molecular fragments. Principally similar to mass spectrometric profiles, fragment histograms generate a fragmentation signature of a molecule. A major difference is that computational fragment profiles only monitor the frequencies of fragments of a given length, but not mass/charge distributions of fragments, as in mass spectrometry.

In our original implementation of the MolBlaster approach, molecular fragmentation was facilitated by random deletion of a pre-set number of rows from connectivity tables of hydrogen-suppressed 2D molecular graphs. This process randomly breaks a specific number of bonds per iteration and is conceptually distinct from retrosynthetic fragmenting for synthesis planning[12] or ligand design.[13] Connectivity tables were calculated with the Molecular Operating Environment,[14] and following deletions, SMILES[15] strings of molecular fragments were exported from these reduced tables for fragment sampling and histogram analysis. This basic fragmentation strategy is also followed here in the design of a fragment-based method for compound database mining.

Exploring Structure−Activity Relationships. A primary goal of our study is to quantitatively compare molecular fragment populations as a measure of molecular similarity in order to identify molecules having similar activity. This brings with it a number of specific requirements. First, we need to devise a generally applicable approach that produces fragment populations of maximal information content, irrespective of the topology and size of the molecules. Second, we need to identify fragment subsets that are a signature of an activity class. Third, a generally applicable scoring function is required that quantitatively accounts for differences in molecular fragment profiles and accurately ranks database compounds according to molecular similarity. Finally, compound fragmentation, representation, and comparison must be amenable to high-throughput calculations as a prerequisite for large-scale virtual screening.

Fragmentation Method. Compound fragmentation in principle depends on two parameters: the maximum number of permitted bond deletions per iteration and, in addition, the total number of iterations that are carried out. The number of bond deletions per step affects the average size and length of the resulting fragments and thus the composition of the fragment population. In addition, the fragment population is determined by the size of test molecules and their connectivity patterns (complexity). It follows that a constant number of permitted deletions per step will very likely not produce fragment profiles that sufficiently account for structural differences between diverse compound classes. Therefore, we introduce a modification for the systematic analysis of structure−activity relationships: we randomize the number of allowed bond deletions during each iteration in order to balance fragment generation for molecules having different degrees of complexity. Different from the number of deletions per iteration, fragmentation results are largely insensitive to the total number of fragmentation iterations as long as a required minimum number of steps are carried out. In our initial study,[8] we could also show that 2000 iterations were generally sufficient to produce stable fragment profiles for different compound sets. The calculations reported herein have been carried out over 3000 iterations, which represented a reasonable compromise between accuracy and computing time.

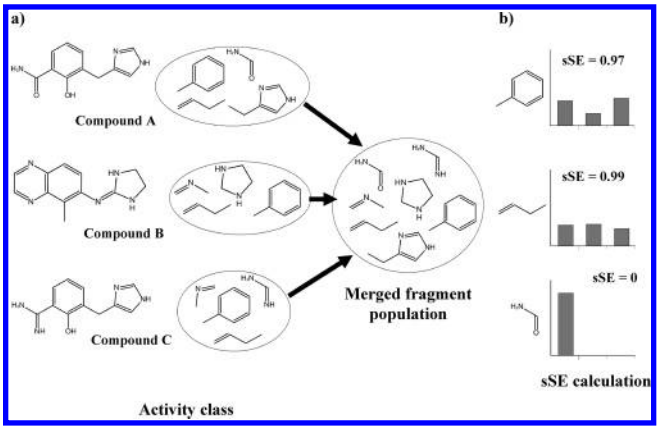

Figure 1. Calculation of class specificity. The figure illustrates how sSE values are calculated to classify fragments on the basis of class specificity. (a) The fragment populations of individual compounds of an activity class are combined into one set, while the frequency of occurrence of all fragments is recorded. Using these frequencies, histograms are generated, as shown in b, which monitor the distribution of fragments over all the compounds. The information content of these distributions is then quantified using the sSE metric.

Entropy-Based Fragment Classification. To represent different compound classes in an informative way and facilitate meaningful comparisons, it is required to generate as many diverse fragments as possible and identify those that are shared by compounds belonging to a particular class. Then, we need to quantify fragment overlap as a measure of similarity. Fragment distributions can be conveniently displayed as histograms. For a set of compounds having equivalent biological activity, fragment populations of all individual compounds are merged into one set, and the frequency of occurrence of all fragments in the combined set is determined. If the frequency of a specific fragment in each compound is recorded in a single histogram, fragment overlap can be quantified by calculating the information content of the histogram representation. This process, which is one of the foundations of our approach, is illustrated in Figure 1. The information content is determined using the Shannon entropy (SE) metric[9] that we adapted for descriptor profiling in compound databases[16] and feature comparisons between different compound databases.[17] Briefly, SE is defined as

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

where p_i is the probability of a data point to fall as a count c into a specific data range i and calculated as $p_i = c_i/\sum c_i$. To obtain SE values that are independent of the histogram representation, we calculate scaled SE values (sSE) that are normalized by the number of histogram bins (here, the numbers of compounds per set). Figure 1 illustrates the

meaning of sSE values. If a fragment only occurs in one compound of a set, the frequency of occurrence is limited to one bin. Thus, following the principles of Shannon entropy analysis, the probability of a data point to fall into this bin is one, and the histogram representation therefore has no information content; that is, the sSE value is zero. By contrast, if a fragment occurs with exactly the same frequency in all n test compounds, the probability of a count to fall into a histogram bin is $1/n$, which corresponds to (maximum data dispersion and) information content, and the resulting sSE value is 1.

For virtual screening calculations reported herein, we accept fragments having a minimal sSE value of 0.75 to represent an activity class. This threshold value was empirically chosen on the basis of test calculations on diverse compound sets. Lowering the threshold value below 0.75 often leads to a lower ranking of active compounds among top-scoring molecules because, at lower sSE values, fragments become specific for single active molecules, rather than a set of different compounds having similar activity. For the assessment of similarity, we introduce the "class-specific frequency" of a qualifying fragment, which is calculated as its average frequency of occurrence for the compounds in that class.

DSE[10] has been introduced as an extension of the Shannon entropy concept to quantify differences in data distributions, even if they are very subtle. 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 $\mathrm{SE}_{\mathrm{AB}}$ 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 synergy in the information content of the individual data distributions. To obtain directly comparable DSE values, we calculate scaled (sDSE) values, analogous to sSE.

For our initial analysis of fragment histograms,[8] we have calculated reciprocal DSE values in order to emphasize differences between DSE values smaller than 1 that are typically produced by DSE calculations:

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

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

Entropy-Based Scoring of Database Compounds. On the basis of the definition of class-specific frequencies of molecular fragments, we can introduce a new entropic scoring scheme designed to evaluate the overall similarity of the fragment profile of an activity class and the fragment population of a single database compound, which we call *Proportional SE* (PSE):

$$\mathrm{PSE} = \sum_i \frac{a}{b}\,\mathrm{sSE}_{\mathrm{AC}}(i)$$

with $a = \min\{\mathrm{Freq}_{\mathrm{AC}}(i);\mathrm{Freq}_{\mathrm{DB}}(i)\}$ and $b = \max\{\mathrm{Freq}_{\mathrm{AC}}(i);\mathrm{Freq}_{\mathrm{DB}}(i)\}$. $\mathrm{Freq}_{\mathrm{AC}}(i)$ is the class-specific frequency of fragment i in a given activity class and $\mathrm{Freq}_{\mathrm{DB}}(i)$ the fragment frequency of a single database compound. $\mathrm{sSE}_{\mathrm{AC}}(i)$ reports the sSE value for fragment i in the activity class (according to Figure 1). Weighting of the sSE value by the frequency proportion favors the detection of database molecules that share many fragments with similar frequency. Such compounds would produce a large cumulative PSE score as an indicator of molecular similarity. Applying this entropic formalism, virtual screening calculations generate a PSE-based ranking of database compounds relative to a class of known active compounds.

Table 1. Compound Activity Classes[a]

code	biological activity	N	N_{baits}	avg Tc	dev
AA2	adrenergic (α 2) agonists	35	12	0.38	0.15
ACE	ACE inhibitors	17	6	0.72	0.12
ANA	angiotensin II-AT antagonists	45	15	0.62	0.12
CAE	carbonic anhydrase II inhibitors	22	7	0.64	0.13
CAL	calpain inhibitors	28	9	0.48	0.14
CHO	cholesterol esterase inhibitors	30	10	0.48	0.18
DD1	dopamine (D1) agonists	30	10	0.54	0.15
ESU	estrone sulfatase inhibitors	35	12	0.61	0.11
GLY	glycoprotein IIb−IIIa receptor−antagonist	34	11	0.59	0.12
HIV	HIV protease inhibitors	18	6	0.67	0.14
KAP	κ agonists	25	8	0.55	0.15
KRA	kainic acid receptor antagonists	22	7	0.56	0.17
LAC	lactamase (β) inhibitors	29	10	0.46	0.18
LDL	upregulator of LDL receptor	30	10	0.49	0.22
SQS	squalene synthetase inhibitors	42	14	0.50	0.17

[a] Compound sets ACE and CAE were assembled from CMC and all other sets from the MDDR, except HIV, which was taken from the literature.[20] N gives the total number of compounds per class and N_{baits} the number of randomly selected template compounds used to generate activity class-specific fragment profiles and frequencies. The remaining active compounds were added to the source database as potential hits in virtual screening calculations. For each set, Tc similarity statistics from pairwise compound comparisons were calculated using MACCS structural keys, and average Tc (avg Tc) values as well as standard deviations (dev) are reported.

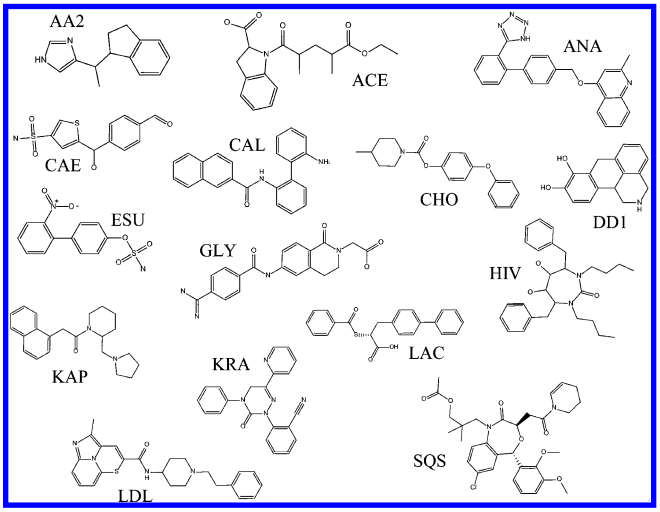

Figure 2. Representative compounds. For each of the activity classes studied here, a randomly selected molecule is shown.

Virtual Screening Trials. For our virtual screening trials, 15 different activity classes were used, as summarized in Table 1. Representative structures are shown in Figure 2. These compound sets were assembled from the Comprehensive Medicinal Chemistry (CMC)[18] database, the Molecular Drug Data Report (MDDR),[19] or taken from the literature.[20] Compound classes were selected on the basis of intraclass average Tanimoto coefficient values[21] calculated using

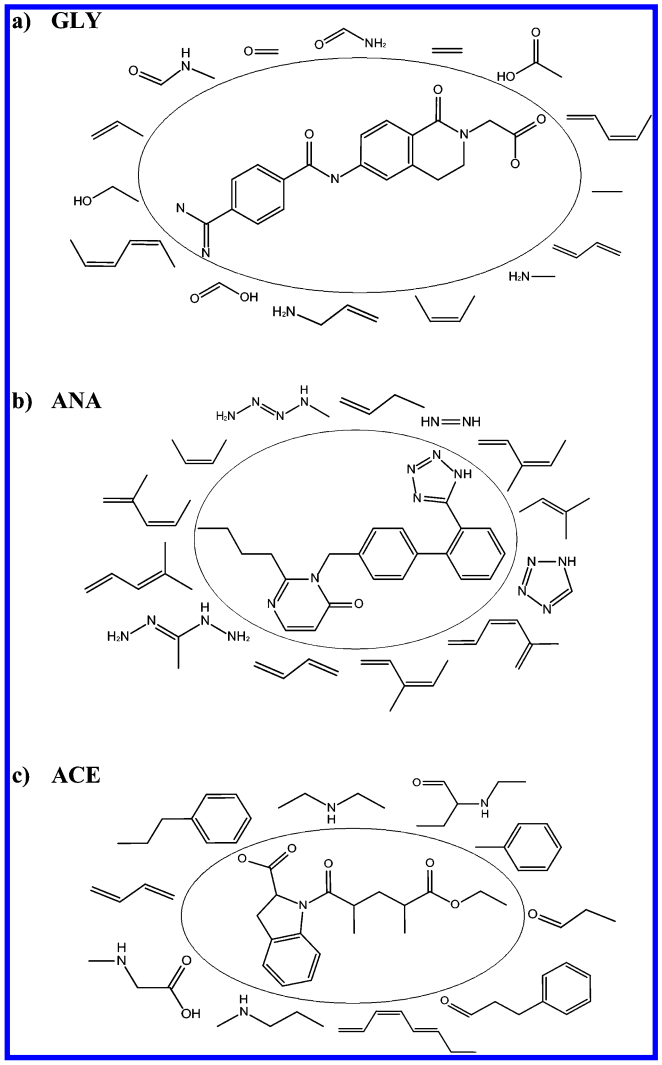

Figure 3. Exemplary fragment populations. Shown are fragments having an sSE value within the range [0.95−1.0] taken from the fragment populations of (a) GLY, (b) ANA, and (c) ACE.

MACCS structural keys.[22] This was done in order to provide a spectrum of compound classes with medium to high structural diversity. Average Tc values for our 15 classes ranged from 0.38 to 0.72 (Table 1).

For each class, about one-third of the compounds were randomly selected as template molecules or "baits" to determine common fragments and class-specific fragment frequencies, and the remaining compounds were added to a source database as potential hits. To reduce chance effects, 10 random bait sets were selected in each case for one series of calculations and subsequently 20 random bait sets for another series. In general, these calculations produced very similar results for either 10 or 20 random bait sets, and hence, we report average results obtained for 10 bait sets. As a source database for virtual screening, a subset of 102 891 compounds with unique 2D molecular graphs was randomly selected from the ZINC collection.[23]

Furthermore, for each activity class, fragment populations were grouped into ranges of decreasing sSE values. We started our investigation with fragments having an sSE value in the range [0.95−1.0] and expanded the lower boundary in steps of 0.05 until reaching the predefined sSE threshold value of 0.75. Figure 3 shows examples of fragments generated from different active compounds having an sSE value within range [0.95−1.0]. Accordingly, for each activity class, calculations on multiple bait sets were carried out using five overlapping sets of signature fragments increasing in size. This was done in order to assess the predictive ability of fragment populations representing different levels of class-specific information content.

Computational Benchmarking. Benchmark calculations revealed that 1000 fragmentation iterations required between ∼0.8 and 1.4 s per molecule on a 3.6 GHz processor. During virtual screening trials, 30−110 database molecules were processed per second of CPU time, depending on the size of the involved fragment populations, thus permitting the screening of large compound sets in a reasonable computing time.

Reference Calculations. To compare the performance of fragment profile searching with that of standard methods in the field, we have carried out reference calculations with different types of 2D fingerprints. As a purely structural fragment-based design, a substructure fingerprint consisting of the publicly available set of MACCS keys[22] (166 bits) was used. Furthermore MPMFP[24] (171 bits) was used, a hybrid fingerprint consisting of structural keys and binary transformed molecular property descriptors. In addition, we included Molprint 2D[25] in the comparison, which is derived from the connectivity table of a molecule and combines large numbers of strings of varying size that represent unique atom environments organized in distance layers (up to $\sim 2^{50}$ strings are theoretically possible). Finally, we also used the Daylight fingerprint,[26] which captures connectivity pathways through molecules and is often considered a gold standard in the field, because it pioneered the development of hashed 2D fingerprints. It was used here in a version consisting of 2048 bits.

Although fingerprinting is conceptually distinct from fragment profiling, fingerprint searches were chosen as reference calculations because they also produce a similarity-based ranking of database compounds. Thus, the number of active molecules occurring within a specific number of top-scoring database compounds can be easily compared. Because fragment profile searching utilizes information from multiple bait compounds, we have used multiple reference structures for fingerprint calculations and applied a multiple template-based search strategy, the centroid approach.[27] When this strategy is followed, an average fingerprint vector is calculated from the individual fingerprints of all available bait compounds and compared to database molecules using the general formulation of the Tanimoto coefficient for numerical values,[21] rather than binary representations.

RESULTS AND DISCUSSION

Evaluation of the Fragmentation Approach. First, we analyzed similarity calculations when randomizing the number of permitted bond deletions per step as opposed to setting a constant number of deletions. For this purpose, we subjected compounds of seven randomly selected classes to pairwise comparisons in 20 fragmentation calculations, each over 3000 iterations, and compared the results to equivalent calculations with varying numbers of constant deletions per step. Compound pairs were ranked according to the rDSE metric, which we used for compound ranking based on structural resemblance in our initial study. The ranked list, from most to least similar pairs, was compared to a Tc-based

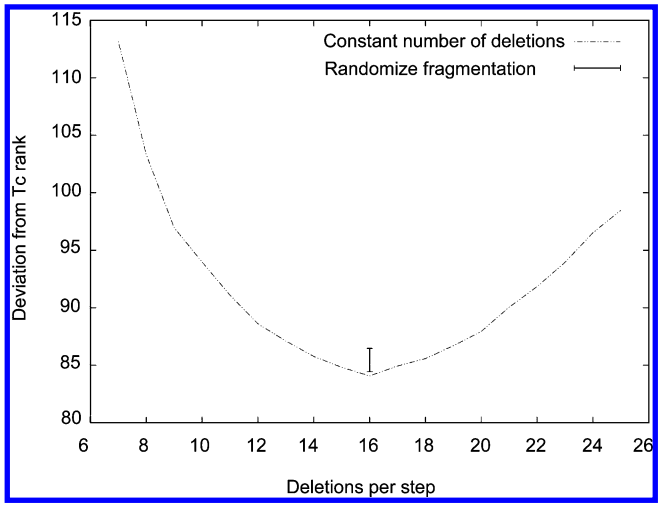

Figure 4. Evaluation of fragmentation methods. For activity class AA2, we report the deviation of rDSE-based rank positions from a Tc-based ranking using MACCS structural keys. The dashed line represents fragmentation using a constant number of deletions per step. The observed range of deviations for the randomized fragmentation scheme is represented by the black bar, positioned at the number of deletions producing the best results.

Table 2. Comparison of Different Fragmentation Methods[a]

	constant no. of deletions			random no. of deletions	
activity class	avg min dev	mean dev	optimal deletion	avg dev	mean dev
AA2	84.2	0.68	16	85.5	0.60
ACE	23.0	0.59	13	26.9	0.24
CHO	46.1	0.28	21	44.5	0.29
ESU	83.6	0.75	10	85.4	0.77
GLY	91.9	0.81	29	99.4	0.91
LAC	61.2	0.26	26	67.2	0.63
SQS	129.2	0.68	32	140.1	1.32

[a] For fragmentation with constant numbers of bond deletions per iteration (constant no. of deletions), we report for each activity class the average minimum deviation (avg min dev) in rank positions from a Tc-based pairwise compound ranking using MACCS structural keys for five independent calculations varying the number of deletions. Also reported are the mean deviation (mean dev) and the number of bond deletions producing the minimum deviation (optimal deletion). For the fragmentation process with randomized deletions per step and 20 independent trials, the average deviation in rank position (avg dev) and the mean deviation (mean dev) are given.

ranking using MACCS keys, and deviations of the entropy-based ranking from the Tc-based ranking were calculated. Thus, in these calculations, we tried to identify a preferred fragmentation routine to reproduce a ranking of molecular pairs according to structural similarity. Representative results are shown in Figure 4, and Table 2 summarizes the results obtained for all seven classes. For each activity class, a different constant number of bond deletions produced lowest deviations from Tc ranking, consistent with differences in size and molecular complexity between these activity classes (as illustrated in Figure 2). Dependent on the activity class, optimal numbers of deletions per step ranged from 10 to 32. Thus, for database mining applications, we required a fragmentation scheme that was less sensitive to compound class-specific features and more generally applicable. Therefore, we explored the randomization procedure. We found that randomizing the number of permitted deletions gave results comparable to those obtained for optimized constant numbers and only slightly higher deviations in six of seven test cases, although no class-directed optimization was carried out. A representative example is shown in Figure 4. The number of permitted deletions per step is the major determinant of the average size of the resulting fragments, and randomization of deletions per step produced more diverse fragment populations than optimization, although observed differences were subtle. Because randomization essentially eliminated class-dependent differences in compound fragmentation, this procedure was clearly the method of choice for virtual screening on diverse compound classes where preferred fragmentation levels are a priori not known. The consistently small standard deviation of the average values reported in Table 2 also confirmed that 3000 fragmentation steps produced stable fragment representations of test compounds.

Virtual Screening Trials. We then tested the approach in virtual screening calculations focusing on the main question of our analysis: can compounds having similar activity be distinguished from very large numbers of database compounds on the basis of their fragmentation profiles? Therefore, we carried out systematic screening trials and recorded recovery rates for the top 100, 500, and 1000 compounds (i.e., for up to 1% of the source database), now using the PSE metric as a scoring function that we specifically developed for this purpose, as described in the Methodology section. Importantly, the PSE metric favors the identification of database molecules that share a significant number of signature fragments with known active compounds. Table 3 reports average results from 10 runs with varying bait sets for all activity classes, and Figure 5 shows representative recall curves for seven of these classes reflecting specific trends. Compound recovery rates varied significantly among the activity classes and were also much influenced by the choice of different sSE ranges and thus fragmentation levels. Overall, the obtained results were promising. For seven classes, recovery rates of >50−100% were obtained for only 100 selected database compounds, and for the remaining classes, the corresponding recovery rates at preferred fragmentation levels ranged from ∼25% to 48%. For the selection of 1% of the databases compounds, nine classes displayed recovery rates of >75−100%. These findings demonstrated that molecular fragment profiles could be successfully used as queries for database searching and that significant recovery rates were achieved for small compound selection sets. The latter aspect is nicely illustrated in Figure 5, showing that recovery rates in general did not dramatically increase when selecting 10 times more database compounds. This suggests that large-scale fragment profile comparisons have a significant degree of specificity.

Figure 5 also illustrates that structurally homogeneous classes such as ACE produced the best compound recovery rates, as one would expect. However, even for distinctly diverse classes such as AA2, recovery rates of ∼20% were achieved within the 100 top-scoring database compounds. These findings suggest that fragment profile searching can explore diverse structure−activity relationships and has scaffold-hopping potential.[28] This conjecture was further supported by the inspection of recovered compounds and bait molecules used to detect them. Representative results are shown in Figure 6 where it can be seen that bait compounds and correctly identified hits frequently had different core structures.

Table 3. Virtual Screening Trials[a]

sSE range	RR_{100}	RR_{500}	RR_{1K}	RR_{100}	RR_{500}	RR_{1K}	RR_{100}	RR_{500}	RR_{1K}
	AA2			ACE			ANA		
[0.95−1.0]	4.8	13.5	19.1	90.0	92.7	92.7	80.3	89.7	93.0
[0.90−1.0]	10.0	18.2	23.9	88.2	91.8	95.5	85.3	93.0	94.3
[0.85−1.0]	17.8	27.4	30.9	91.8	96.4	100	87.0	93.3	94.3
[0.80−1.0]	20.0	32.2	35.2	94.5	98.2	100	92.0	93.3	94.3
[0.75−1.0]	27.8	35.7	37.4	95.5	100	100	93.0	93.7	94.7
	CAE			CAL			CHO		
[0.95−1.0]	62.7	75.3	79.3	11.1	17.4	24.2	10.5	17.0	20.0
[0.90−1.0]	58.0	72.7	78.0	23.2	38.4	48.4	29.0	38.0	40.5
[0.85−1.0]	57.3	72.0	76.0	28.4	44.7	55.3	34.5	43.0	47.0
[0.80−1.0]	55.4	72.7	75.3	34.2	47.4	53.7	47.0	57.0	58.0
[0.75−1.0]	53.3	72.7	73.3	40.5	48.9	56.9	47.0	55.0	57.0
	DD1			ESU			GLY		
[0.95−1.0]	0	2.6	7.9	48.3	58.3	61.7	62.6	80.5	86.6
[0.90−1.0]	25.8	49.0	63.2	78.7	81.3	82.2	46.5	66.1	74.4
[0.85−1.0]	32.6	57.9	70.5	85.7	88.3	88.3	47.0	63.0	69.6
[0.80−1.0]	50.0	72.6	76.8	85.7	87.0	89.6	39.1	53.5	60.0
[0.75−1.0]	56.3	72.1	78.4	85.7	90.0	91.3	40.0	57.8	62.2
	HIV			KAP			KRA		
[0.95−1.0]	98.3	100	100	17.7	28.2	32.9	11.3	27.3	33.3
[0.90−1.0]	100	100	100	28.8	45.9	51.8	16.7	27.3	32.7
[0.85−1.0]	99.2	100	100	35.9	50.0	58.8	19.3	26.7	32.7
[0.80−1.0]	98.3	100	100	37.1	55.9	60.0	18.7	23.7	36.6
[0.75−1.0]	98.3	100	100	38.3	55.3	65.0	17.3	24.7	27.4
	LAC			LDL			SQS		
[0.95−1.0]	17.9	40.5	46.8	2.5	6.0	8.5	25.7	47.1	56.8
[0.90−1.0]	16.3	32.6	43.7	5.5	16.0	23.0	31.8	42.9	49.3
[0.85−1.0]	21.6	41.1	48.7	18	26.0	30.0	35.7	47.5	52.2
[0.80−1.0]	21.6	38.9	45.3	29.5	38.0	41.5	36.1	47.2	52.5
[0.75−1.0]	24.7	38.9	43.7	39	44.5	46.0	35.4	46.5	52.9

[a] Average recovery rates (RR; in percent) are reported for the top-scoring 100 (RR_{100}), 500 (RR_{500}), and 1000 (RR_{1K}) compounds for each activity class and different sSE ranges of the fragment populations.

Reference Calculations. To compare the results our test calculations to standard methods, we carried out multiple template-based 2D similarity search calculations using four different 2D fingerprints of greatly varying complexity. Table 4 reports the recovery rates for these reference calculations relative to the results obtained in our fragment screening trials for selection of the top-scoring 100 database compounds. Observed differences in search performance were overall comparable in a number of cases. For example, for seven activity classes, differences between all four fingerprints and the fragment profiling method were within ∼15%. Overall, Molprint 2D performed best in these calculations using multiple reference structures and the fingerprint centroid approach, followed by MPMFP, fragment profiling, Daylight, and MACCS structural keys. Of the fingerprints used here for comparison to our methodology, Molprint 2D represents the by far highest level of complexity. However, differences in search performance between fingerprints were often subtle. For example, for 10 of 15 classes, differences between Molprint 2D and Daylight were within ∼15%. Thus, the 2D fingerprints used here frequently produced comparable search results, despite significant differences in their design and complexity. Random fragment profiling gave better results than using a dictionary of defined MACCS structural keys in 11 of 15 cases, although differences in recovery rates were here also within ∼15% for 13 activity classes. Thus, taken together, fragment profile searching showed performance in multiple template-based searching at least comparable to that of established and widely used 2D similarity search tools, without the need to use conventional or high-complexity descriptors.

Fragment-Qualifying sSE Ranges. After confirming that the fragment method was capable of consistently detecting active compounds in database search calculations and assessing its search capacity, we analyzed the influence of the size of computed fragment populations on compound recovery. In the majority of—but not all—cases, extension of the fragment-qualifying sSE range increased compound recovery rates, which indicated that larger fragment populations have higher information content, leading to similarity analysis at higher resolution and increasingly accurate compound ranking. Preferred sSE ranges for fragment selection substantially varied among the different activity classes. Furthermore, we found that differences in search performance between single runs for an activity class, which we also observed, could be attributed to differences in fragment populations derived from distinct bait sets. Table 5 summarizes the recovery rates obtained for 10 unique bait sets selected for activity class ESU. Recovery rates were very similar in 6 of 10 cases but significantly lower in four others. Thus, although many bait sets produced comparable recovery rates, differences as a consequence of bait set composition could be observed. An extreme example is displayed in Figure 7 that compares compound recovery rates obtained for two of the ESU bait sets. Here, one randomly chosen bait set produced recovery rates of ∼80% already for fewer than 50 selected compounds, whereas the other essentially failed to recover active compounds. We determined that the fragment population

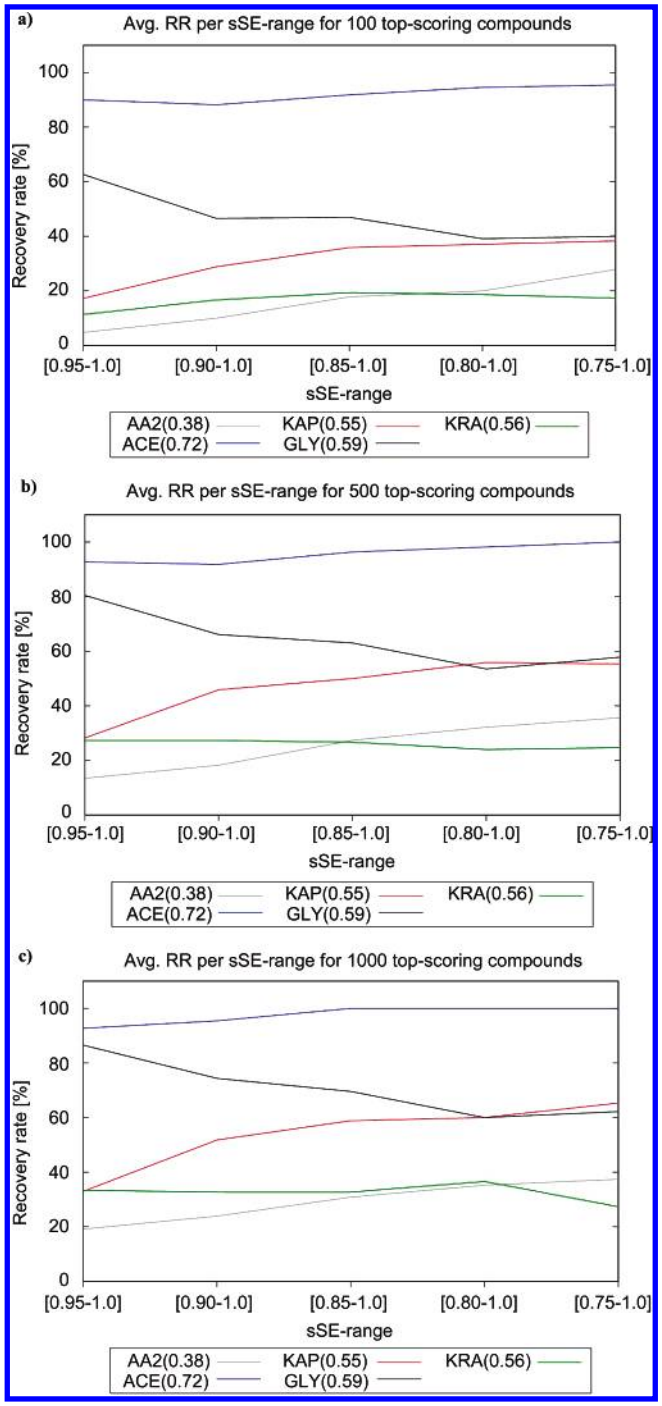

Figure 5. Results of virtual screening trials. The figure reports average recovery rates for different sSE ranges and the top-scoring (a) 100, (b) 500, and (c) 1000 compounds. Average Tc values of the activity classes are reported in parentheses to illustrate differences in intraclass structural diversity.

derived from the well-performing bait set was nearly 3 times larger than the other. Taken together, these findings indicated that differences in fragment numbers and distributions were of significant importance for accurate compound ranking.

Composition of Fragment Profiles and Preferred sSE Ranges. In light of the findings discussed above, we further analyzed the sSE range-dependent fragment distributions. Table 6 reports for each class the average number of fragments for each sSE range. In addition, a percentage of these structures is given that belongs to a set of the most frequently occurring database fragments. This set was

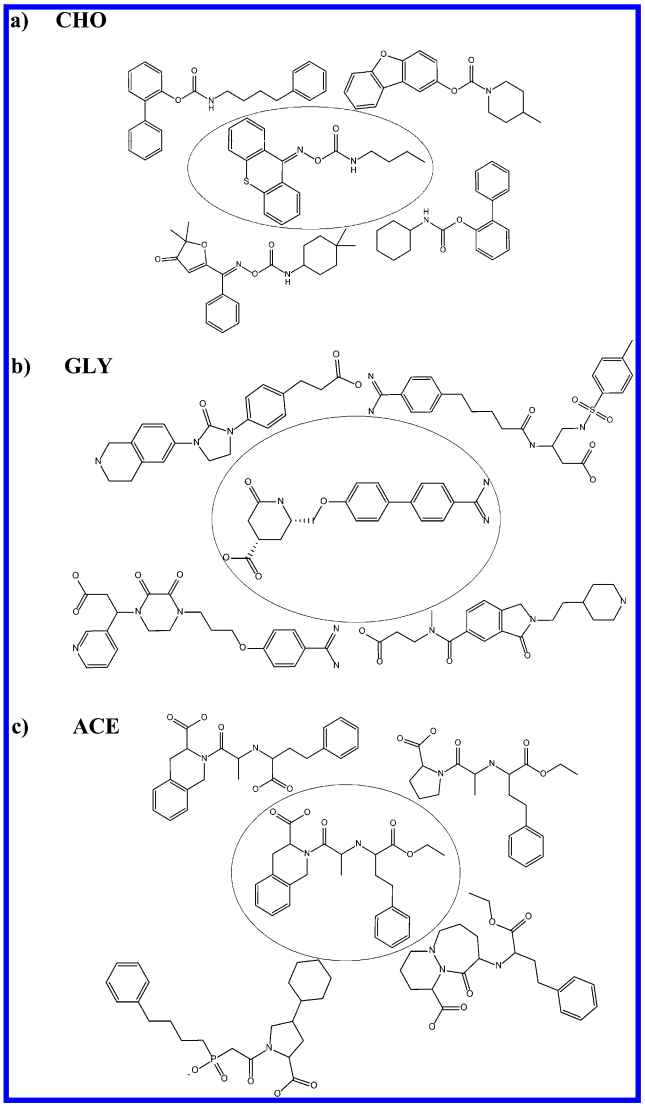

Figure 6. Retrieval of active compounds. Shown are examples of correctly identified hits among the top-scoring 100 database compounds for activity classes (a) CHO, (b) GLY, and (c) ACE.

Table 4. Reference Calculations[a]

class	frag profile	MACCS	MPMFP	Daylight	Molprint 2D
AA2	27.8	20.9	30.9	28.3	33.9
ACE	95.5	81.1	95.5	80.9	100.0
ANA	93.0	65.7	79.4	91.7	85.0
CAE	62.7	56.7	82.0	37.8	52.0
CAL	40.5	26.8	38.4	22.6	53.7
CHO	47.0	49.0	50.0	58.0	60.5
DD1	56.3	61.5	72.0	26.0	59.0
ESU	85.7	79.6	80.0	65.2	95.2
GLY	62.6	48.3	65.7	16.5	51.7
HIV	100	70.0	82.5	93.3	74.3
KAP	38.3	25.9	38.8	24.7	31.8
KRA	19.3	22.7	38.7	51.3	63.3
LAC	24.7	35.8	37.7	51.1	36.0
LDL	39.0	28.0	46.5	43.5	52.0
SQS	36.1	34.3	42.9	40.7	44.6

[a] Reported are best recovery rates (in percent) for the top-scoring 100 compounds using fragment profiling (frag profile) in combination with our PSE similarity metric and four different fingerprints. For similarity searching using these fingerprints, the centroid approach was applied and Tc values were calculated for compound ranking.

calculated by determining all fragments that occurred in at least 10 000 ZINC compounds. These fragments would be

Table 5. Recovery Rates for Individual Bait Sets[a]

bait set	avg. Tc	RR_{100}	RR_{500}	RR_{1K}
1	0.61	69.7	87.0	91.3
2	0.60	4.3	21.7	21.7
3	0.59	0.0	4.3	13.0
4	0.64	13.0	26.1	34.8
5	0.65	78.3	82.6	82.6
6	0.65	78.3	82.6	82.6
7	0.64	78.3	82.6	82.6
8	0.60	73.9	82.6	82.6
9	0.61	8.7	30.4	43.5
10	0.65	78.3	82.6	82.6

[a] For activity class ESU, recovery rates for 10 randomly selected unique bait sets are reported within sSE range [0.95-1.0]. Average MACCS Tc values are reported for a pairwise comparison of bait compounds.

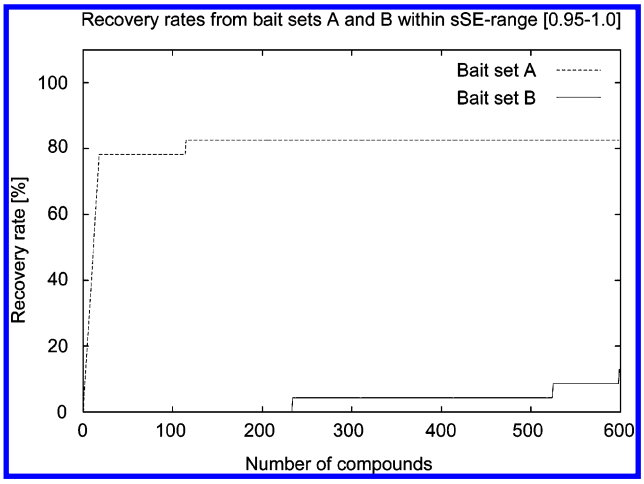

Figure 7. Importance of bait set composition. Reported are recovery rates for two different ESU bait sets. The average Tc values for bait sets A and B are 0.65 and 0.59, respectively.

considered as "nonspecific" in our search calculations or as fragmentation background. As can be seen, the number of unique fragments consistently increased when the sSE range was expanded, as one would expect. However, the number of fragments for each sSE range significantly differed between classes. For example, in the sSE range [0.95−1.0], ACE generated 72 fragments but LDL only six. Furthermore, as illustrated in Figure 3, generated fragments also varied in size. Variations in fragment numbers and their average size were found to be largely due to relative differences in intraclass structural diversity. There is a correlation between the average Tc values of activity classes reported in Table 1 and the number and size of fragments per class: compounds in structurally homogeneous classes tend to produce more and larger shared fragments than increasingly diverse classes. Importantly, Table 6 also shows that each activity class had a significant number of fragments that did not occur in the fragmentation background. This has been a key finding, as it rationalizes the observed specificity and performance of the search calculations. Furthermore, as shown in Table 3, expanding the sSE range often—but not always—increased

Table 6. Size and Composition of Fragment Sets[a]

sSE range	avg no. frgmts	% freq set	avg no. frgmts	% freq set	avg no. frgmts	% freq set
	AA2		ACE		ANA	
[0.95−1.0]	6.6	100	71.5	61.2	39.8	79.8
[0.90−1.0]	11.6	99.6	114.2	48.7	98.5	50.5
[0.85−1.0]	20.2	97.1	154.6	41.7	164.7	34.4
[0.80−1.0]	30.3	92.1	181.2	37.0	236.4	26.8
[0.75−1.0]	41.9	84.0	218.7	32.6	309.9	22.3
	CAE		CAL		CHO	
[0.95−1.0]	18.4	54.9	18.1	95.8	7.5	97.0
[0.90−1.0]	24.9	51.0	36.1	91.0	20.3	91.2
[0.85−1.0]	33.2	50.4	57.3	81.5	38.1	78.6
[0.80−1.0]	45.0	44.6	76.5	71.9	65.9	67.1
[0.75−1.0]	52.3	44.1	116.7	57.8	94.9	55.8
	DD1		ESU		GLY	
[0.95−1.0]	19.4	92.5	28.2	58.9	16.2	99.4
[0.90−1.0]	39.9	79.3	60.9	47.1	28.3	94.2
[0.85−1.0]	59.4	66.9	94.1	40.9	42.8	86.3
[0.80−1.0]	79.6	59.4	123.6	37.1	62.0	73.8
[0.75−1.0]	100.5	52.5	150.6	32.7	86.9	61.1
	HIV		KAP		KRA	
[0.95−1.0]	67.7	59.6	16.1	96.5	13.0	95.2
[0.90−1.0]	93.3	50.5	34.5	90.8	23.3	86.1
[0.85−1.0]	144.8	40.7	54.0	80.6	39.5	76.4
[0.80−1.0]	175.0	35.8	73.9	69.5	56.6	67.9
[0.75−1.0]	212.8	31.4	91.0	61.8	79.6	58.9
	LAC		LDL		SQS	
[0.95−1.0]	6.9	99.3	5.7	98.5	9.5	100
[0.90−1.0]	15.7	96.6	13.1	92.9	20.3	96.8
[0.85−1.0]	26.5	93.8	27.9	85.6	38.6	88.4
[0.80−1.0]	37.6	88.79	45.1	77.0	54.0	77.4
[0.75−1.0]	53.3	80.7	74.3	66.9	67.5	67.8

[a] For each activity class, the average number of fragments (avg no. frgmts) for specified sSE ranges is reported and so is the percentage of fragments belonging to the most frequently occurring set of fragments (% freq set). This set consists of all fragments that also occur in at least 10 000 of the ZINC database compounds used here.

compound recovery. For 8 of 15 classes, the [0.75−1.0] range gave the best results. In these cases, recovery rates increased when the sSE range was expanded as the percentage of class-specific fragments was increasing. For other classes, more narrowly defined sSE ranges were preferred, reflecting class-specific differences in fragment profiling. For example, for CAE and GLY, the smallest sSE range gave the best results and in other cases such as ESU or SQS an intermediate one. However, with only one exception (GLY) among the seven classes that produced the best recovery rates at small or medium size sSE ranges, the results were similar to those obtained for the [0.75−1.0] range. Therefore, for virtual screening trials, the sSE range [0.75−1.0] should present a reasonable choice for targeting many different activity classes in fragment profile searching.

Scientific Context. Among similarity-based search approaches, our fragment methodology is conceptually unique. Its major components are a random fragmentation and sampling scheme and an especially designed information-theoretic similarity metric for fragment profile comparison. Fragment profiles are distinct from structural fragment-based molecular fingerprints because profiles are not based on a bit string scheme, do not require the use of predefined or categorized descriptors, and are evaluated and compared on the basis of information entropy. For comparison, we have carried out similarity search calculations using 2D fingerprints because they also rank database compounds according to similarity criteria. Because fragment profiles are calculated from connectivity tables, 2D fingerprints were selected for our purpose and not 3D fingerprints. It was also interesting to investigate how profiles capturing random fragment distributions would compare in similarity searching to bit string representations representing defined 2D fragments such as MACCS keys. There has been much debate in the literature about the superiority of either 2D or 3D descriptor-based methods, generally revealing a strong test case dependence of the relative performance of 2D and 3D methods.[2,6] On the basis of data available thus far, it cannot be concluded that 3D approaches are in principle superior to 2D methods because 3D descriptors might capture more molecular information than 2D representations.[6] Regardless, for comparison with fragment profiling, descriptor dimensionality becomes largely irrelevant because fragment distributions are independent of conformational features and classical 2D or 3D classification schemes for molecular descriptors[2] do not apply to randomly generated fragments.

In the context of molecular fragment-type approaches, it is worth noting that Pedersen and colleagues have recently introduced a molecular fragment method for the assessment of Tanimoto similarity.[29] To improve the performance of MACCS keys in cluster analysis, these investigators have applied a well-established hierarchical definition of molecular substructures (i.e., rings, linkers, and side chains) to divide conventional Tc calculations into separate ones for ring systems, linkers, and side chains, the weighted sum of which is used as a similarity metric. The derivation of this similarity metric is strictly hierarchical in nature and lacks any random elements. In the context of molecular information theory, Graham and colleagues have reported a study involving randomization steps.[30] These investigators have generated signature patterns of synthetic compounds using atom−bond−atom units extracted from molecular graphs by random walks. This approach is akin to the systematic exploration of connectivity patterns in molecules.

CONCLUSION

In this study, we have introduced the generation and use of compound class-specific fragment profiles for similarity searching. A class-independent fragmentation scheme was developed together with a novel information entropy-based similarity metric. In virtual screening trials, our new methodology displayed significant specificity and produced promising recovery rates for diverse active compounds. The composition of fragment profiles was found to be a major determinant for the predictive value of fragment comparisons, and preferred fragmentation levels could be determined. A key feature of the approach is that random fragment generation and sampling under the conditions established herein produced signature fragments for diverse compound classes that were not present in the fragment background of the screening database, thus enabling specific compound recognition. Correctly identified hits were among high-scoring compounds and often had core structures different from the bait compounds, indicating the potential of fragment profile searching for scaffold hopping. The fragment method is one of very few currently available molecular similarity-based approaches that do not depend on the use of predefined descriptors of molecular structure and properties and the design of chemical reference spaces.

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) Xue, L.; Bajorath, J. Molecular Descriptors in Chemoinformatics, Computational Combinatorial Chemistry, and Virtual Screening. *Comb. Chem. High Throughput Screening* **2000**, *3*, 363−372.
(3) Todeschini, R.; Consonni, V. In *Methods and Principles in Medicinal Chemistry − Volume 11 − Handbook of Molecular Descriptors*; Mannhold, R., Kubinyi, H., Timmerman, H., Eds.; Wiley: New York, 2000; pp 1−667.
(4) 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.
(5) 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.
(6) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882−894.
(7) Bender, A.; Glen, R. C. A Discussion of Measures of Enrichment in Virtual Screening: Comparing the Information Content of Descriptors with Increasing Level of Sophistication. *J. Chem. Inf. Model.* **2005**, *45*, 1369−1375.
(8) Batista, J.; Godden, J. W.; Bajorath, J. Assessment of Molecular Similarity from the Analysis of Randomly Generated Structural Fragment Populations. *J. Chem. Inf. Comput. Sci.* **2006**, *46*, 1937−1944.
(9) Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Technol. J.* **1948**, *27*, 379−423.
(10) Godden, J.; 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.
(11) Aebersold, R.; Goodlett, D. R. Mass Spectrometry in Proteomics. *Chem. Rev.* **2001**, *101*, 269−295.
(12) Lewell, Q. X.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP -Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511−522.
(13) 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.

(14) *MOE (Molecular Operating Environment)*; Chemical Computing Group Inc.: Montreal, Quebec, Canada, H3B 3X3, 2005.
(15) 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.
(16) 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.
(17) Godden, J. W.; Bajorath, J. Chemical Descriptors with Distinct Levels of Information Content and Varying Sensitivity to Differences between Selected Compound Databases Identified by SE-DSE Analysis. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 87−93.
(18) *Comprehensive Medicinal Chemistry Database (CMC)*; MDL Information Systems Inc.: San Leandro, CA, 2005.
(19) *Molecular Drug Data Report (MDDR)*; MDL Information Systems Inc.: San Leandro, CA, 2005.
(20) Xue, L.; Bajorath, J. Accurate Partitioning of Compounds Belonging to Diverse Activity Classes. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 7575−764.
(21) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(22) *MACCS Structural Keys*; MDL Information Systems Inc.: San Leandro, CA, 2002.
(23) Irwin, J. J.; Shoichet, B. K. ZINC − A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.
(24) Xue, L.; Godden, J.; Stahura, F.; Bajorath, J. Design and Evaluation of a Molecular Fingerprint Involving the Transformation of Property Descriptor Values into a Binary Classification Scheme. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1151−1157.
(25) Bender, A.; Mussa, Y.; Glen, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708−1718.
(26) James, C. A.; Weininger, D. *Daylight Theory Manual*; Daylight Chemical Information Systems Inc.: Irvine, CA.
(27) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391−405.
(28) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894−2896.
(29) Jorgensen, A.; Langgard, M.; Gundertofte, K.; Pedersen, J. T. A Fragment-Weighted Key-Based Similarity Measure for Use in Structural Clustering and Virtual Screening. *QSAR Comb. Sci.* **2006**, *3*, 221−234.
(30) 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.

CI600377M