Topological Fragment Index for the Analysis of Molecular Substructures and Their Topological Environment in Active Compounds

Eugen Lounkine and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology & Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrassse 2, D-53113 Bonn, Germany

Received July 30, 2008

We report the development and application of the Topological Fragment Index (ToFI), a measure for the complexity of the topological environment of defined molecular fragments in active compounds. On the basis of ToFI calculations, RECAP fragments are organized in dependency hierarchies that capture fragment co-occurrence and facilitate the identification of topology clusters and activity class characteristic RECAP fragments. By combining structural and topological environment information through ToFI, RECAP fragments that are a signature of compounds active against one of several closely related targets are consistently identified.

INTRODUCTION

Molecular substructures are widely recognized as powerful descriptors that can be associated with biological activity and are interpretable in a chemically intuitive manner.[1−3] For example, in virtual screening, fragments are often used as structural keys for the detection of similar molecules. Such structural keys are usually encoded as bit string representations that record the presence and absence of individual fragments in a molecule. For pairwise comparison of molecules, bit string overlap is quantified using similarity coefficients, which serves as a measure of molecular similarity.[4]

In addition to using fragments as structural keys, the analysis of fragments that are recurrent in compounds having similar activity across protein target families has led to the identification of so-called privileged substructures.[5−7] Furthermore, co-occurrence of selected fragments is often a signature of individual compound classes,[8] and fragment pairs or triplets have been found to carry much activity class-specific information.[9]

Compound sets can be organized based on different chemotypes[10] or hierarchies of molecular fragments.[11,12] By doing so, individual substructures can be identified that are associated with specific biological activities. Recently reported substructure organization schemes include the Scaffold Tree introduced by Schuffenhauer et al.[11] and hierarchical organization on the basis of fragment co-occurrence in chemical databases.[12] The latter approach organizes fragments based on their conditional probabilities of occurrence. Fragment dependency relationships are encoded in dependency graphs that incorporate activity information of fragments and enable the identification of activity class-specific fragment pathways.[10] This approach has revealed that random fragment populations contain activity class-specific information that is associated with fragment combinations and frequencies of fragment occurrence.[12,13]

Herein we introduce the Topological Fragment Index (ToFI) that quantitatively accounts for the topological environment of substructures in test molecules. This is achieved by a systematic analysis of the bonding pattern of a given fragment within its source molecule. ToFI is applicable to any type of molecular fragments, regardless of how they are derived. On the basis of ToFI calculations, user-defined fragments can also be organized in hierarchies that are based on fragment co-occurrence and define fragment pathways that are specific for individual activity classes. Here we apply ToFI to RECAP[14] fragments that were originally designed on the basis of retrosynthetic considerations. For five different sets of related targets, RECAP fragments were consistently identified that distinguished inhibitors of each target from other closely related ones. ToFI analysis can be practically applied to mine however defined fragment populations for fragments that are specific for compound activity classes.

METHODS

Topological Fragment Index (ToFI). We introduce the Topological Fragment Index (ToFI) as a measure of the topological environment information of a fragment within its molecular context. For a given fragment and molecule, all instances of the fragment in the molecule are identified using subgraph matching. In principle, three parameters account for fragment generation and topological environment information:

1. the total number of bonds in the molecule (n),

2. the number of bonds that must be cleaved in order to obtain the fragment (k), and

3. the number of bonds that are not permitted to be cleaved because atoms connected by these bonds constitute the fragment (l).

On the basis of these parameters and for a given fragment instance, ToFI is calculated as follows

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

10.1021/ci8002599 CCC: $40.75
Published on Web 11/14/2008

Table 1. Data Set Composition[a]

name	description	activities	activity codes	molecules	mapped fragments	graph fragments	ACCRF
CCK	CCK A/B agonists/antagonists	4	42705, 42706, 42712, 42713	730	1268	203	59
dopamine	dopamine D1−4 antagonists	4	07702, 07701, 07703, 07710	1557	1789	406	111
MAO	MAO A/B inhibitors	2	08410, 08420	252	544	148	80
opioids	$\kappa/\delta/\mu$ agonists	3	01131, 01132, 01133	711	1480	231	62
PDE	phosphodiesterase I−IV inhibitors	4	78415, 78416, 78417, 78418	2567	1836	444	167

[a] Five supersets containing a varying number of activity classes ("activities") against related targets, as defined by MDDR "activity codes", were used for the generation of dependency graphs. Reported are the total number of molecules per superset ("molecules"), the number of successfully mapped substructures ("mapped fragments"), and the number of fragments remaining in the graph after deletion of paths that did not terminate at a molecule ("graph fragments"). "ACCRF" refers to the number of Activity Class Characteristic RECAP Fragments found in the graph.

$$ToFI_{Instance}(n,m,k) = \frac{1}{n+1}\sum_{i=0}^{n-m}\frac{1}{\binom{n}{k+i}}\binom{n-m}{i}$$

Here m is the sum of the number of bonds that must be deleted or are not allowed to be deleted ($m = k + l$) to isolate the fragment from its source compound; the variable i denotes the number of other bonds that are deleted in addition to k but do not influence the separation of the given fragment instance. The quotient term before the sum renders ToFI calculations independent of the overall number of bonds in a molecule. Figure 1A illustrates exemplary ToFI calculations. The overall ToFI value of a given fragment is the sum of ToFI values for its instances. In order to calculate ToFI for a given pair of a molecule and fragment, the fragment is first mapped onto the molecule, the bonds relevant for ToFI calculations are identified as shown in Figure 1A, and the function is applied. The larger the ToFI value becomes, the less complex is the topological environment of a fragment. In the context of ToFI calculations, the complexity of a fragment is increasing with the number of bonds within the fragment and the number of bonds between fragment atoms and other atoms in the molecule. Therefore, it is also possible to interpret ToFI values as the likelihood for any given fragment to be isolated from source molecules by randomized bond cleavage.

In order to control computational cost and provide easily interpretable ToFI scores, floating point ToFI values are transformed into integers by multiplying each calculated value with an empirically chosen constant of 10^6 and rounding the resulting value to the nearest integer. According to this interpretation, a ToFI value of five means that if the compound was randomly fragmented one million times, the fragment of interest would have been generated five times. Figure 1B shows three exemplary ToFI calculations for a benzene fragment in three different compounds. Although the fragment frequency for the fragment is two in all three cases, its topological environment is different in these molecules, which is reflected in distinct ToFI values.

RECAP Derived Substructures. We performed RECAP[14] analysis of the MDDR using the Molecular Operating Environment (MOE).[15] Figure 2 illustrates an exemplary RECAP fragmentation, and Supporting Information Figure 1 reports the extended RECAP fragmentation rules and the sixteen atom environment categories utilized in MOE RECAP analysis. In RECAP analysis, bonds are cleaved between any two atoms belonging to the same environment category according to retrosynthetic criteria. In Figure 2, two distinct environments, "4 urea" and "1 amide" are identified, and the corresponding bonds are cleaved. Source molecules and fragments were used in hydrogen suppressed form. Fragments that were generated less than five times or that contained more than twenty heavy atoms were omitted, yielding a total number of 10,246 RECAP fragments derived from ~159,000 MDDR compounds. We encoded each RECAP fragment using recursive SMARTS,[16] which provides not only structural but also chemical environment information of those atoms forming bonds that are cleaved during RECAP fragmentation. For substructure representation, isotope-like labels were used to denote individual atom environments. Figure 2 also illustrates the mapping of generated RECAP fragments on test molecules.

Dependency Graphs. A previously described methodology[12] has been adopted in order to organize RECAP fragments in directed acyclic dependency graphs. Dependency graphs have been designed to account for fragment co-occurrence in molecules. In other words, they report subsets of fragments that only occur together with others, i.e. a fragment A must be present (conditional) for a fragment B to occur (dependent). Fragment pathways identify fragments that co-occur in active compounds. For the purpose of our analysis, only fragment pathways are considered that terminate at complete molecules. In pathways that are annotated with compound activity information, activity class characteristic RECAP fragments can be identified. Fragment dependencies are quantified based on ToFI values using the following formalism. Given N compounds, each fragment is represented as an N-dimensional vector where each component reports the ToFI score for a specific molecule in the data set. For each fragment, its dependency on other fragments is expressed as the quotient of the respective ToFI components. A fragment *dep* only depends on a fragment *cond* if all vector components of *dep* are smaller or equal to the respective components of *cond* (and at least one component of *dep* is smaller than the corresponding component of *cond*).

$$dependency(dep, cond) = \delta\sum_{i=1}^{N}\frac{ToFI_i(dep)}{ToFI_i(cond)}$$

Here, $ToFI_i$ is the overall ToFI (sum over all instances) in molecule i and N is the total number of molecules. The δ operator controls the case when $ToFI(cond)$ is smaller than $ToFI(dep)$.[12] Molecules are treated as "superfragments" in the dependency calculation. Accordingly, the vector component that reports the frequency for such a molecule is set

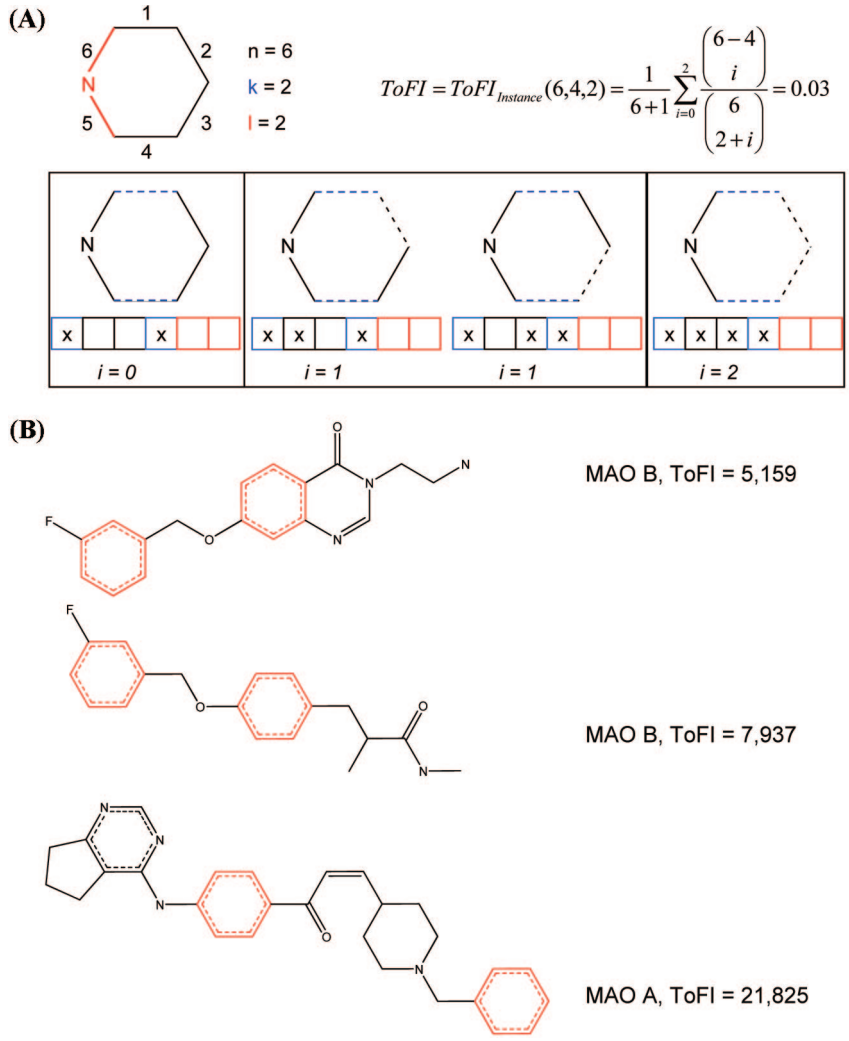

Figure 1. Exemplary ToFI calculation. (A) For piperidine, shown at the top left, the ToFI calculation for a dimethylamine fragment (red) is presented. Bonds in the molecule are numbered 1 to 6. Bonds 5 and 6 are not allowed to be deleted because the bonded atoms constitute the fragment, whereas bonds 1 and 4 must be deleted in order to isolate the fragment. Accordingly, the parameters n, k, and l, are determined, as described in the text. The lower panel shows all possible bond deletion combinations that give rise to the fragment. The small boxes represent the bonds, and an "x" is placed within a box whenever the corresponding bond is deleted. Blue boxes (1 and 4) always contain an "x", and red boxes (5 and 6) are never allowed to contain an "x". For the remaining two black boxes (2 and 3), all combinations of possible deletions are enumerated. In total, two, three, or four bond deletions are possible, and two combinations exist for the deletion of three bonds. (B) For two MAO B and one MAO A inhibitor, ToFI values were calculated for a benzene fragment. The ToFI values differ, which reflects the different topological context the benzene fragments occur in (whereas frequency counts are "two" in all three cases).

to 1 and all others are set to 0. For each fragment and molecule, a set of conditional fragments is extracted based on maximal dependency values. The fragment(s) that *dep* maximally depends on become its conditional fragments. These relationships are represented in a directed acyclic graph. In this graph, edges connect fragments that are dependent on each other. Fragment pathways in the graph that do not terminate at a molecule are deleted because they contain fragments that have low signature character for active molecules. The fragments are annotated based on their occurrence in compounds with different biological activity. Activity Class Characteristic RECAP Fragments (ACCRF) are defined as those fragments that exclusively occur in an individual activity class (and are color-coded in graph representations). Dependency graph depictions were generated using the freely available software Tulip.[17]

Graph Characterization. In order to describe the topology of a dependency graph, we have analyzed the graph coherence and distribution of ACCRF in subgraphs. These subgraphs are defined on the basis of root fragments. Each subgraph consists of all nodes and edges that originate from the corresponding root fragment. For each subgraph, the number of fragments and the distribution of ACCRF belonging to different activity classes have been determined. Subgraph overlap is quantified based on the number of shared fragments and/or molecules. Furthermore, edges are annotated with substructure relationship information. Four categories of substructure relationships are distinguished for a pair of a conditional (parent) and a dependent (child) fragment: (1) the parent is a substructure of the child, (2) the child is a substructure of the parent, (3) the fragments are identical but differ in the environmental annotation of individual atoms, and (4) no structural relationship exists.

Data Sets. Compound activity classes were assembled from the Molecular Drug Data Report (MDDR)[18] on the basis of the MDDR Activity Codes. For five "supersets" consisting of between two and four related targets (for example, dopamine receptors D1-D4), sets of compounds reported to be active against each individual target were selected, as summarized in Table 1.

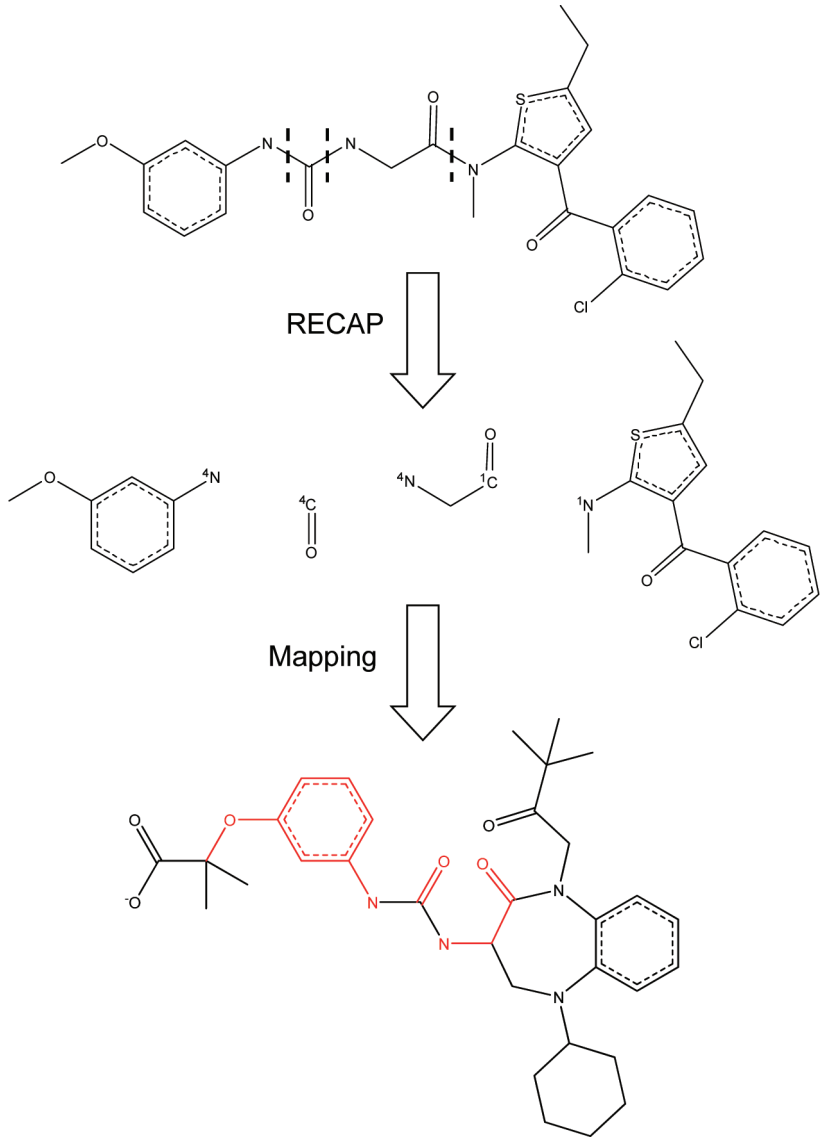

Figure 2. Exemplary RECAP fragmentation and mapping. The RECAP fragmentation of the compound at the top is shown. Three bonds (two urea and one amide bond) are cleaved according to RECAP rules, leading to four fragments with annotated environments. These fragments are mapped onto another compound at the bottom (red substructures). The isotope-like labels define RECAP atom types.

Table 2. Substructural Relationships of Dependent Fragments[a]

	CCK	dopamine	MAO	opioids	PDE	average
no subgraph	67.2	60.2	56.4	63.7	52.4	60.0
parent in child	25.2	28.4	33.4	27.0	37.6	30.3
child in parent	0.3	0.7	0.7	0.7	0.2	0.5
equal	7.2	10.8	9.5	8.6	9.8	9.2

[a] For each superset, the percentage of edges is shown that connect fragments having four distinct structural relationships: "no subgraph", no sub- or supergraph relationship; "parent in child", the parent is a subgraph of the child; "child in parent", the child is a subgraph of the parent; "equal", structural equivalence of fragments.

Table 3. Topology Cluster Environment Distribution[a]

environment	CCK	dopamine	MAO	opioids	PDE	average
1	83.8	96.5	95.0	97.6	94.0	93.4
2	100.0	100.0	100.0	100.0	91.7	98.3
3		53.7	91.3	86.8	90.0	80.5
4	100.0		100.0	90.0	93.8	95.9
5	86.1	85.1	78.0	92.3	66.7	81.6
6		78.6	83.3	100.0	100.0	90.5
8				100.0	87.8	93.9
10		100.0			98.4	99.2
11	100.0				75.0	87.5
12					100.0	100.0
15				100.0	84.6	92.3

[a] For each superset, the percentage of fragments is reported that contains at least one atom having the same topological environment (designated according to Supporting Information Figure 1) as the corresponding root fragment. Eleven of 16 possible RECAP atom environments were present in root fragments.

RESULTS AND DISCUSSION

Topological Fragment Index (ToFI). ToFI is designed to be a quantitative measure of the topological environment of a fragment. The smaller the ToFI value becomes, the more complex is the topological environment of a given fragment. Fragments with equal frequency of occurrence are further distinguished by ToFI, if they occur in distinct topological contexts. Hence, ToFI extends frequency counts of fragments in molecules by incorporating information about the bonding patterns surrounding the fragment. Examples are provided in Figure 1B. Thus, ToFI allows for the assessment of compound specific topological fragment information, which

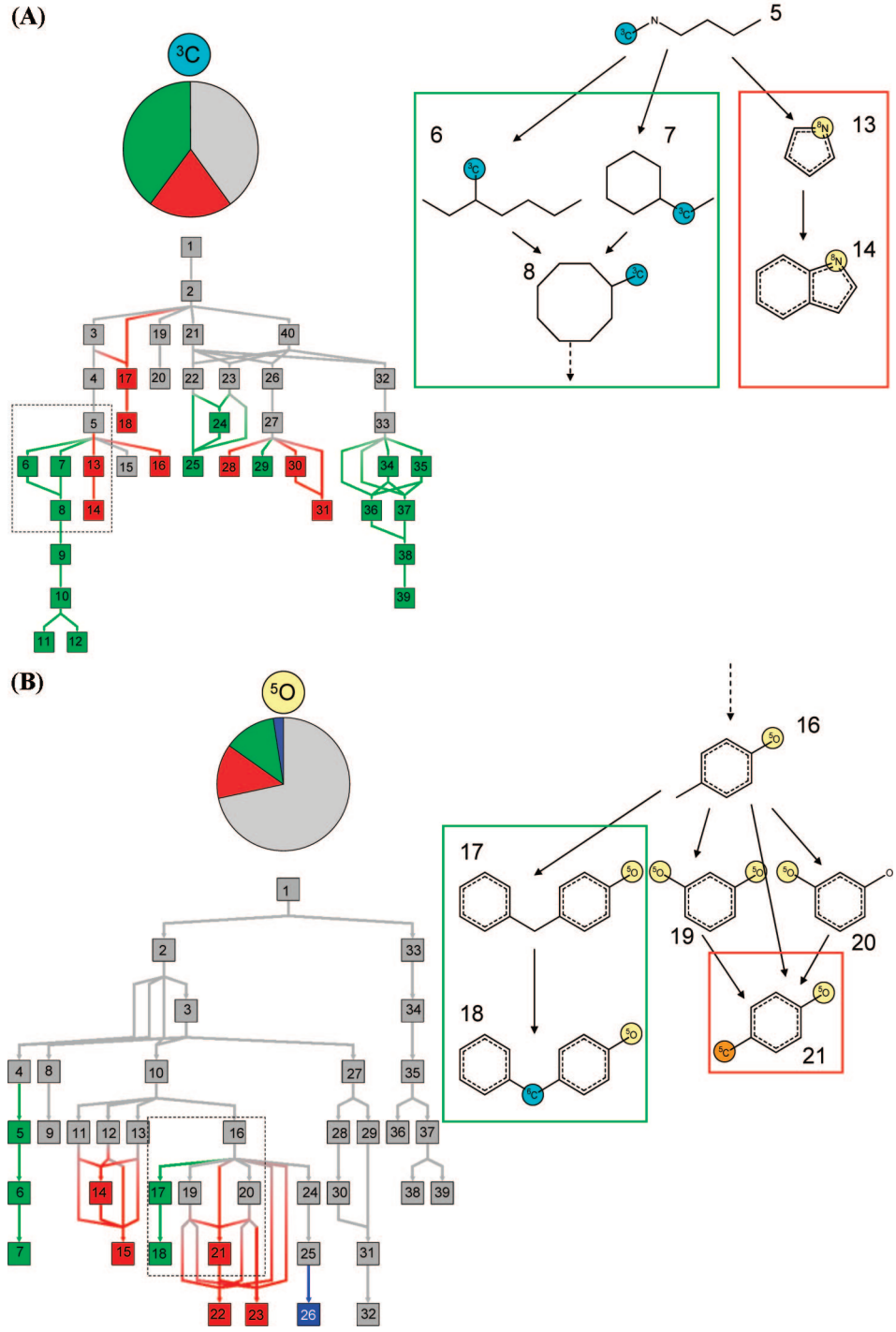

Figure 3. Exemplary subgraphs. (A) For the superset "MAO", the subgraph defined by root fragment "^{3}C" is shown. Numbers are fragment identifiers. The pie chart above the subgraph represents its fragment distribution. Grey segments correspond to generic fragments, i.e. fragments that occur in more than one activity class. Color-coded segments report the distribution of ACCRF (red: MAO A, green: MAO B). On the right, fragments are shown that correspond to the boxed region in the subgraph. The isotope-like labels define RECAP atom types. Atom environments in fragments are also color-coded (blue: "^{3}C", yellow: "^{8}N"). (B) For the superset "opioids", the subgraph defined by the root fragment "^{5}O" is shown. The ACCRF segment color-code is red: κ, green: δ, blue: μ; the atom environment color-code is blue: "^{6}C", yellow: "^{5}O", orange: "^{5}C".

cannot be detected by substructure counts. We have applied ToFI to a set of more than 10,000 substructures derived from RECAP fragmentation of the MDDR. The calculations were performed for five compound supersets active against closely related targets. RECAP derived substructures generated on the basis of the MDDR were identified that exclusively occurred in individual activity classes.

Dependency Hierarchies. We have assessed ToFI value distributions of RECAP fragments among our activity classes using a previously described methodology for the hierarchical organization of fragments.[12] The hierarchy is encoded in a directed acyclic graph that reports dependency relationships between fragments. Fragment dependency is calculated based on the systematic comparison of fragment ToFI value distributions among active compounds. Quantifying topological fragment similarity also accounts for fragment co-occurrence, which is visualized in dependency graphs, because ToFI incorporates information about both the

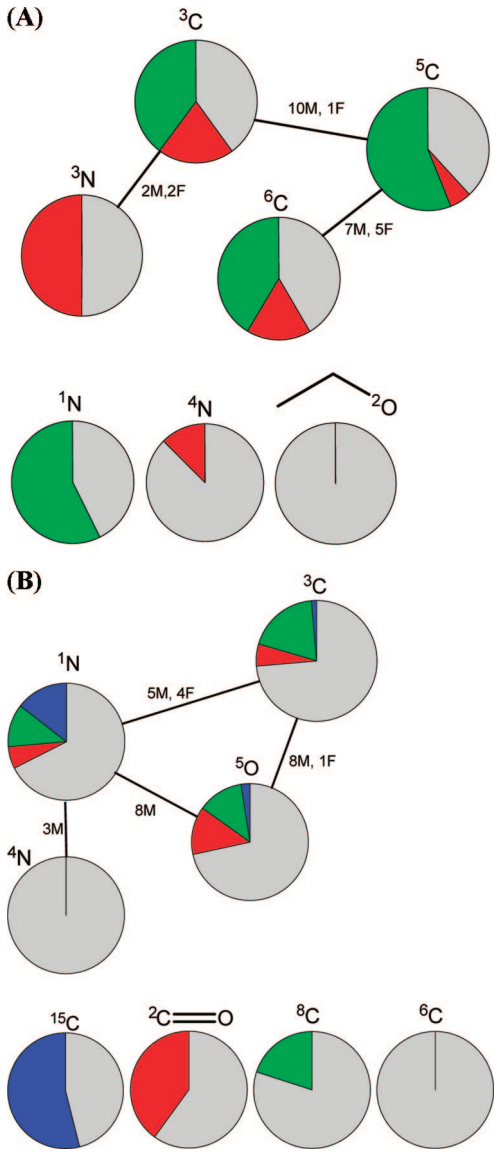

Figure 4. ToFI topology clusters. Each topology cluster is represented as a pie chart according to Figure 3. Pie charts are connected if the clusters share nodes and are annotated with the numbers of shared fragments (F) and/or molecules (M): (A) "MAO" and (B) "opioids".

frequency and topological context of fragments within active compounds. Fragments that are connected in the dependency graph most strongly depend on each other, i.e. they are most similar with respect to their topological context and distribution among active compounds. An edge between a fragment A and fragment B in the dependency graph indicates two relationships: first, fragment A is always present in a molecule, if fragment B is present; second, of all fragments that co-occur with fragment A, fragment B has the highest likelihood to do so.

We have assessed the structural relationships between dependent fragments. As reported in Table 2, on average, ~60% of the detected dependencies did not correspond to a structural (sub- or supergraph) relationship, as illustrated in Figure 3A for fragments 5 and 13. The second largest subset of fragment dependencies (30%) included fragment pairs where the dependent (child) fragment contained the conditional (parent) fragment. An example is shown in Figure 3A (i.e., fragments 13 and 14). This finding demonstrates that

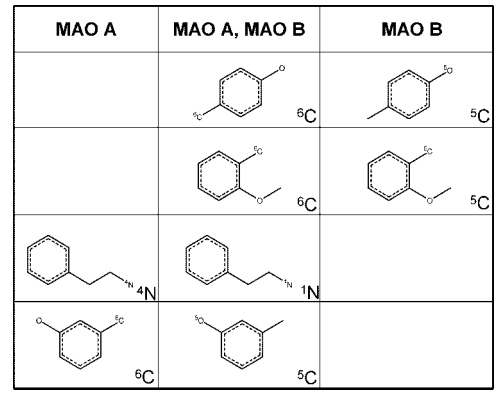

Figure 5. Topology cluster distribution of ACCRF. For the "MAO" superset, four examples of structurally identical fragments are shown that belong to different topology clusters identified by ToFI and that differ in their biological activity. The root fragment of the topology cluster is reported in each cell.

ToFI-based fragment hierarchies go beyond a description of substructure matches but rather reflect fragment similarity in terms of distribution and topological environment criteria among active compounds. Figure 3A,B shows exemplary subgraphs for the supersets "MAO" and "opioids", respectively. Dependency graph examples for all supersets and depictions of the corresponding fragments are provided in Supporting Information Figure S2.

Activity Class Characteristic RECAP Fragments. ToFI-based RECAP fragment hierarchies enable the identification of RECAP fragments that are characteristic of individual activity classes (ACCRF). These fragments constitute activity class specific fragment dependency pathways because they exhibit ToFI distributions that have signature character for different sets of active compounds. For the supersets "MAO", "opioids", and "dopamine" ACCRF have been identified for all activity classes, and for both "CCK" and "PDE", ACCRF were found for three of four classes (except for CCK B agonists and PDE II inhibitors).

Fragment Topology Clusters. ToFI organization of fragments makes it possible to analyze subgroups of topologically related fragments independently of each other. These ToFI "topology clusters" in a dependency graph correspond to root fragment-defined subgraphs, as described in the Methods section, and are not independent graph representations. The organization of exemplary topology clusters is illustrated in Figure 4A,B for "MAO" and "opioids", respectively. Topology clusters generally share only few, if any fragments. Supporting Information Figure S3 provides the topology clusters for all five supersets with charts that are scaled in size according to the total number of fragments and molecules per cluster. It should be noted that ToFI dependency graphs focus on fragment distributions but do not compare individual molecules.

Systematic analysis of ToFI-based fragment dependencies revealed that ToFI value distributions accurately described the interdependence of RECAP fragments containing atoms with equivalent environment information. Table 3 shows that most fragments (approximately 90%) in ToFI topology clusters shared the same chemical environment of individual atoms with their root fragment. Thus, ToFI value distributions reflect similar topological contexts of fragments that have atoms with the same chemical environment in common.

Distribution of ACCRF in Topology Clusters. We have also analyzed the distribution of ACCRF in topology clusters. In Figure 4A,B, the distribution of ACCRF in individual clusters and cluster overlap are reported. Topology clusters usually have distinct ACCRF distributions. For example, in Figure 4A the cluster "^{1}N" does not contain any ACCRF for MAO A inhibitors, whereas clusters "^{6}C" and "^{3}C" show similar ACCRF distributions, although they do not share any fragments. We further analyzed the presence of activity class characteristic topological environments identified using ToFI by systematic comparison of structurally identical fragments containing atoms with different topological environments. Figure 5 provides examples of such fragments that occur in different MAO activity classes and distinct ToFI topology clusters. For all supersets, structurally identical fragments belonging to different ToFI topology clusters and activity classes were identified.

CONCLUSIONS

The Topological Fragment Index is introduced as a measure of the topological environment and complexity of user-defined fragments in active compounds. We have applied ToFI to RECAP fragments in the context of five supersets of compounds active against closely related targets. ToFI value distributions among active compounds have been assessed using a hierarchical organization of fragments from which RECAP fragments have been identified that are characteristic of individual activity classes (ACCRF). We have shown that ToFI accounts for topology clusters of fragments with distinct distributions of ACCRF and for activity class inherent fragment relationships that go beyond structural resemblance.

ACKNOWLEDGMENT

We thank José Batista for helpful discussions concerning fragment dependency calculations and critical review of the manuscript.

Supporting Information Available: Extended RECAP fragmentation rules, exemplary subgraphs, and topology clusters for all compound supersets (Figures S1, S2, and S3, respectively). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Barnard, J.; Downs, G. Chemical Fragment Generation and Clustering Software. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141–142.
(2) Merlot, C.; Domine, D.; Cleva, C.; Church, D. J. Chemical Substructures in Drug Discovery. *Drug Discovery Today* **2003**, *8*, 594–602.
(3) Lounkine, E.; Batista, J.; Bajorath, J. Random Molecular Fragment Methods in Computational Medicinal Chemistry. *Curr. Med. Chem.* **2008**, *15*, 2108–2121.
(4) Willett, P.; Barnard, J.; Downs, G. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
(5) Erlanson, D. A.; McDowell, R. S.; O'Brien, T. Fragment-Based Drug Discovery. *J. Med. Chem.* **2004**, *47*, 3463–3482.
(6) Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. Are Target-Family-Privileged Substructures Truly Privileged? *J. Med. Chem.* **2006**, *49*, 2000–2009.
(7) Aronov, A. M.; McClain, B.; Moody, C. S.; Murcko, M. A. Kinase-likeness and Kinase-Privileged Fragments: Toward Virtual Polypharmacology. *J. Med. Chem.* **2008**, *51*, 1214–1222.
(8) Lameijer, E.; Kok, J. N.; Bäck, T.; Ijzerman, A. P. Mining a Chemical Database for Fragment Co-Occurrence: Discovery of "Chemical Clichés". *J. Chem. Inf. Model.* **2007**, *46*, 553–562.
(9) Lounkine, E.; Auer, J.; Bajorath, J. Formal Concept Analysis for the Identification of Molecular Fragment Combinations Specific for Active and Highly Potent Compounds. J. Med. Chem. **2008**, *51*, 5342–5348.
(10) Medina-Franco, J. L.; Petit, J.; Maggiora, G. M. Hierarchical Strategy for Identifying Active Chemotype Classes in Compound Databases. *Chem. Biol. Drug Des.* **2006**, *67*, 395–408.
(11) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M.; Waldmann, H. The Scaffold Tree - Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, *47*, 47–58.
(12) Batista, J.; Bajorath, J. Mining of Randomly Generated Molecular Fragment Populations uncovers Activity-Specific Fragment Hierarchies. *J. Chem. Inf. Model.* **2007**, *47*, 1405–1413.
(13) 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.
(14) 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.
(15) *Molecular Operating Environment (MOE), version 2007.09*; Chemical Computing Group, Inc.: Montreal, Quebec, Canada, 2007.
(16) James, C. A.; Weininger, D. *Daylight Theory Manual*; Daylight Chemical Information Systems Inc.: Aliso Viejo, CA, 2008.
(17) Auber, D. Tulip: A Huge Graph Visualisation Framework. In *Graph Drawing Softwares, Mathematics and Visualization*; Mutzel, P., Jünger, M., Eds.; Springer-Verlag: 2003; pp 105−126.
(18) *Molecular Drug Data Report (MDDR), version 2005.2*; Symyx Software: San Ramon, U.S.A.

CI8002599