Mining of Randomly Generated Molecular Fragment Populations Uncovers Activity-Specific Fragment Hierarchies

José Batista and Jürgen Bajorath*

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

Received March 28, 2007

We introduce a methodology to analyze random molecular fragment populations and determine conditional probability relationships between fragments. Random fragment profiles are generated for an arbitrary set of molecules, and each observed fragment is assigned a frequency vector. An algorithm is designed to compare frequency vectors and derive dependencies of fragment occurrence. Using calculated dependency values, random fragment populations can be organized in graphs that capture their relationships and make it possible to map fragment pathways of biologically active molecules. For sets of molecules having similar activity, unique fragment signatures are identified. The analysis reveals that random fragment profiles contain compound class-specific information and provides evidence for the existence of activity-specific fragment hierarchies.

INTRODUCTION

Defined molecular fragments and substructures have long been used as descriptors for chemical similarity searching or cluster analysis and continue to be very popular tools for computational analysis of small molecules.[1−9] In many applications, structural or fragment-type descriptors perform well,[6−9] which is often attributed to the fact that they implicitly capture much information about chemical and activity-related properties of small molecules.[8] In addition to their use as descriptors, substructures have also been associated with specific biological activities and privileged binding motifs of therapeutic target families.[5,10,11] For the derivation of substructures, hierarchical[12] or retrosynthetic[13,14] computational fragmentation schemes have been devised. For text-based molecular representations,[15] fragmentation methods have also been introduced[16,17] to complement or replace fragment dictionaries[16] or formulate structure−activity rules.[17]

Regardless of their details, a characteristic feature of fragment or substructure methods is that they generally operate on the basis of well-defined structural elements. We have been interested in the design of molecular similarity methods that do not depend on the use of predefined structural or property descriptors and have introduced a methodology termed MolBlaster[18] that generates random fragment populations of molecules, represents them as histograms profiles, and uses an information-theoretic metric[19] to quantitatively compare these fragment profiles as a measure of molecular similarity. Fragmentation is facilitated by randomly deleting bonds in connectivity tables of hydrogen-suppressed 2D molecular graphs. Through comparison of fragment profiles we were able to accurately reproduce similarity relationships between diverse active molecules.[18] We then combined individual fragment profiles of active molecules, generated profiles for different activity classes, and applied these profiles in similarity searching.[20] Comparison of the fragment profile of an activity class with individual profiles of database molecules was made possible through the application of an extension of the Shannon Entropy concept[19] that we termed Proportional Shannon Entropy (PSE).[20] On the basis of these profile comparisons, we were able to correctly identify molecules belonging to a variety of activity classes in database search calculations.[20] Even for activity classes of high structural diversity, significant compound recall was observed, and reference molecules and correctly identified hits were found to contain different scaffolds. These findings indicated that fragment profile searching had considerable scaffold hopping potential.[20]

These studies showed that a gross comparison of random fragment populations and their information content was sufficient to consistently detect similarity relationships between active molecules and distinguish them from other database compounds. Thus, random fragment profiles could serve as a measure of molecular similarity, which was a key finding of these investigations. However, these results also implied that random molecular fragment populations must contain specific yet unknown information about characteristic compound features. This raised a number of questions. What is this information? How is it “hidden” in random fragment distributions? How can we explore and identify specific elements? What can we perhaps learn about class-specific features from information contained in random fragment populations? These questions have prompted us to study the information contained in random fragment populations and develop an approach for the detailed analysis of fragment profiles. Essentially, these investigations lead us back to an exploration of compound class-specific structural features, given the fact that random fragment profiles could discern

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

10.1021/ci700108q CCC: $37.00 © 2007 American Chemical Society
Published on Web 06/22/2007

structure–activity relationships. The results of our analysis reveal the presence of activity-specific fragmentation patterns and fragment hierarchies and rationalize the ability to distinguish between molecular similarity relationships on the basis of fragment profile comparisons. Novel types of molecular signatures could be identified by mining randomly generated fragment populations. This approach departs from the conventional way of designing structural fragments or descriptors. A key feature of uncovered fragment hierarchies is that individual fragments and their conditional probabilities of occurrence determine unique pathways and become signature patterns of active molecules.

METHODOLOGY

Random Fragmentation through MolBlaster. First we introduce the MolBlaster method[18] that is used here to generate fragment profiles of test molecules as raw data for our analysis. Random fragment populations are generated through iterative deletions of randomly selected rows from the connectivity table of hydrogen-suppressed 2D graph representations. This procedure corresponds to breaking bonds in molecules. From the reduced connectivity tables, SMILES[15] strings of molecular fragments are exported, and their frequency of occurrence is monitored over subsequent iterations. The composition of computed fragment populations generally depends on two parameters. The first one is the number of permitted bond deletions per iteration, which affects average fragment size. The second parameter is the total number of fragment-producing iterations that determines the total number of diverse fragments and their frequency of occurrence. We showed that randomizing the number of bond deletions during each step was a generally preferred fragmentation scheme.[20] A total of 2000 iterations were found to produce characteristic fragment populations for different classes of molecules.[20]

Information Entropy-Based Fragment Profile Analysis. Next we discuss the application of information entropy analysis to generate fragment profiles for activity classes and compare these class profiles to fragment profiles of individual database compounds. For the representation of sets of molecules having similar biological activity, we combine individual populations into a single fragment population. Fragment selection is based on evaluating the frequency distribution of each generated fragment over all molecules. For this purpose, we calculate a scaled form of Shannon entropy,[19] defined as

$$\mathrm{sSE(frag)} = -\sum_{i=1}^{N} p_i(\mathrm{frag})\log_N(p_i(\mathrm{frag}))$$

Here, N is the number of molecules forming the activity set. The probability p_i is calculated as

$$p_i(\mathrm{frag}) = \mathrm{freq}(\mathrm{frag},i)/\sum_{j=1}^{N}\mathrm{freq}(\mathrm{frag},j)$$

and represents the likelihood of a fragment to originate from molecule i. The function freq(frag,i) gives the frequency of a specific fragment frag within the fragment population of molecule i. Scaled Shannon entropy values are high for fragments that occur in all individual fragment populations with comparable frequency, and we select fragments with an sSE value of at least 0.75 to build the fragment profile of an activity set. Applying this criterion, we could achieve significant compound recovery rates for various activity classes in database searching.[20]

Database search calculations involve the comparison of a class profile with fragment profiles of individual database compounds. For this purpose, we have introduced another variant of Shannon entropy called *proportional SE* (PSE):[20]

$$\mathrm{PSE} = \sum_i \frac{a}{b} * \mathrm{sSE_{AC}}(i)$$

with $a = \min\{\mathrm{Freq_{AC}}(i);\mathrm{Freq_{DB}}(i)\}$ and $b = \max\{\mathrm{Freq_{AC}}(i);\mathrm{Freq_{DB}}(i)\}$. $\mathrm{Freq_{AC}}(i)$ is the average frequency of occurrence of fragment i in a given activity class, and $\mathrm{Freq_{DB}}(i)$ is the fragment frequency for a single database compound. $\mathrm{sSE_{AC}}(i)$ reports the scaled SE value for fragment i within the activity class. Weighting of the sSE value by the frequency proportion enables the detection of database molecules that share many fragments with similar frequency with the class profile. Such compounds produce a high PSE score as a measure of molecular similarity.

Algorithm for Mining Fragment Populations. After discussing our approach to generate and compare random fragment profiles, we introduce a novel method for the detailed analysis of fragment populations. We are not only interested in studying the composition of a fragment population at the molecular level of detail but also in aiming to systematically explore whether there are conditional probabilities of fragment occurrence and unique fragment combinations or pathways. Therefore, we attempt to identify and quantify fragment dependencies and organize fragment populations according to dependence relationships.

Let MS = $\{\mathrm{Molecule}_1,\mathrm{Molecule}_2,...,\mathrm{Molecule}_N\}$ be an ordered set of N arbitrarily assembled molecules and FragPop(MS) be the merged fragment set derived from the single fragment populations, as described above. Each fragment frag is associated with a frequency vector, where the ith item is defined as

$$\mathrm{fv(frag)}(i) = \mathrm{freq(frag},i)$$

The generation of frequency vectors is illustrated in Figure 1. These vectors are used to map fragments into an N-dimensional reference system constituted by MS (see Figure 2). Each axis of this coordinate system represents the frequency of occurrence of a fragment in a specific molecule during the fragmentation process.

In principle, the occurrence of a fragment $\mathrm{frag_{dep}}$ can only depend on another fragment $\mathrm{frag_{cond}}$ if $\mathrm{frag_{dep}}$ is generated less frequently in a given subset of molecules, from which $\mathrm{frag_{cond}}$ is derived. This is a necessary but not sufficient condition for a dependence relationship and can be formulated as follows:

$$\forall\ 1 \le i \le N{:}\mathrm{fv(frag_{dep})}(i) \le \mathrm{fv(frag_{cond})}(i)$$

and

$$\exists\ 1 \le i \le N{:}\mathrm{fv(frag_{dep})}(i) < \mathrm{fv(frag_{cond})}(i)$$

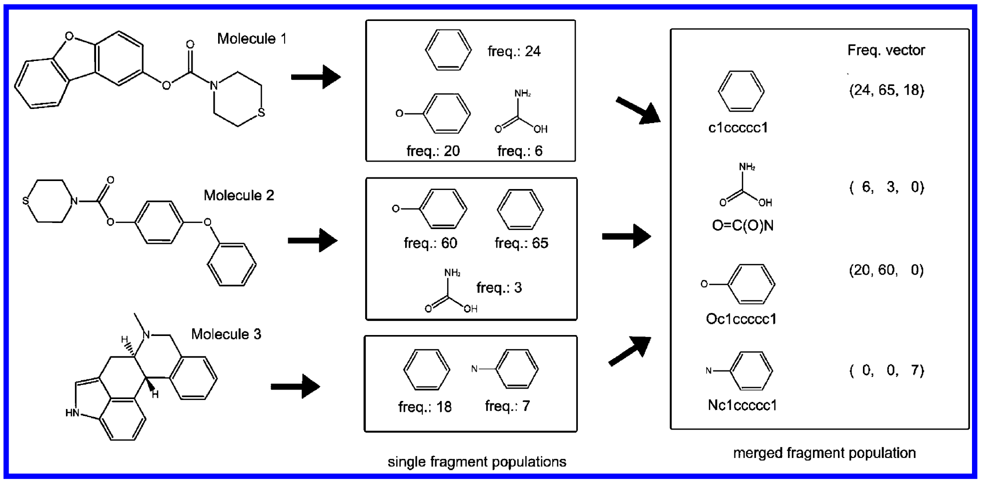

Figure 1. Generation of a fragment profile. An ordered set of molecules is subjected to random fragmentation. After generation of individual fragment populations the frequency of occurrence is determined for each fragment. When individual fragment populations are combined, frequency vectors are generated that record the relative frequency of occurrence in all molecules.

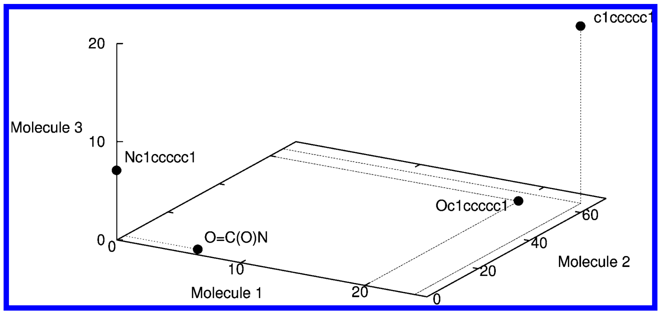

Figure 2. Mapping of fragments into a molecular coordinate system. For the molecule set in Figure 1, a reference system is established where fragments are positioned based on their frequency vectors. Dots indicate fragment positions.

∀ and ∃ are logical quantification symbols and mean "for any" and "there exists", respectively. In order to quantify fragment dependencies, we introduce the function dep:

$$\mathrm{dep}(\mathrm{frag}_{\mathrm{dep}},\mathrm{frag}_{\mathrm{cond}}) = \delta \sum_{i=1}^{N} \mathrm{fv}(\mathrm{frag}_{\mathrm{dep}})(i)/\mathrm{fv}(\mathrm{frag}_{\mathrm{cond}})(i)$$

The delta-operator is defined as

$$\delta = \begin{cases} 1 \text{ if } \forall\ 1 \le i \le N{:}\mathrm{fv}(\mathrm{frag}_{\mathrm{dep}})(i) \le \mathrm{fv}(\mathrm{frag}_{\mathrm{cond}})(i) \\ \wedge\ \exists\ 1 \le i \le N{:}\mathrm{fv}(\mathrm{frag}_{\mathrm{dep}})(i) < \mathrm{fv}(\mathrm{frag}_{\mathrm{cond}})(i) \\ 0 \text{ else} \end{cases}$$

and formalizes the dependency condition stated above.

Applying the dep function, we systematically calculate for all possible fragment pairs in a fragment population FragPop(MS) the dependency values. Nonzero dependency values indicate a conditional probability relationship between two fragments. For a conditional fragment, the fragment with the largest dependency value is most closely related to and dependent on it. Table 1 reports the calculated dependency values for the example fragments shown in Figure 1. We can for all fragments determine a set of conditional fragments CF(frag) yielding the largest dependency values for dependent fragments. This set is defined as

$$\mathrm{CF}(\mathrm{frag}_i) = \{\mathrm{frag}_k \in \mathrm{FragPop} \mid \neg\exists\ \mathrm{frag}_j{:}\mathrm{dep}(\mathrm{frag}_i,\mathrm{frag}_j) > \mathrm{dep}(\mathrm{frag}_i,\mathrm{frag}_k)\}$$

Table 1. Quantification of Fragment Dependency[a]

dependent fragments	conditional fragments			
	c1ccccc1	O=C(O)N	Oc1ccccc1	Nc1ccccc1
c1ccccc1		0	0	0
O=C(O)N	0.30		**0.35**	0
Oc1ccccc1	**1.75**	0		0
Nc1ccccc1	**0.39**	0	0	

[a] For the fragments shown in Figure 1, dependency values are calculated for all possible fragment pair combinations. Maximum values for dependent fragments are written in bold.

In Figure 3, we illustrate the so derived maximal dependency relationships between individual fragments, and Figure 4 shows a treelike graph capturing these relationships. For a fragment population derived from any set of molecules, a corresponding graph structure can be built after calculating dependency values. In order to determine essential fragment pathways leading to individual molecules, we complement the fragment population with SMILES representations of all test molecules. These "superfragments" are assigned a frequency vector where the dimension representing each individual molecule is set to one, whereas all other values are set to zero. This guarantees that intact molecules are always positioned at the end of a fragment hierarchy. Furthermore, it simplifies the tree representation because pathways that do not terminate with an individual molecule can be eliminated because such fragment paths cannot have signature character.

Fragment Profile Analysis. In order to mine random fragment populations and establish fragment dependencies, a number of test sets consisting of different known active and other database compounds were assembled (further described below) and subjected to random fragmentation using MolBlaster. For each set of molecules, a fragment profile was generated, as described above, and fragment dependency values were systematically calculated using the dep function. Using these dependency values, fragment populations were organized in graphs that were annotated with activity information. Each test set consisted of 15 compounds, two times five belonging to two different activity classes of the Molecular Drug Data Report (MDDR)[21] and five randomly taken from ZINC.[22] For comparison of active compounds, Tanimoto similarity[23] (Tanimoto coefficient, Tc)

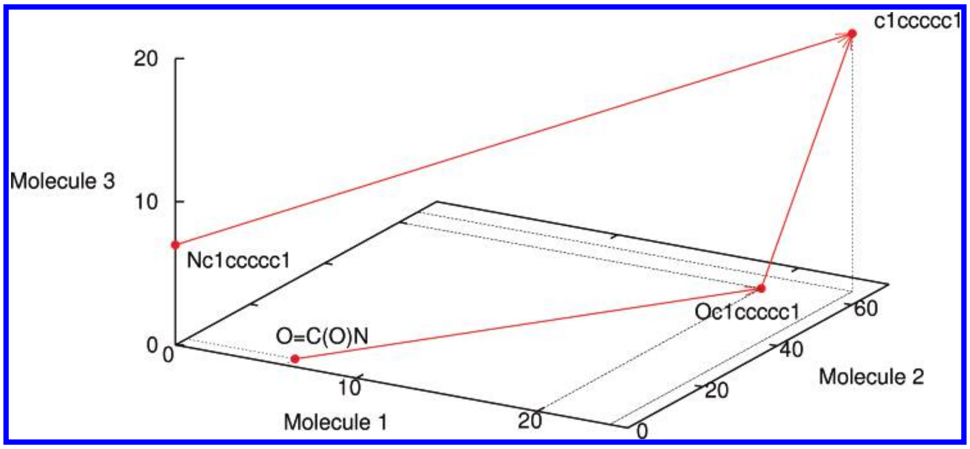

Figure 3. Dependency relationships. Maximal dependency relationships are shown based on the dependency values reported in Table 1. Red dots indicate fragment positions and arrows relationships.

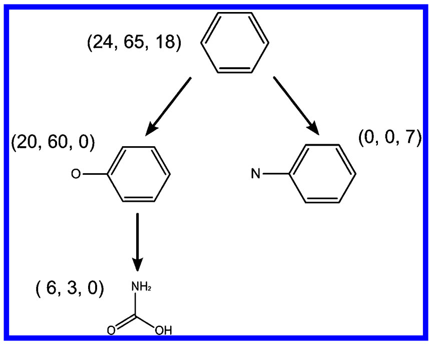

Figure 4. Dependency graph. Shown is a graph representation capturing the fragment dependencies in Figure 3, as described in the text.

was calculated using a fingerprint consisting of 166 MACCS structural keys.[24] Potential activities of the ZINC compounds are unknown, and they were thus considered decoys in our analysis. All test molecules were fragmented over 3000 MolBlaster iterations, and tree graphs were drawn with graphviz.[25]

MolBlaster analysis and fragment profiling can handle much larger molecule sets than the ones studied here.[18,20] The major reason to limit the size of the molecule sets for our current analysis was that graphs for representing fragment dependencies quickly become very large and complex, which hinders intuitive visual analysis. Therefore, we used relatively small sets consisting of three compound classes in order to control the complexity of the representations. We have also carried out calculations on larger molecule sets and more than three classes per set. These calculations indicated that the findings reported herein were in general not much influenced by test set composition and size. In order to further simplify the evaluation of dependency graphs, we have classified the fragments into three distinct classes on the basis of their frequency vectors, two classes containing all fragments uniquely associated with the activity classes, and a third class containing all remaining fragments. Edges in the dependency graphs are color-coded to reflect this classification scheme and permit the inspection of fragment pathways of subsets of molecules having similar activity.

ANALYSIS OF RANDOM FRAGMENT POPULATIONS

Fragment Dependency. The graph shown in Figure 4 is a rudimentary prototype for large tree structures representing frequency-based fragment dependency relationships. What is the meaning of fragment dependencies? The three molecules shown in Figure 1 produce a pathway leading from benzene to the phenol ring and the hypothetical amino acid (methyl carbamate precursor) on the left in Figure 4. Random fragmentation of all three molecules produces benzene. Only a subset of these molecules produces the phenol ring and the hypothetical amino acid (methyl carbamate precursor). Whenever the phenol ring is generated, the fragment profile contains benzene. Furthermore, whenever the hypothetical amino acid is produced, the fragment profile contains the phenol ring. This means that the occurrence of the phenol ring in the fragment profile is dependent on the presence of benzene that is generated more frequently within this subset of molecules. Similarly, the occurrence of the amino acid is dependent on the presence of the phenol ring. Importantly, pathways like the left branch in Figure 4 can only exist for subsets of test molecules. Therefore, the simultaneous presence of the three fragments discussed above becomes a signature of two of the three test molecules shown in Figure 1. Fragment dependencies are not necessarily substructure relationships. For example, there is a frequency-based dependency relationship between benzene and the hypothetical amino acid in Figure 4, but no structural relationship because the smaller fragment is not a substructure of benzene. Thus, the fragment relationships analyzed herein go beyond the assessment of structural resemblance and capture the conditional probability of fragment occurrence.

Test Sets. We have analyzed fragment populations produced by a number of different test sets and consistently obtained results revealing the same general trends. Therefore, we present the results for one exemplary compound set and provide the data for additional test sets in Supporting Information Figures 1−3. The exemplary set discussed herein consists of neurokinin NK2 receptor antagonists (NK2), thromboxane antagonists (THR), and ZINC compounds. The structures of these molecules are shown in Figure 5.

Shape of Fragment Dependency Graphs. The graph representing all fragment dependencies for the fragment profile of the NK2-THR-ZINC set is shown in Figure 6. It is evident that calculated fragment pathways are complex even for relatively small sets of test molecules and that detailed analysis of dependency pathways requires the evaluation of subgraphs. Root fragments at the top of the dependency graph, where pathways begin, are usually small

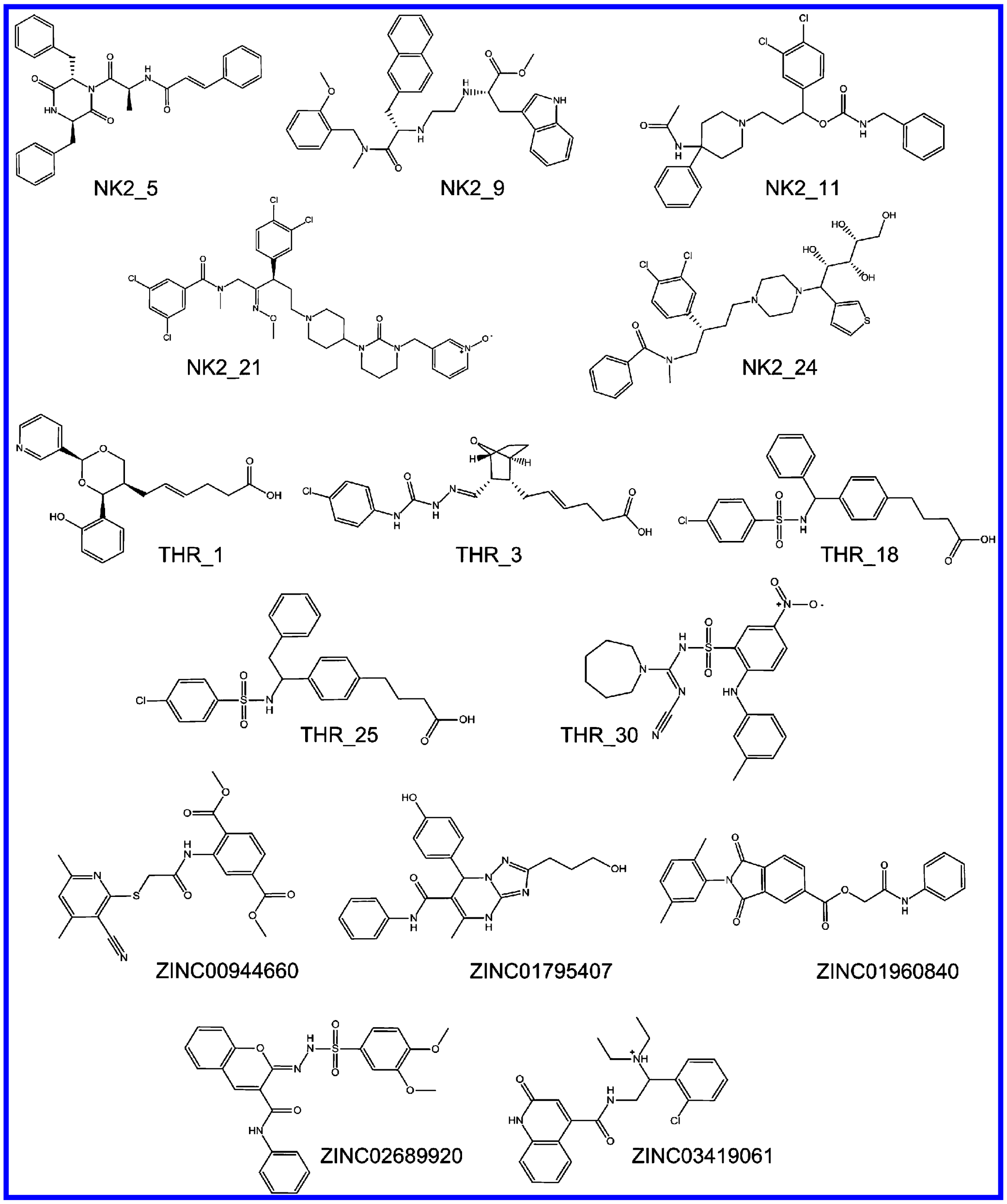

Figure 5. Molecular test set. Structures of neurokinin NK2 receptors antagonists (NK2), thromboxane antagonists (THR) and ZINC molecules are shown, for which random fragment populations and a profile were generated. Database identity codes are provided.

and often single atoms like "C" (as in this case), "N", or "O". Small fragments occur in all individual fragment populations but with varying frequency. At the bottom of the graph, pathways end at individual molecules. Fragment pathways split at many levels and become increasingly specific for subsets of test molecules. As pathways proceed toward target molecules, fragments have the tendency to become larger in size and thus less generic (and more specific).

Annotation with Compound Activities. Random fragment profiles and dependency graphs can be calculated for arbitrary sets of molecules, but the presence of known active compounds makes it possible to annotate the fragment pathways with activity data, as also shown in Figure 6. Through color-coding it becomes evident that specific fractions of all fragment pathways become characteristic for subsets of molecules having similar activity. The NK2-THR-ZINC graph contains a total of 888 fragments. Only seven of 888 fragments are produced by all 15 test molecules, and 182 fragments occur in all molecule subsets that are separated along the dependency graph. By contrast, 169 fragments (19.0%) are uniquely associated with activity class NK2 and 125 (14.1%) with class THR. Thus, about a third of all randomly generated fragments in this test set are activity class-unique. This is a key observation because it demonstrates that random MolBlaster profiles contain compound class-specific fragments and provides an explanation for the ability of profile comparisons to distinguish between different structure−activity relationships[18] and identify active molecules in virtual screening.[20]

Class-Specific Subgraphs. Fragment pathways that are unique to NK2 and THR are shown in Figures 7 and 8, respectively. These subgraphs reveal how different dependency pathways converge on molecules belonging to an activity class. These pathways will be discussed in detail below. A noteworthy feature of these subgraphs is that their topology significantly differs. This is a direct consequence of differences in fragment distribution among active mol-

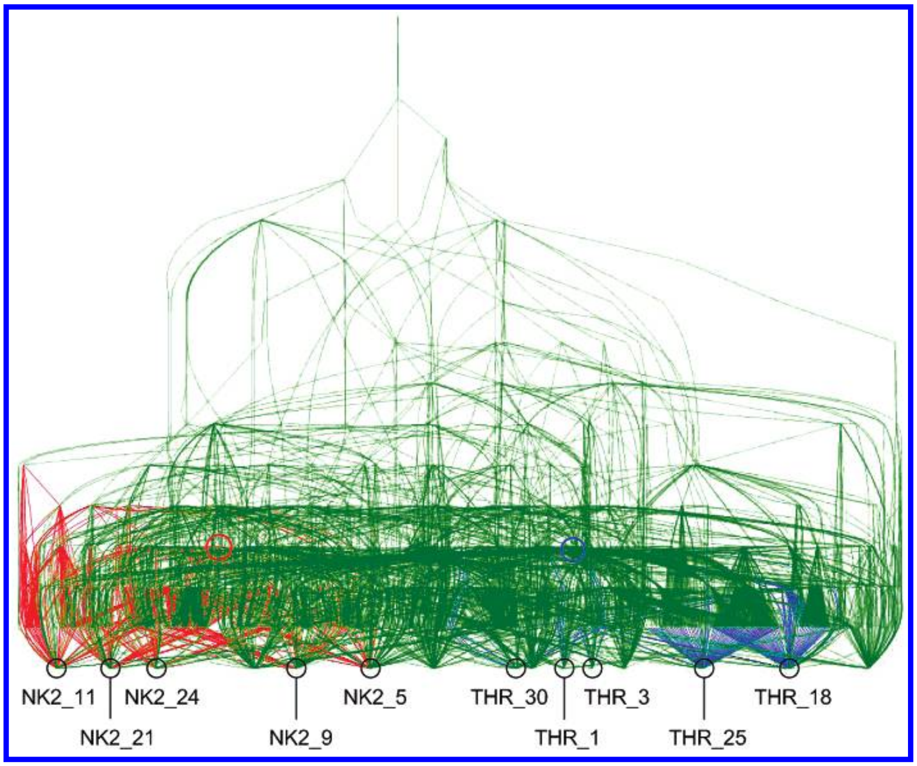

Figure 6. Dependency graph of a fragment profile. The graph for the NK2-THR-ZINC fragment profile is shown. NK2- and THR-specific fragment pathways are colored red and blue, respectively. All other pathways are colored green. In Figures 6−8, the positions of selected corresponding NK2- and THR-specific fragments are indicated by red and blue circles, respectively.

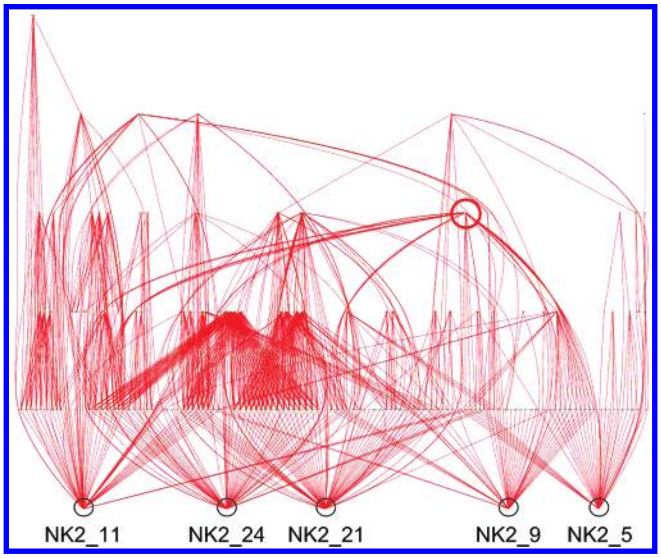

Figure 7. NK2-specific fragment pathways. The NK2-specific subgraph is shown according to Figure 6. Terminal nodes at the bottom represent individual molecules.

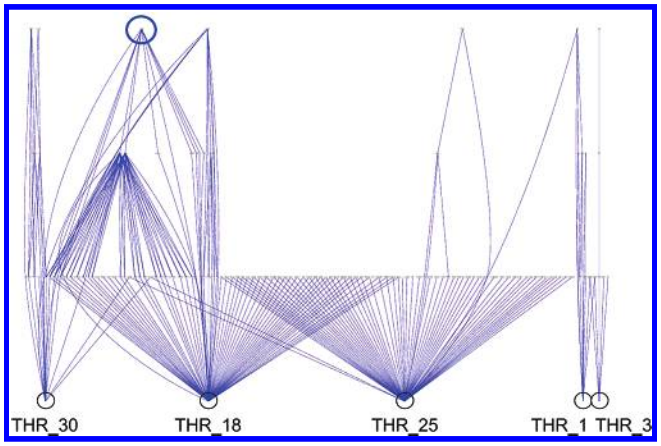

Figure 8. THR-specific fragment pathways. The THR-specifc subgraph is shown according to Figure 6. Terminal nodes at the bottom represent individual molecules. The NK2- and THR-specific subgraphs have distinct differences in topology, as discussed in the text.

ecules. For example, the NK2 subgraph in Figure 7 is compact and densely populated by different pathways. This is the case because a number of conditional fragments are generated by all active compounds such as, for example, fragment N(C)CCCC═C. The dependent fragment N(C)-(CC)CCC═C then leads to molecules NK2_21 and NK2_24 and dependent fragment N(CC)(CCC)CCC to NK2_21 and NK2_5. Thus, NK2_21 represents a transitional state between NK2_5 and NK2_24. In contrast to the NK2 subgraph, the THR subgraph in Figure 8 is sparsely populated and does not contain many "hub" fragments that are shared among active molecules. For example, molecules THR_1 and THR_3 are only connected with other THR compounds through the single conditional fragment C/C═C/CCCC. Thus, the topology of "active subgraphs" reflects intraclass structural homogeneity and molecular size. The average MACCS Tc value for pairwise comparison of the NK2 molecules is 0.55, and the corresponding average value for THR molecules is 0.43. NK2 molecules consist of on average 42.0 non-hydrogen atoms, whereas THR compounds consist of 29.6. Therefore, structurally more similar and larger molecules such as NK2 produce more shared signature fragments and more densely connected subgraphs than structurally more diverse and smaller compounds such as THR (see also Figure 5).

Unique Fragments and Compound-Specific Pathways. The activity class-specific subgraphs in Figures 7 and 8 reveal a key feature of fragment hierarchies. The subset of

Table 2. NK2 Fragment Pathway[a]

fragment	frequency vector
C	(21265, 26335, 22592, 21147, 28088, 16443, 15883, 16876, 15845, 13144, 14829, 15427, 18451, 18408, 16585)
CC	(3602, 3825, 3108, 3636, 3573, 2918, 2627, 3248, 2667, 2869, 1817, 2565, 2716, 2249, 3103)
C=C	(1396, 878, 936, 1834, 1724, 1314, 865, 1405, 790, 582, 666, 984, 1282, 1312, 992)
CCC	(336, 584, 351, 343, 433, 349, 479, 460, 551, 587, 240, 226, 334, 71, 118)
C/C=C/C	(34, 44, 23, 85, 55, 39, 72, 66, 50, 64, 64, 26, 65, 25, 22)
C=C(C)CC	(9, 5, 22, 10, 23, 22, 1, 23, 0, 0, 0, 4, 0, 0, 6)
NCCCC=C	(4, 2, 7, 5, 4, 0, 0, 4, 0, 0, 0, 0, 0, 0, 5)
N(C)CCCC=C	(4, 1, 3, 1, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
NCCNCCC	(0, 0, 3, 1, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
O=C[C@@H](NC)CC	(0, 0, 0, 1, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

[a] Reported is a summary of a selected fragment pathway for molecule NK2_5. SMILES strings of participating fragments are associated with their corresponding frequency vectors. The first five frequency vector coordinates represent molecules from activity class NK2, the next five coordinates represent molecules from activity class THR, and the last five coordinates represent ZINC compounds.

Table 3. THR Fragment Pathway[a]

fragment	frequency vector
C	(21265, 26335, 22592, 21147, 8088, 16443, 15883, 16876, 15845, 13144, 14829, 15427, 18451, 18408, 16585)
CC	(3602, 3825, 3108, 3636, 3573, 2918, 2627, 3248, 2667, 2869, 1817, 2565, 2716, 2249, 3103)
C=C	(1396, 878, 936, 1834, 1724, 1314, 865, 1405, 790, 582, 666, 984, 1282, 1312, 992)
CCC	(336, 584, 351, 343, 433, 349, 479, 460, 551, 587, 240, 226, 334, 71, 118)
C/C=C/C	(34, 44, 23, 85, 55, 39, 72, 66, 50, 64, 64, 26, 65, 25, 22)
C=C/C=C/C	(15, 13, 12, 73, 35, 15, 7, 27, 28, 0, 10, 3, 8, 13, 14)
[S+1]([O-])([O-])C=C	(0, 0, 0, 0, 0, 9, 0, 26, 9, 0, 0, 0, 0, 13, 0)
[S+]([O-])(N)C=C	(0, 0, 0, 0, 0, 9, 0, 8, 6, 0, 0, 0, 0, 10, 0)
[S+1]([O-])([O-])C(=)C	(0, 0, 0, 0, 0, 3, 0, 6, 1, 0, 0, 0, 0, 9, 0)
[S+]([O-])(N)/C(/C=C)=C/C	(0, 0, 0, 0, 0, 2, 0, 4, 1, 0, 0, 0, 0, 0, 0)
[S+1]([O-])([O-])(NC)c1ccccc1	(0, 0, 0, 0, 0, 2, 0, 2, 1, 0, 0, 0, 0, 0, 0)
S(NC)C(=)C	(0, 0, 0, 0, 0, 0, 0, 2, 1, 0, 0, 0, 0, 0, 0)

[a] Summary of a selected fragment pathway leading to molecule THR_30. The representation is according to Table 2.

fragments that is unique to an activity class is organized in pathways that are specific for individual molecules. Table 2 summarizes a pathway that is specific for an NK2 molecule, and Table 3 summarizes another one that is specific for a THR compound. These pathways begin at the root level and extend over 10 and 12 fragments, respectively, which either co-occur in different compound classes or are class-unique. The frequency vectors show that about half of these fragments are very narrowly distributed. The bottom three fragments in Tables 2 and 3 are NK2-unique and THR-unique, respectively. Clearly, the presence of class-unique fragments in fragment profiles helps to rationalize why these profiles could be successfully used for the identification of active compounds. This requires that unique fragments found in individual fragment profiles are also selected for the class profile, which is accomplished by the sSE metric that selects fragments generated for all active compounds with comparable frequency. As shown in Tables 2 and 3, these fragments include class-unique fragments but also generic fragments that occur not only in active molecules. When PSE scoring is applied, these widely distributed fragments can favor the selection of false-positives. Dependency graphs make it possible to omit generic fragments from class profiles. This is expected to further increase the accuracy of database search calculations. As shown in Tables 2 and 3, class-unique fragments occur in more than one active molecule, but their conditional probabilities vary and their pathways are molecule-specific. Figures 9 and 10 illustrate for subsets of class-unique fragments how different dependency pathways lead to an individual NK2 and THR molecule, respectively. Unique fragments participate in multiple specific pathways for individual molecules. These pathways involve different sets of class-unique fragments having different conditional probabilities and relative importance in forming fragment pathways. For example, some fragments are conditional for the occurrence of many other fragments and thus have a “hub” function for the formation of alternative dependency pathways.

Molecular Signatures. The results discussed above demonstrate that random molecular fragment populations contain structural elements and hierarchies that have the character of molecular signatures. We can distinguish between two signature levels. First, combinations of unique fragments represent a compound class signature. Second, specific fragment pathways represent signatures of individual molecules within a class. Since multiple pathways exist for each molecule, they could also be combined to produce signature patterns.

In addition to these specific signatures, fragment profiles can also reveal pharmacophore information. For example, fragment profiles of carbonic anhydrase inhibitors (see the Supporting Information) contain the sulfonamide group, which is crucially important for inhibitory activity, with high frequency of occurrence. Furthermore, we can also derive common subgraphs from fragment profiles. By definition, MolBlaster fragments are 2D molecular subgraphs. Hub fragments are produced by all molecules within a class and thus represent common subgraphs. The combination of hub fragments can also produce the maximum common subgraph for a set of compounds.

Concluding Remarks. We have introduced a methodology to mine and organize randomly generated molecular fragment populations. On the basis of conditional probabilities of fragment occurrence, fragment hierarchies could be established. Our analysis has identified fragments in random populations that are uniquely produced by compounds having similar activity. Although the composition of unique fragment subsets is expected to be influenced by the nature of compound classes and molecular data sets, our analysis has demonstrated that random fragment profiles

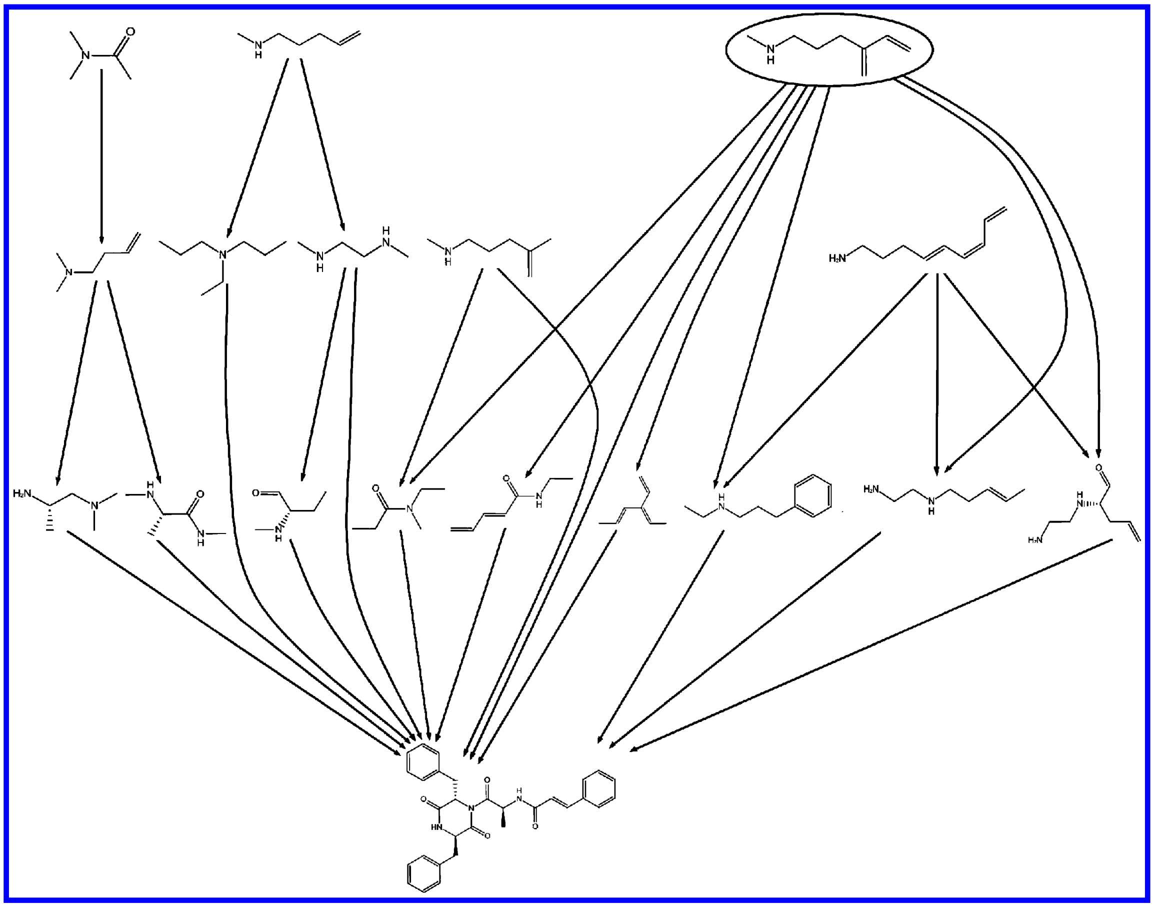

Figure 9. NK2 compound-specific fragment pathways. Shown is a subset of specific fragments and their dependency pathways leading to molecule NK2_5. The fragment in the circle is the one whose position is indicated by the red circles in Figures 6 and 7.

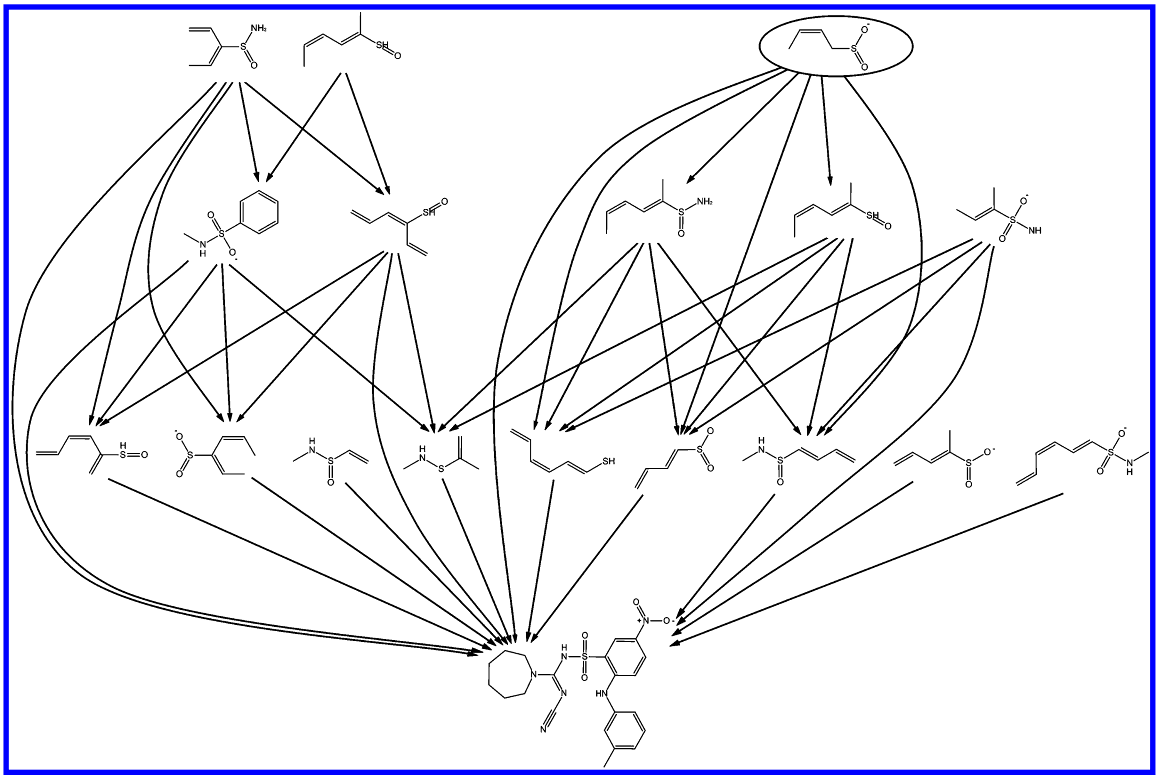

Figure 10. THR compound-specific fragment pathways. Shown is a subset of specific fragments and their dependency pathways leading to molecule THR_30. The fragment in the circle is the one whose position is indicated by the blue circles in Figures 6 and 8.

contain class-unique elements. These findings explain why the comparison of random fragments profiles can distinguish between different structure−activity relationships and serve as a measure of molecular similarity in compound database searching. Moreover, the organization of random fragment populations into dependency graphs has revealed the presence of "active subgraphs" of different topology and many molecule-specific fragment pathways. These random fragment pathways provide previously unconsidered molecular information specifically associated with active compounds. The results of our analysis suggest that it might be possible to systematically extract class-unique fragments and specific pathways from random fragment profiles of active molecules, which would provide an opportunity for the design of novel types of structural descriptors.

Supporting Information Available: Fragment profiles and pathways for three additional test sets each consisting of two compound activity classes and database compounds (Figures 1−3). This material 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) 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.

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

(6) Xue, L.; Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Comb. Chem. High Throughput Screening* **2000**, *3*, 363−372.

(7) 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.

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

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

(10) Bondensgaard, K.; Ankersen, M.; Thogersen, H.; Hansen, B. S.; Wulff, B. S.; Bywater, R. P. Recognition of privileged structures by G-protein coupled receptors. *J. Med. Chem.* **2004**, *47*, 888−899.

(11) Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. Are target-family-privileged substructures truly privileged? *J. Med. Chem.* **2006**, *49*, 2000−2009.

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

(13) 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.

(14) 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.

(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) Vidal, D.; Thormann, M.; Pons, M. LINGO, an efficient holographic text based method to calculate properties and intermolecular similarities. *J. Chem. Inf. Model.* **2005**, *45*, 386−393.

(17) Karwath, A.; De Raedt, L. SMIREP: predicting chemical activity from SMILES. *J. Chem. Inf. Model.* **2006**, *46*, 2432−2444.

(18) 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.

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

(20) 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.

(21) *Molecular Drug Data Report (MDDR);* MDL-Elsevier: San Leandro, CA, U.S.A., 2006.

(22) 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.

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

(24) *MACCS structural keys*; MDL-Elsevier: San Leandro, CA, U.S.A., 2002.

(25) Gansner, E. R.; North, S. An open graph visualization system and its applications to software engineering. *Software-Pract. Experience* **2000**, *30*, 1203−1233.

CI700108Q