Mapping of Activity-Specific Fragment Pathways Isolated from Random Fragment Populations Reveals the Formation of Coherent Molecular Cores

Eugen Lounkine, José Batista, and Jürgen Bajorath*

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

Received July 16, 2007

Randomly generated molecular fragment populations are investigated as a source for compound-class-dependent chemical information. The analysis of conditional probabilities of fragment co-occurrence in random fragment populations reveals that different classes of active compounds produce series of fragments whose presence depends on each other. Such relationships constitute a fragment hierarchy that becomes a signature of a compound class. We find that such sets of fragments isolated from random populations are typically found to form coherent molecular cores in active compounds. Thus, class-specific random fragment hierarchies encode meaningful structural information. Characteristic core regions already formed by small numbers of substructures remain stable when more fragments are added. These findings provide a structural rationale for the signature character of activity-specific fragment pathways. Thus, randomly generated fragment populations can be mined for combinations of substructures that characterize activity classes. It follows that compound-class-directed structural descriptors can be isolated from random fragment populations that do not depend on the application of predefined fragmentation or design schemes.

INTRODUCTION

In chemoinformatics and drug design, molecular substructures are widely recognized as information-rich and powerful descriptors[1−4] for various applications including molecular similarity analysis, chemical database mining, and compound classification.[5−7] Different types of structural descriptors have been introduced[1,2,4,8] that are generally based on a defined molecular organization[9,10] or synthetic criteria.[11,12] For example, substructures might be derived by dividing molecules in a hierarchical manner into ring-containing core structures, substituents, and linkers[9] or applying retrosynthetic fragmentation schemes.[11] Regardless, the well-defined nature of substructure compendiums[1−3] is a characteristic feature of contemporary structural-fragment-type descriptors.

Recently, randomly generated molecular fragment populations have been analyzed for the first time to evaluate their potential to capture molecule-specific information.[13,14] An initial study carried out in our laboratory introduced a method termed MolBlaster designed to generate random fragment populations of test molecules.[13] MolBlaster fragments are obtained through series of random deletions of bonds in connectivity tables of hydrogen-suppressed two-dimensional (2D) molecular graphs. Careful comparison of the information content of histogram representations of such fragment populations made it possible to successfully detect molecular similarity relationships[13] and identify active compounds through systematic comparison of fragment profiles of known reference molecules and large numbers of database compounds.[14] These studies provided a first use of randomly generated fragments in molecular similarity analysis. Of course, the ability to successfully detect and distinguish between different structure−activity relationships on the basis of random fragment sets raises important questions: what are the elements in random fragment populations that carry compound-class-selective information and how can we characterize and/or identify these components? In order to investigate these and other questions, an algorithm was developed to analyze conditional probabilities of fragment occurrence in random populations.[15] Large-scale mining of fragment populations revealed that profiles of different compound sets contained class-specific fragment pathways.[15] Such fragment pathways are characterized by layers of fragments having different degrees of complexity whose co-occurrence strictly depends on each other. Thus, these fragments represent a conditional hierarchy. In tree representations of random fragment populations, class-specific fragment hierarchies represent unique subgraphs and such distinct patterns of fragment co-occurrence are the major distinguishing feature between random fragment profiles of different compound classes.[15]

Class-specific fragment hierarchies—or, in other words, specific combinations of fragments—are in essence statistically determined, i.e., by conditional probabilities of fragment co-occurrence in large populations. Moreover, fragments per se are not a class signature, because they might appear outside given compound classes, as one should expect. However, unique combinations of fragments, whose presence depends on each other, confer activity class specificity. The dependency relationship means that only if fragment A occurs fragment B is also observed because it is derived from A. Given the statistical nature of fragment hierarchies, their potential meaning at the molecular level of detail has thus far remained elusive.[15] Therefore, we have developed an approach to systematically correlate class-specific fragment

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

10.1021/ci700251b CCC: $37.00
Published on Web 10/16/2007

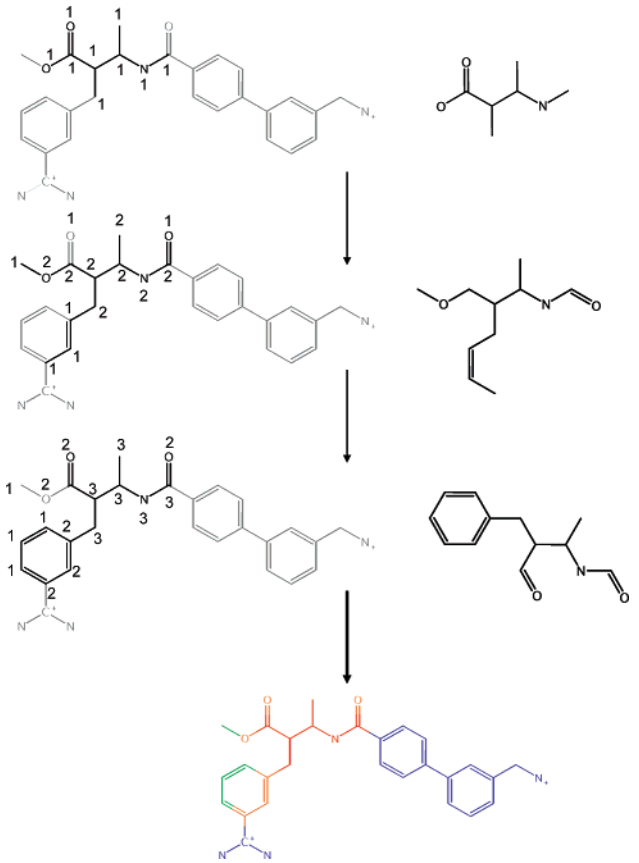

Figure 1. Substructure mapping and core derivation. The procedure of iterative core derivation is illustrated for a hypothetical ACCS set consisting of three fragments. Each fragment is mapped onto the 2D graph of the test molecule and the atom count for matched atoms is updated after each step. In this example, three distinct cores can be distinguished (bottom): $core_{90}$ (red), $core_{60}$ (red, orange), and $core_{30}$ (red, orange, green). Parts of the molecule that were not matched are shown in purple. Atom match rates are color coded following a continuous spectrum from purple to red.

hierarchies with molecular structure. Molecular maps of fragment sets representing unique pathways are found to delineate well-defined core structures in compound activity classes that distinguish them from others. Thus, random fragment profiles can be mined for activity-class-specific combinations of substructures that often delineate central parts of active molecules. These findings rationalize the predictive value of activity-specific fragment hierarchies at the molecular level of detail. Furthermore, our results extend presently available approaches to descriptor design by establishing that sets of activity-class-directed structural descriptors can be isolated from randomly generated fragment populations.

MATERIALS AND METHODS

Compound Activity Classes. In this study, we have systematically analyzed a total of 45 different activity classes assembled from major compound repositories or the literature. These compound sets and their source information are reported in Supplementary Table S1. The activity classes consist of between 11 and 30 molecules (for the majority of classes). In total, the 45 activity classes studied here contain 1025 active molecules.

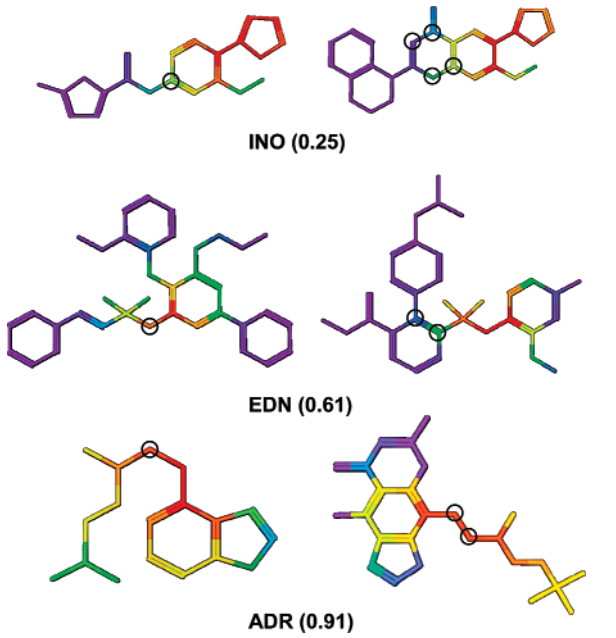

Figure 2. Cores of classes having different center scores. For three representative activity classes with low (INO), intermediate (EDN), and high (ADR) core center scores, two molecules are shown. The color code is according to Figure 1; i.e., red indicates $core_{90}$ and purple indicates atoms that are not matched at any core level. In each molecule, the most central atom(s) is(are) encircled.

Molecular Fragmentation and Characteristic Substructures. For each test molecule, random fragment populations were generated using MolBlaster[13] with a previously evaluated fragmentation protocol[14] with 3000 fragmentation iterations per molecule and randomized numbers of bond deletions per step. This fragmentation scheme was found to produce characteristic fragment populations for different classes of active compounds.[14] For each compound class, we defined a set of activity-class-characteristic substructures (ACCS) by identifying those fragments that occurred in the populations of at least two active molecules but none of the compounds of other activity classes or 500 background compounds randomly chosen from the ZINC database.[16] Up to six different background sets were used in these calculations in order to study the influence of background molecules on the composition of the fragment populations of active molecules. The resulting ACCS sets were used for mapping studies. The ACCS definition used here can be readily modified depending on the conditions of particular applications.

Molecular Mapping of ACCS. Characteristic substructures were mapped back on active molecules by performing a subgraph search for each fragment using the PerlMol package.[17] The mapping procedure is illustrated in Figure 1. Whenever a fragment matched an atom of a test molecule, a counter for the matched atom was increased by 1. For each atom, division of its final counter state by the total number of matched substructures gave its *match rate*.

Core Definition and Analysis. For each active molecule, alternative core regions, $core_x$, were defined as the set of all atoms of the molecule that have an ACCS match rate greater than x% (with x referred to as the *core level*). For each molecule, 10 cores ($core_{90}$ to $core_0$) were calculated.

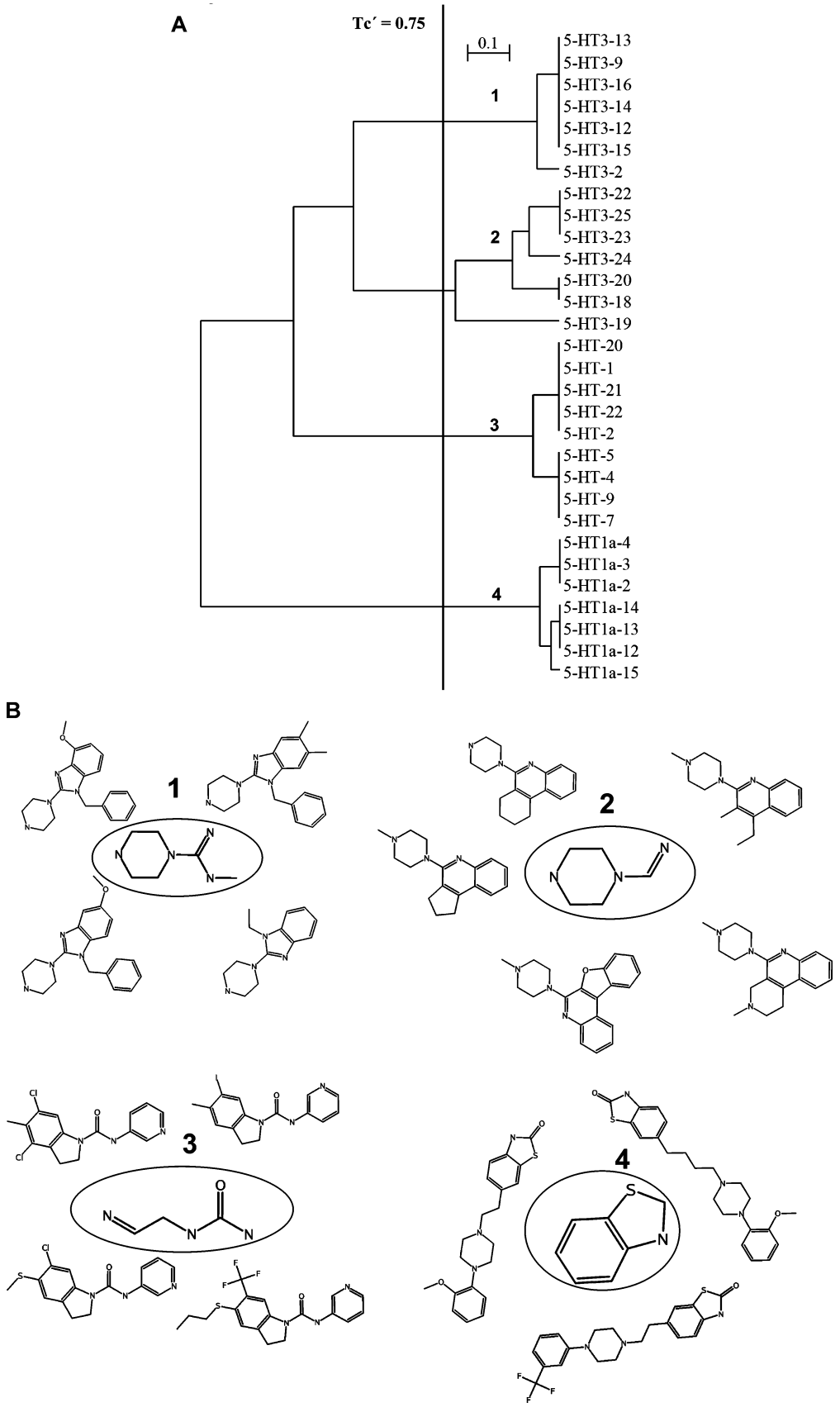

Figure 3. ACCS cluster analysis. (A) Exemplary dendrogram for class 5HT clustered at the $core_0$ level. The vertical black line represents a *Tc′* cutoff value of 0.75, which corresponds to a branch length distance of 0.25. In (B), the $core_{70}$ of four clusters labeled 1−4 in (A) is circled and surrounded by examples of molecules belonging to the cluster.

Figure 1 shows a color code for cores of increasing match rate. Accordingly, *$core_x$ set* is the ACCS subset that contributes atoms to $core_x$ of an active molecule and the *$core_x$ set fraction* is the size of the $core_x$ set divided by the size of the ACCS set. In order to compare the size of individual cores, we calculate a *relative core size*:

$$\text{relative core}_x\ \text{size}(i) = |\text{core}_x|/\text{number of atoms in molecule } i$$

Furthermore, we calculate a *core set regularity score*:

$$\text{regularity score} = 1 - ((|\text{core}_0\ \text{set}| - |\text{core}_{90}\ \text{set}|)/|\text{ACCS set}|)$$

The maximum possible regularity score of 1 means that the core set of a molecule does not change over varying core levels, whereas low scores indicate increases in the core set size.

The evaluation of core set regularity was complemented by studying *core continuity* accounting for the number of distinct core regions that are mapped in a molecule over all core levels. This was accomplished by determining graph coherence of cores. If the graph was split, distinct core regions existed, otherwise the core was contiguous. Graph coherence on the basis of SMILES[18] strings was computed using the Molecular Operating Environment (MOE).[19] Analysis of SMILES strings permits easy detection of nonbonded fragments. Computer graphical representations of core regions were also generated with MOE.

As a measure for the location of cores in a molecule, we also calculated a *core center score*. This scoring function was designed to reflect whether cores mapped to central regions of a test molecule, i.e., its "true" core regions, or peripheral moieties. For each molecule, the most centrally located atoms were determined using built-in graph functions of MOE; from the 2D molecular graph, the vertices with minimal eccentricity were selected. Then the smallest cores were mapped by stepwise addition of all atoms bonded to the central atoms until the core was covered. Then the center score for a molecule was calculated as

$$\text{core center score} = 1 - ((\text{extended central atoms} - 1)/\text{total number of atoms})$$

If a core maps to the center of a molecule, a score close to 1 is obtained. By contrast, scores close to 0 indicate a peripheral location. Figure 2 shows representative examples of varying core locations in test molecules and corresponding center scores.

Generation of ACCS Fingerprints and Cluster Analysis. In order to quantitatively compare the relationships between core sets in each activity class, bit string representations were generated, where each bit position accounted for the presence or absence of an individual fragment of the corresponding ACCS set. As a distance metric for cluster analysis of these prototypic fingerprints, a modified version of the Tanimoto coefficient (*Tc*)[20] was used that averages over bits that are set on and off:

$$Tc'(\text{A,B}) = (Tc(\text{A,B}) + Tc(\text{not A, not B}))/2$$

The conventional Tanimoto coefficient is defined as $Tc(\text{A,B}) = c/(a + b - c)$, with a being the number of bits set on in fingerprint A, b the number of bits set on in B, and c the number of bits set on in A and B. In the Tc' variant, "not A" and "not B" refer to bits set off in A and B, respectively.

Fingerprints were clustered using the Weighted Pair Group Method with Arithmetic Mean (WPGMA).[21] Here the conventional branch lengths ($d/2$) were doubled so that a Tc' cutoff value could be easily incorporated into dendrograms calculated for visualization.[22] Figure 3A shows a representative example. A program originally developed to compare phylogenetic trees[23] was used to compare the dendrograms resulting from ACCS clustering. The program evaluates and compares tree topologies and returns a score from 0% to 100%. The top score reflects the topological identity of the two tree representations. For each activity class, we carried out pairwise comparisons of dendrograms at subsequent core levels (i.e., 0 with 10, 10 with 20, ..., 80 with 90). Dendrograms with scores greater than 95% were considered equivalent (i.e., producing the same or nearly the same substructure grouping), and the smallest core producing a cluster distribution equivalent to core_0 is termed the *stable core*.

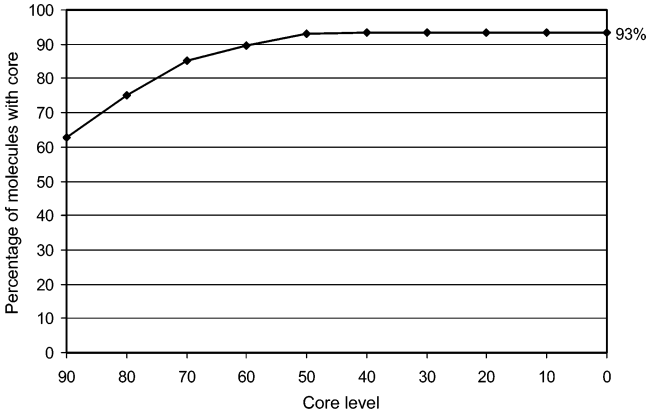

Figure 4. First core occurrence. The fraction of all active molecules with a core having at least one atom is reported over all core levels. The graph shows that 93% of the active molecules form a core after fragment mapping and that more than 60% of all molecules have a core_{90} (i.e., at the most stringent core level). Furthermore, all molecules forming cores do so at the latest at the core_{40} level (where the curve reaches its plateau).

RESULTS AND DISCUSSION

Core Formation. Systematic substructure mapping applying the procedure illustrated in Figure 1 revealed that for many different activity classes regions of fragment overlap were not scattered over molecules, but formed coherent cores. Representative examples are shown in Figure 2. These findings suggested that combinations of substructures that represent activity-class-specific pathways in randomly generated fragment populations[15] have a structural meaning and encode molecule-specific information through the delineation of characteristic core regions. Therefore, we subdivided mapped cores into 10 levels of increasing substructure overlap density, with core_{90} having the highest and core_0 the lowest, and carried out a detailed analysis of core formation by ACCS sets.

Substructure and Core Statistics. A summary of substructure and core statistics is provided in Table 1 and Supplementary Table S2 reports the results for all 45 activity classes. ACCS sets were found to vary considerably in size, depending on the activity class; they contained as few as 39 and as many as 2070 fragments. Changing the ZINC background molecule sets that were cofragmented with active compounds led to very little, maximally 5%, variation in ACCS-dependent core formation. Thus, the distribution of

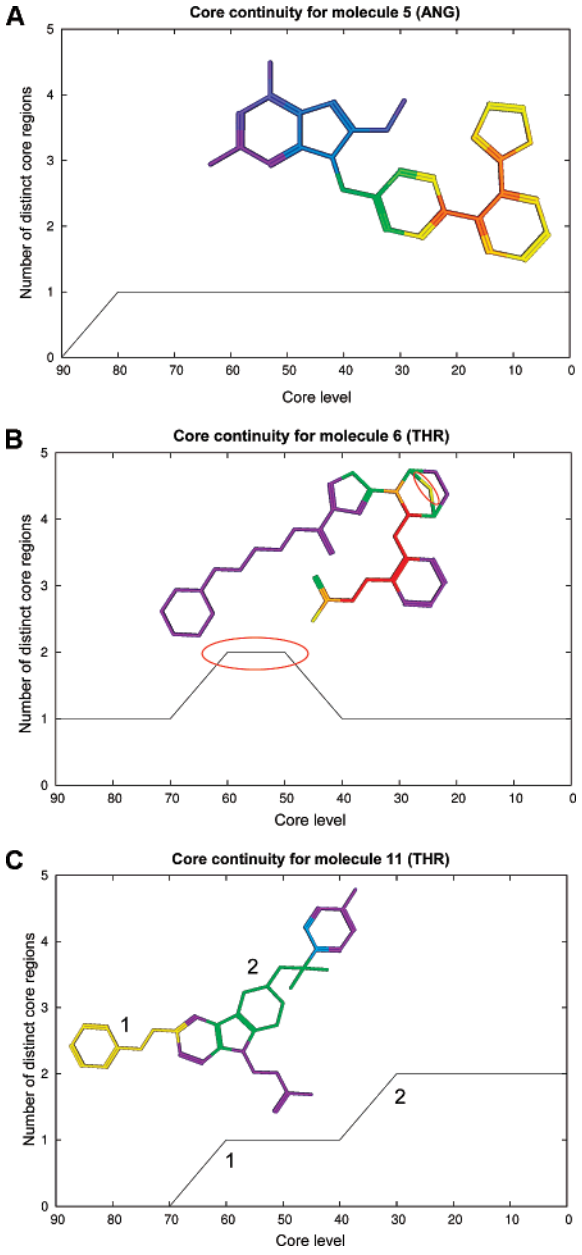

Figure 5. Core continuity. "Core development graphs" are shown for three exemplary molecules. These graphs report the core levels where distinct ACCS core regions are formed and depict these cores. The example in (A) represents the case most commonly observed in our analysis: a single core grows contiguously starting at the $core_{80}$ level. In (B) a second $core_{60}$ is formed (indicated by the red ellipse) and merges with the first core at the $core_{40}$ level. (C) shows an example where two distinct cores are formed at the $core_{70}$ (1) and $core_{40}$ (2) levels that do not merge. This example represents the least frequently observed case.

characteristic substructures was not sensitive to the composition of randomly chosen background database molecules. Figure 4 shows that 93% of all active molecules formed a core after fragment mapping. However, Table 1 also shows that for an individual molecule on average only one-third of an ACCS set, and in no case more than 66%, contributed to the formation of cores, as further discussed in the following. These observations indicated that cores were specifically formed by ACCS subsets.

Table 2 reports average core sizes at different core levels on a per molecule basis, and Supplementary Table S3 reports core sizes for all activity classes. From the most stringent

Table 1. Substructure and Core Statistics for 45 Activity Classes

	ACCS set size	core set fraction	regularity score	center score	stable core level
mininum	39.00	0.15	0.43	0.25	0.00
maximum	2070.00	0.66	1.00	0.92	90.00
mean	553.24	0.33	0.84	0.64	49.33
SD	474.61	0.15	0.16	0.19	27.75

Table 2. Average Core Size[a]

core level	90	80	70	60	50	40	30	20	10	0
core size	9.35	16.28	24.76	32.97	40.57	49.03	56.46	63.98	71.48	82.67

[a] For each core level, the average core size of all active compounds is reported as the percentage of the total number of atoms per molecule.

Table 3. Core Continuity[a]

max no. of distinct core regions	0	1	2	3	4
molecules (%)	5.85	79.12	14.54	0.20	0.29

[a] The percentage of active molecules displaying one to four distinct core regions (over all core levels) is reported. The table shows that more than 79% of all molecules only have a single contiguously growing core region.

($core_{90}$) to the most generic ($core_0$) levels, about 10%−80% of atoms in a molecule contributed to the formation of cores. Thus, those cores that were shared by most compounds belonging to an activity class (e.g., $core_{90}$, $core_{80}$) represented small regions of recurrent fragment overlap patterns.

Core Set Regularity and Core Continuity. Table 1 shows that the average core set regularity was high (0.84). This means that core sets did not substantially change over different core levels. Figure 4 shows that more than 60% of all active molecules had a $core_{90}$. All cores appeared at core levels equal to or greater than $core_{40}$. When core levels change, cores grow through addition of atoms with a lower match rate, but the corresponding fragment sets remain largely unchanged. Table 3 and Supplementary Table S4 report average and class-dependent core continuity, respectively. The data show that 79% of all active molecules only have a single core that exclusively grows through the addition of atoms that are directly bonded to core atoms. An example of this by far most frequently occurring case is shown in Figure 5A. By contrast Figure 5B,C provides examples that only occur in less than 20% of test molecules. In these cases, merging or distinct cores are mapped. Distinct cores were only very rarely seen. Our observations that core sets did not significantly change over different core levels and that mostly single contiguous cores were mapped provided clear indications of class-specific core formation. This was consistent with our findings that class-specific fragment pathways generally represented unique substructure combinations.

Cluster Analysis and Core Stability. The generation of fingerprints for all ACCS sets enabled us to carry out clustering of characteristic substructures for each activity class on the basis of Tanimoto-like similarity. Systematic cluster analysis, as illustrated in Figure 3A, produced dendrograms for activity classes that displayed highly conserved topology beginning at specific core levels. Conserved topology of dendrograms reflects constant core set distributions within a compound class, which is a measure of core stability and further complements the analysis of core

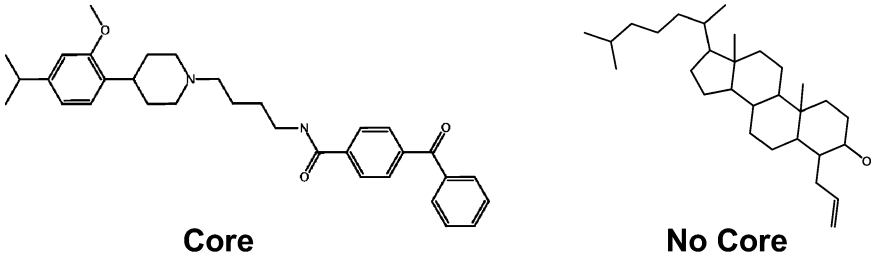

Figure 6. Molecular flexibility and core formation. Two compounds from activity class LDL are shown. In this case, core formation was observed for the more flexible molecule but not for the compound having the rigid steroid-type scaffold.

set regularity and core continuity. Stable cores were often observed at stringent core levels. Supplementary Table S2 shows that, for six activity classes, $core_{90}$ was already stable and that many others had stable cores at high levels. On average, dendrogram topology did not change for cores beyond $core_{40}$ (Table 1), consistent with the results discussed above.

Core Locations. The predominant formation of single cores by characteristic substructures also raised the question of where these regions were located in active molecules. With the center score, we devised a metric that related mapped cores to the most centrally located atoms in test compounds, as illustrated in Figure 2. For many but not all activity classes, center score calculations detected a tendency of cores to map to central regions of active molecules. Supplementary Table S2 shows that 20 compound classes reached center scores between 0.75 and 0.92 that indicate the presence of centrally located cores. However, there were also classes with low center scores: eight classes had a center score of less than 0.4. These findings demonstrate that central atoms did not always contribute to cores, although they displayed a general tendency to do so. Systematic graphical analysis of cores revealed that they did not resemble hierarchically derived molecular scaffolds.[9,10] Rather, these cores were often limited to regions close to the central atoms of active molecules. Typically, these regions have high topological complexity, and we found that they were a preferred source of ACCS. However, not every central and/or topologically complex region was part of a core. This also showed that ACCS were typically focused on limited molecular regions that are shared by active molecules and confer class-specific information. Topologically complex regions in molecules are often rigid. Comparison of molecules within activity classes having significantly different flexibilities revealed that core formation was not systematically influenced by such differences. In some cases, flexible molecules formed cores, whereas rigid ones did not. An example is shown in Figure 6. Furthermore, fragment mapping does not provide information about parts of molecules other than formed cores. For example, recurrent structural patterns might exist outside mapped cores that are shared by active molecules but are not part of class-specific fragment pathways. However, the design of our mapping analysis precluded the possibility that recurrent structural patterns were identified that simultaneously occurred in different activity classes.

CONCLUSIONS AND PERSPECTIVE

We have analyzed substructures from class-specific pathways in random molecular fragment populations. Such pathways can be isolated from fragment populations of many different classes of compounds. Systematic mapping of these substructures for more than 1000 molecules belonging to 45 diverse activity classes revealed that these substructures usually form coherent molecular core regions. Thus, fragment combinations that are statistical signatures of compound classes were found to have a defined structural meaning: they encode well-defined common cores in active molecules that are often stable when increasing numbers of fragments are mapped. These findings demonstrate that random fragment populations encode specific structural information and explain why random fragment profiles can be used to detect and distinguish between different structure−activity relationships.[13] Moreover, the results presented herein provide a basis for the derivation of compound-class-directed sets of structural descriptors that do not depend on conventional design schemes such as hierarchical or retrosynthetic molecular fragmentation. Substructures are known to be powerful descriptors for similarity searching, compound classification, and the study of structure−activity relationships. Our findings extend the repertoire of currently available substructures and demonstrate that it is possible to identify sets of substructures to target selected activity classes.

Supporting Information Available: Supplementary Tables S1−S4 report the composition and source of each activity class, their ACCS set and core properties, individual core sizes, and core continuities, respectively. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Barnard, J. M.; Downs, G. M. Chemical fragment generation and clustering software. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141−142.
(2) 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.
(3) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.
(4) Merlot, C.; Domine, D.; Cleva, C.; Church, D. J. Chemical substructures in drug discovery. *Drug Discovery Today* **2003**, *8*, 594−602.
(5) Brown, R. D.; Martin, Y. C. An evaluation of structural descriptors and clustering methods for use in diversity selection. *SAR QSAR Environ. Res.* **1998**, *8*, 23−39.
(6) Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233−245.
(7) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882−894.
(8) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity searching using reduced graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338−345.
(9) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(10) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The scaffold tree—visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* **2007**, *47*, 47−58.

(11) Lewell, X. Q.; 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.
(12) 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.
(13) Batista, J.; Godden, J. W.; Bajorath, J. Assessment of molecular similarity from the analysis of randomly generated structural fragment populations. *J. Chem. Inf. Model.* **2006**, *46*, 1937−1944.
(14) Batista, J.; Bajorath, J. Chemical database mining through entropy-based molecular similarity assessment of randomly generated structural fragment populations. *J. Chem. Inf. Model.* **2007**, *47*, 59−68.
(15) Batista, J.; Bajorath, J. Mining of randomly generated molecular fragment populations uncovers activity-specific fragment hierarchies. *J. Chem. Inf. Model.* **2007**, *47*, 1405−1413.
(16) 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.
(17) PerlMol—Perl Modules for Molecular Chemistry. http://www.perlmol.org (accessed April 2007).
(18) 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.
(19) *Molecular Operating Environment (MOE)*; Chemical Computing Group, Inc.: Montreal, Quebec, Canada, 2006.
(20) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(21) Eidhammer, I.; Jonassen, I.; Taylor, W. R. Multiple Global Alignment and Phylogenetic Trees. In *Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure Analysis*, 1st ed.; John Wiley & Sons Ltd.: West Sussex, England 2004; p 83.
(22) Perrière, G.; Gouy, M. WWW-Query: an on-line retrieval system for biological sequence banks. *Biochimie* **1996**, *78*, 364−369.
(23) Nye, T. M.; Liò, P.; Gilks, W. R. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics* **2006**, *22*, 117−119.

CI700251B