Scaffold Hopping Using Clique Detection Applied to Reduced Graphs

Edward J. Barker,[†] David Buttar,[‡] David A. Cosgrove,[‡] Eleanor J. Gardiner,*[,†] Paula Kitts,[‡] Peter Willett,[†] and Valerie J. Gillet[†]

Department of Information Studies and Krebs Institute for Biomolecular Research, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom, and AstraZeneca, Mereside, Alderley Park, Macclesfield, Cheshire, United Kingdom

Received August 26, 2005

Similarity-based methods for virtual screening are widely used. However, conventional searching using 2D chemical fingerprints or 2D graphs may retrieve only compounds which are structurally very similar to the original target molecule. Of particular current interest then is *scaffold hopping*, that is, the ability to identify molecules that belong to different chemical series but which could form the same interactions with a receptor. Reduced graphs provide summary representations of chemical structures and, therefore, offer the potential to retrieve compounds that are similar in terms of their gross features rather than at the atom−bond level. Using only a fingerprint representation of such graphs, we have previously shown that actives retrieved were more diverse than those found using Daylight fingerprints. Maximum common substructures give an intuitively reasonable view of the similarity between two molecules. However, their calculation using graph-matching techniques is too time-consuming for use in practical similarity searching in larger data sets. In this work, we exploit the low cardinality of the reduced graph in graph-based similarity searching. We reinterpret the reduced graph as a fully connected graph using the bond-distance information of the original graph. We describe searches, using both the maximum common induced subgraph and maximum common edge subgraph formulations, on the fully connected reduced graphs and compare the results with those obtained using both conventional chemical and reduced graph fingerprints. We show that graph matching using fully connected reduced graphs is an effective retrieval method and that the actives retrieved are likely to be topologically different from those retrieved using conventional 2D methods.

INTRODUCTION

Similarity searching is a widely applied technique, especially in the early stages of drug discovery when little is known about the biological target. The rationale for similarity searching is based on the similar property principle, which states that structurally similar compounds tend to have similar properties.[1] Thus, a compound which is structurally similar to a known active compound has an increased likelihood of sharing the same activity, relative to a compound selected at random. Although counter examples exist, where small changes in structure lead to large changes in activity, evidence for the structure property principle is provided by Martin et al.,[2] and also by the analogue series that exist in many in-house databases within pharmaceutical companies and in external databases such as the MDL Drug Data Report (MDDR).

Similarity methods require numerical descriptors to represent the compounds and a similarity coefficient to quantify the degree of similarity based on the descriptors.[3] The first similarity methods to be reported were based on the 2D fragment bitstrings originally developed for substructure search. These descriptors, in combination with the Tanimoto coefficient, continue to be popular and have been shown to be effective in enrichment studies. However, while 2D fragment descriptors are well-suited to identifying close analogues, they are less effective at identifying compounds that represent different lead series yet share biological activity, a technique that has become known as scaffold hopping.[4,5] Such hits can be more valuable than close analogues for a number of reasons: they allow a drug discovery program to move out of the patent space of the initial query compound, they provide alternative lead series should one series fail because of poor absorption, distribution, metabolism, and excretion properties, and they provide an alternative route when a lead compound has difficult chemistry.

Recently, a great deal of effort has been applied to developing new ways of describing molecules for similarity searching.[6] Several methods are based on three-dimensional properties of compounds, ranging from alignment-free methods such as pharmacophoric fingerprints through to shape- and field-based methods that typically require the compounds to be aligned prior to measuring their similarity, for example, FBSS[7] and ROCS.[8] The aim of these approaches is to capture features of molecules that are important for activity without considering the full atom−bond framework of the molecules. While not always better than 2D descriptors in terms of enrichment studies, these methods have been shown to be more effective at identifying diverse hits.[9,10] However, a disadvantage of 3D similarity methods is that the conformational flexibility of the compounds should be taken into account. This is a nontrivial problem since the

* To whom all correspondence should be addressed. E-mail: e.gardiner@sheffield.ac.uk.
[†] University of Sheffield.
[‡] AstraZenca.

10.1021/ci050347r CCC: $33.50
Published on Web 01/10/2006

bioactive conformations of the molecules are not usually known and there can be many conformers that are accessible for a given molecule. Furthermore, structure generation programs vary in their ability to identify bioactive conformations.[11]

Thus, there has also been considerable interest in 2D descriptors that aim to capture pharmacophoric features of molecules rather than all the atoms and bonds. These descriptors are sometimes referred to as topological pharmacophores to emphasize that the relationships between the features are encoded in 2D rather than 3D, often using through-bond distances. Thus, they avoid the need to consider conformational flexibility. Examples include the modified atom pairs introduced by Kearsley et al., where atoms are represented by their binding properties such as cation, anion, donor, acceptor, and so forth rather than specific element types.[12] More recent descriptors include the CATS descriptors[5] and Similog keys.[13]

We have previously described the use of reduced graphs for similarity searching. Reduced graphs can also be considered as topological pharmacophore descriptors since they provide summary representations of molecules. Groups of connected atoms are reduced to single nodes with the connectivity between the nodes reflecting the connectivity of atoms in the original chemical graph. Many different criteria can be used to generate reduced graphs;[14] however, when applied to similarity searching, the graph reduction process focuses on features of molecules that are likely to be important for interaction with a receptor, such as ring systems, charged groups, and groups having hydrogen bonding ability. In previous work, we have explored a variety of different ways both of generating reduced graphs and also of mapping them to fingerprints for similarity searching. Harper et al. have developed an alternative approach to measuring similarity between reduced graphs on the basis of a string edit distance methodology.[15] Here, we focus on the use of graph-matching techniques for quantifying the similarity of reduced graphs. Takahashi et al. have previously described a related approach to comparing reduced graphs.[16] The method was applied to a set of five structurally diverse antihistamines and to a set of six antipsychotropic agents, and in both cases, some of the structural similarities were found. Although our work is similar in concept to that of Takahashi, to our knowledge, their method has not been applied to the virtual screening of large data sets, as is described here. There are also some differences in the definitions of the reduced graphs used. For example, in Takahashi's method, the nodes are defined using a fragment dictionary, which allows for overlapping fragments so that atoms can belong to more than one node. Furthermore, the edges in their reduced graph are weighted by the distances between the nodes so that, when structures contain rings, multiple paths between functional groups lead to multiple edge weights.

In this paper, we apply different graph-matching techniques to the comparison of reduced graphs in a virtual screening context. We compare our methods with conventional 2D fingerprints using enrichment studies, the ability to scaffold hop, and computational cost.

Table 1. Reduced Graph Node Types

node description	node code
acyclic inert node	L
acyclic feature node, acceptor only	A
acyclic feature node, donor only	D
aromatic ring, no hydrogen bonding	Ar
aromatic ring, hydrogen bond acceptor	ArA
aromatic ring, hydrogen bond donor	ArD
alicyclic ring, hydrogen bond donor	RD
alicyclic ring, hydrogen bond acceptor	RA
alicyclic ring, no hydrogen bonding	R
acid feature	Ac
base feature	B
acyclic feature node, both donor and acceptor	D/A
aromatic ring, both donor and acceptor	ArD/A
alicyclic ring, both donor and acceptor	RD/A

METHODS

Graph Reduction. In previous work, we have described the generation of different types of reduced graphs using different levels and degrees of reduction.[17,18] Here, we use a particular graph reduction, similar to the ring/feature/link reduction/level 4 described previously[17] with the additional prior identification of acid and base nodes. Other modifications include the treatment of a hydrogen bond donor/acceptor atom as both a hydrogen bond donor and a hydrogen bond acceptor and the explicit inclusion of nonreactive terminal groups. This level of graph reduction was determined to be the best-performing overall in a highly detailed series of experiments.[19] Thus, 14 node types are defined, listed in Table 1. The node definitions used in generating the reduced graphs are user-definable and are determined from SMARTS definitions read from an input file. This gives the method a great deal of flexibility. In the results presented here, the definitions of acids, bases, and so forth are those provided by AstraZeneca. Rings are identified using Daylight ring definitions.[20]

The graph reduction process first identifies all acidic and basic groups. The molecule is then partitioned into cyclic and acyclic fragments. Ring nodes may be aromatic or alicyclic. Within these definitions, a ring may be a hydrogen bond donor, a hydrogen bond acceptor, both a hydrogen bond donor and acceptor, or have no donor/acceptor characteristics. An acyclic feature node may be a hydrogen bond donor, a hydrogen bond acceptor, both a hydrogen bond donor and acceptor, or have no donor/acceptor characteristics. An acyclic node which has no donor/acceptor characteristics is termed a *linker* node. Thus, in this terminology, which has historical origins, terminal groups are described as linker nodes. The generation of the reduced graphs is described in detail by Barker.[19] Some examples are given in Figure 1.

Previously, we developed several fingerprint representations for reduced graphs. In this work, we use the node−bond pair fingerprint.[17] A node−bond pair is defined as *node type*−*distance*−*node type*, where distance is the number of bonds on the shortest path between the two closest atoms, one from each node, in the original chemical graph. The fingerprint contains bits for all possible node pairs and all distances up to a maximum path length of 14. We do not include joint *donor and acceptor* node types in the fingerprint. Instead, we allow joint donor and acceptor nodes to set bits for both *donor* nodes and *acceptor* nodes. Thus, a donor node sets a subset of the bits set by a joint donor and

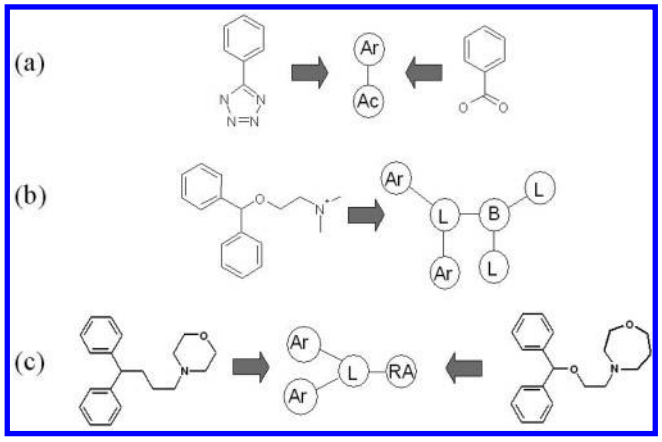

Figure 1. Example of reduced graphs. (a) Acids take precedence over rings; (b) terminal nodes are described as linkers; (c) linker nodes may include heteroatoms if they have no hydrogen-bonding character.

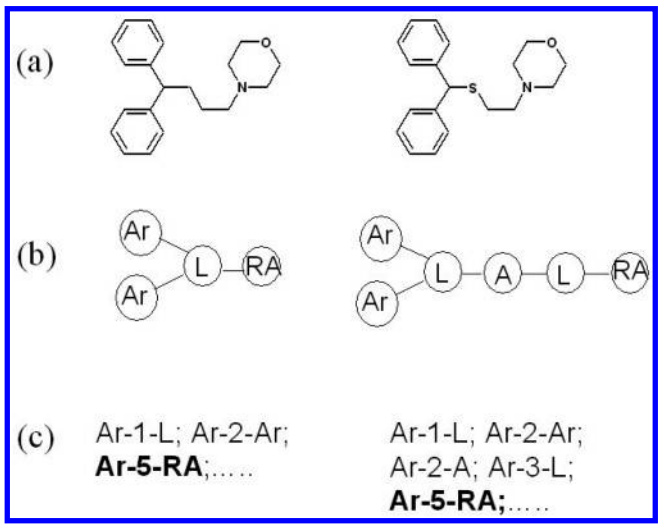

Figure 2. Reduced graph fingerprints maintain a relationship between distant features. (a) Molecules; (b) reduced graphs; (c) examples of the fingerprint bits set. Despite different numbers of intervening nodes, both reduced graphs set a bit for an aromatic ring five bonds from an alicyclic ring.

acceptor node of the same type. For example, all the bits set by an aromatic ring donor node (type ArD) are set by an aromatic ring joint donor and acceptor node (type ArD/A), which also sets additional bits for its acceptor characteristics. The node−bond pair fingerprint enables the relationship between distant features to be maintained when the corresponding reduced graphs differ in the number of intervening nodes. For example, both the reduced graphs in Figure 2 set a bit for an aromatic ring five bonds from an alicyclic acceptor ring. Although the reduced graphs possess different numbers of intervening nodes, both original graphs contain five bonds between the aromatic and alicyclic rings.

Similarity between a pair of reduced graphs, A and B, is measured using the Tanimoto coefficient.

$$\text{Similarity}_{AB} = \frac{c}{a + b - c} \qquad (1)$$

where a is the number of bits set by reduced graph A, b the number of bits set by reduced graph B, and c the number of common bits.

Maximum Common Subgraphs. The fingerprint representation of reduced graphs includes some distance information but loses information about the topology of the original molecules. Since we have a graphical molecular representation, it is natural to look for graph-based measures to determine the similarity between a pair of reduced graphs. One such is the maximum common subgraph (MCS), the largest (by some criterion) subgraph which two graphs have in common. Here, we consider two different versions of the MCS. A *vertex-induced subgraph* is a set of vertices and all the edges which connect them. An *edge-induced subgraph* is a set of edges and all the vertices which are at their end points. An MCIS between a pair of graphs is the largest vertex-induced common subgraph. An MCES between a pair of graphs is the largest edge-induced common subgraph. The presence of an extra edge in one graph but not the other can mean that a node with some common edges can be excluded from a vertex-induced common subgraph. For example, in Figure 3a, the subgraph induced by the vertex set {1,2,3,4,5} is the entire graph, G_1, and the subgraph induced by the vertex set {1′,2′,3′,4′,5′} is the entire graph, G_2. Plainly the two graphs are nonisomorphic because of the edge between 4 and 5 in G_1 and the edge between 3′ and 4′ in G_2. The MCIS is the subgraph induced by the vertex set {1,2,3,5} (or equivalently, {1′,2′,3′,5′}), as shown in Figure 3b. In contrast, node 4 (or 4′) is present in the MCES (Figure 3c).

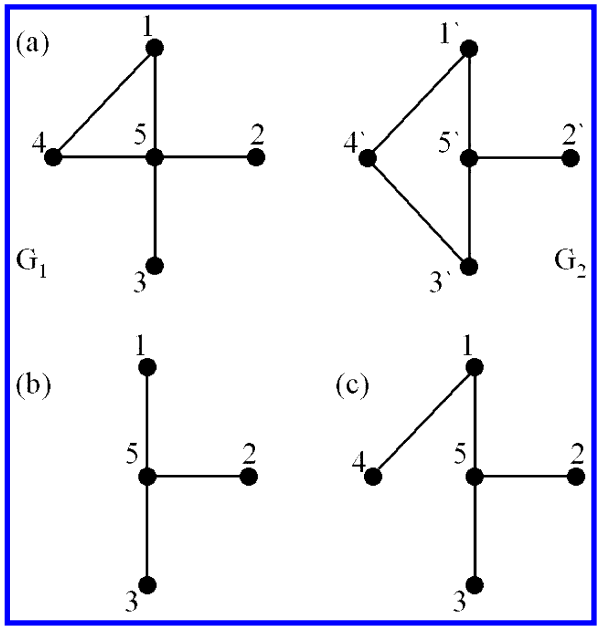

Figure 3. Example MCIS and MCES. (a) Two graphs; (b) MCIS; (c) MCES.

MCS calculations are expensive and may scale exponentially in proportion to the size of the graphs being compared. Recent work has shown that the RASCAL algorithm, when used in an MCES graph matching, is a highly efficient MCS algorithm for the comparison of ordinary chemical graphs.[21,22] Previously, the algorithm of choice has been the Bron−Kerbosch[23] algorithm, which is generally used to find the MCIS. Both RASCAL and Bron−Kerbosch are used to find the maximum cliques (i.e., maximum fully connected subgraphs) of a composite graph. These cliques are mapped back to the original graphs to give the MCS. It is the method of formation of the composite graph which determines whether an MCES or MCIS is found.

It has been suggested that the MCES is a more meaningful measure of the similarity between a pair of molecules than

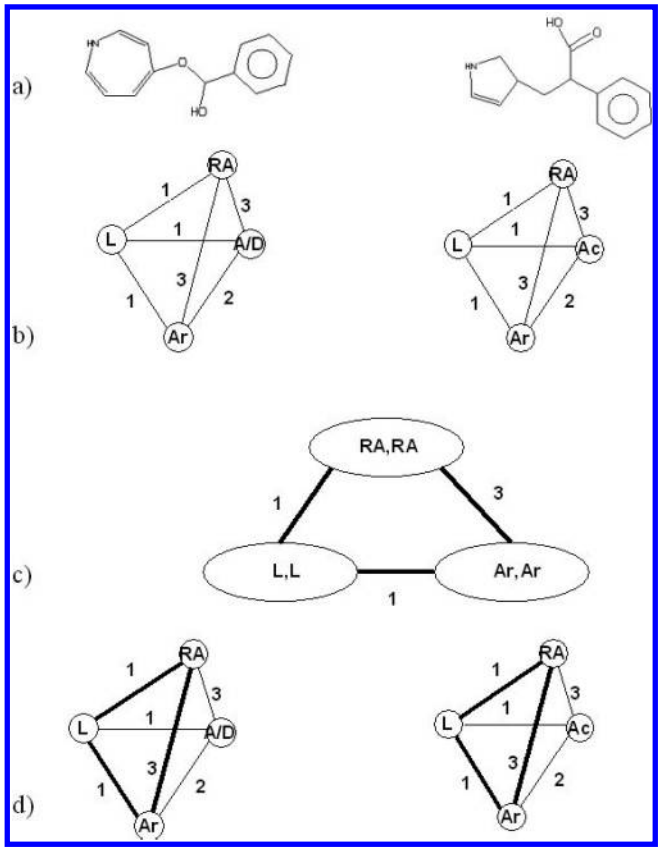

Figure 4. Finding an MCIS. (a) Chemical graphs; (b) FC reduced graphs; (c) correspondence graph; (d) MCIS in bold. The similarity between the reduced graphs is $3/(4 + 4 - 3) = 0.6$.

the MCIS, since the edges represent chemical bonds. These bonds provide constraints on the possible conformations of molecules and, hence, encode some element of three-dimensional molecular recognition. However, it is not clear that the MCES will be a better measure for the similarity between a pair of reduced graphs, since the edges of a reduced graph do not have any such meaning.

Finding the MCIS between a Pair of Fully Connected Reduced Graphs. The process of finding the MCIS is illustrated in Figure 4. Given a pair of molecules (Figure 4a), we first find their reduced graphs. We label the edges of the reduced graphs with the shortest bond distance between atoms in the chemical graphs, giving fully connected reduced graphs (Figure 4b). We focus on finding the MCIS between a pair of fully connected reduced graphs since this ameliorates the problem of intervening nodes causing partial graph matches to be missed. For convenience, we will refer to such graphs as *FC* reduced graphs, while the nonfully connected versions will be called *TC* (*topologically connected*) reduced graphs. Next, we form their *correspondence graph*. The nodes of the correspondence graph are ordered pairs of nodes, one from each reduced graph, whose node labels are of the same type. Two vertices of the correspondence graph are connected if the distance between the first elements of the ordered pairs in their original graph is the same as the distance between the second elements of the ordered pairs in their original graph. The edges of the correspondence graph are labeled with this distance (Figure 4c). The maximum cliques of the correspondence graph are mapped back to the original graphs to give the MCIS (Figure 4d).

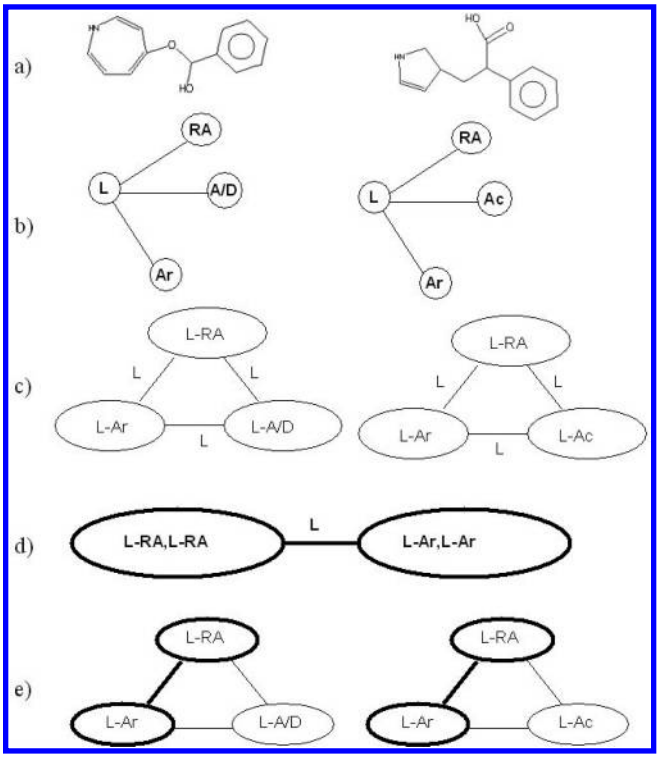

Figure 5. Finding an MCES. (a) Chemical graphs; (b) TC reduced graphs; (c) line graphs; (d) modular product graph; (e) MCES in bold. The similarity between the reduced graphs is $(3 + 2)^2/(4 + 3)(4 + 3) = 0.51$.

The MCIS similarity between a pair of graphs, G_A and G_B, is measured by a graph Tanimoto coefficient

$$\text{Similarity}_{AB} = \frac{|G_{AB}|}{|G_A| + |G_B| - |G_{AB}|} \qquad (2)$$

where $|G_A|$ is the number of vertices in G_A, $|G_B|$ is the number of vertices in G_B, and $|G_{AB}|$ is the number of vertices in the common subgraph.

Finding the MCES between a Pair of Reduced Graphs. The process of finding the MCES is illustrated in Figure 5. For simplicity of description, we focus on the topologically connected case. Given a pair of molecules (Figure 5a), we first find their TC reduced graphs (Figure 5b). We then form the line graphs of the reduced graphs. A line graph is a graph whose vertices are the edges of the original (reduced) graph, labeled with their original endpoints. Line graph vertices are joined if the edges they represent have a common vertex in the original graph (Figure 5c); the line graph edge is labeled with the common vertex label. Then, we form a correspondence graph as before, from the pair of line graphs (Figure 5d). The correspondence graph formed in this way is frequently called a *modular product* graph[21,22] in the literature. The maximum cliques of the modular product graph map back to the common subgraphs of the line graphs (Figure 5e), which in turn map back to the MCES of the reduced graphs. This final mapping is only true if a particular graph interchange, termed a ΔY *interchange*, has not occurred. In a ΔY interchange, triangles of the two line graphs are isomorphic, whereas the original graphs are not, as illustrated in Figure 6. Possible ΔY interchanges are taken into account by the RASCAL program.[21,22]

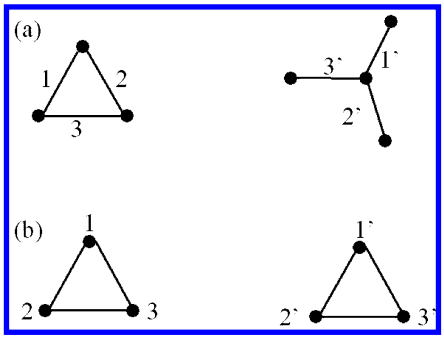

Figure 6. ΔY interchange. (a) Two nonisomorphic graphs; (b) their line graphs are isomorphic.

The similarity calculation for the MCES takes into account the number of edges as well as the number of vertices. This score[22] is given by

$$\text{Similarity}_{AB} = \frac{[|V(G_{AB})| + |E(G_{AB})|]^2}{[(|V(G_A)| + |E(G_A)|)(|V(G_B)| + |E(G_B)|)]} \quad (3)$$

where $|V(G_A)|$ is the number of vertices in G_A and $|E(G_A)|$ is the number of edges in G_A, $|V(G_B)|$ is the number of vertices in G_B and $|E(G_B)|$ is the number of edges in G_B, and $|V(G_{AB})|$ is the number of vertices and $|E(G_{AB})|$ is the number of edges in the common subgraph. (N.B. all these variables refer to the reduced graphs, not to the line graphs.)

We have also implemented the RASCAL algorithm to find the MCES when the reduced graphs are fully connected; again, similarity is measured using eq 3.

Choice of Similarity Scores. In preliminary experiments, we calculated enrichment factors for test sets of 10 targets for each of 11 activity classes (described in detail below) using both eqs 2 and 3 for both the topologically connected and fully connected MCES formulations, which we term *MCES−TC* and *MCES−FC*, respectively, and also for the fully connected MCIS formulation, termed *MCIS−FC*. At the 1% recall level, on average, the enrichment using eq 3 was 22% higher than that obtained using eq 2 for MCES−TC and 87% higher for MCES−FC. In contrast, the average enrichment for the MCIS−FC was 5% higher using eq 2 than it was using eq 3. In Raymond and Willett's study, a number of different similarity measures were investigated.[24] They found a weighted version of eq 3 for the MCES problem to be more effective, the motivation being that, when fragmented, a MCES should be penalized. It is not clear that this is the case for reduced graphs. Thus, we used the original Raymond score, eq 3, for both MCES−TC and MCES−FC and the graph Tanimoto, eq 2, for MCIS−FC.

Simulated Virtual Screening Experiments. We simulated virtual screening using the MDL Drug Data Report (http://www.mdli.com). The database was first filtered using a set of filters from AstraZeneca to remove compounds classified as "ugly". As well as the usual Lipinski-type filters, based on molecular properties such as molecular weight, log *P*, number of rotatable bonds, and so forth, SMARTS rules obtained from a survey of medicinal chemists were used to remove compounds considered undesirable.[25] The remaining database comprised 57 189 "inactive" compounds plus the compounds in 11 activity classes. Details of these classes, which have been used extensively in recent screening experiments,[26,27] are given in Table 2. The renin inhibitors are the most homogeneous, as measured by Hert et al.[26] using the mean of the pairwise similarities obtained using both the Tanimoto coefficient and Unity fingerprints,[28] and the cyclooxygenase inhibitors the least homogeneous of the activity classes.

Table 2. MDDR Classes

code	activity class	number	unique AF	mean sim
5H3	5 HT3 antagonist	706	418	0.35
5HA	5 HT1A agonist	720	415	0.34
5HR	5 HT reuptake inhibitor	313	165	0.35
D2A	dopamine (D2) antagonist	316	223	0.35
Ren	renin inhibitor	327	207	0.57
Ang	angiotensin II AT1 antagonist	305	179	0.40
Thr	thrombin inhibitor	587	341	0.42
SPA	substance P antagonist	470	237	0.40
HIV	HIV-1 protease inhibitor	223	153	0.45
Cyc	cyclooxygenase inhibitor	512	251	0.27
Kin	protein kinase C inhibitor	234	109	0.32

[a] Unique AF is the number of unique atomic frameworks[29] present in the class. [b] Mean sim is the mean pairwise intra-activity class similarity given by Hert et al.[26]

The virtual screening was carried out as follows: For each activity class, choose an active target structure at random (*T*) and rank the database (consisting of the inactives plus the actives in the same class) in order of decreasing similarity to the target. Count the number of actives in the top 1% of the database. Calculate an enrichment factor based on the expected number of actives retrieved if the compounds were randomly ordered. Repeat this procedure using 30 randomly chosen actives as targets.

The similarity methods used in the comparison were Daylight fingerprints, reduced graph fingerprints, and reduced graph matching using MCES−TC, MCES−FC, and MCIS−FC.

Results were compared in terms of enrichment factors, and also in terms of their ability to identify a range of different scaffolds. One possible measure of the number of different scaffolds retrieved in a similarity search is the number of different atomic frameworks[29] found. This was one of the methods used by Hert et al. to quantify the diversity of hits retrieved in a comparison of topological descriptors using the same MDDR activity classes.[26] Table 2 contains a count of the unique atomic frameworks found in each activity class, and also the mean intraclass similarity as reported by Hert et al. Figure 7 shows some example compounds alongside their atomic frameworks.

The scaffold-finding experiment was carried out as follows:

I. For each activity class

I.A. For each of 10 active target structures chosen at random

I.A.1. Rank the database (consisting of the inactives plus the actives in the same class) in order of decreasing similarity to the target.

I.A.2. Note the identifiers of all the actives in the top 1% of the database.

I.B. Accumulate all the actives found by all the targets. Any active is only counted once, no matter how many times it is found.

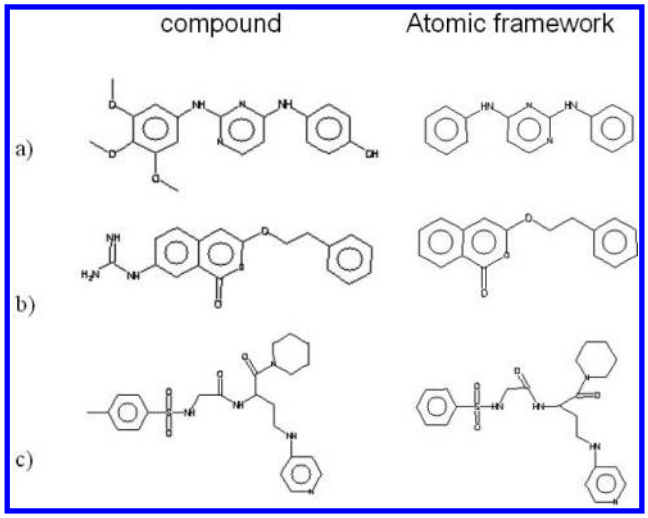

Figure 7. Example atomic frameworks. (a) Protein kinase C inhibitor; (b and c) thrombin inhibitors.

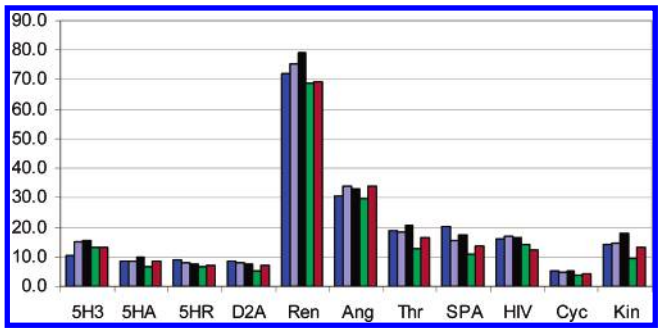

Figure 8. Comparison of virtual screening methods. Bright blue bars, Daylight fingerprints; light blue bars, reduced graph fingerprints; black bars, MCIS−FC; green bars, MCES−TC; red bars, MCES−FC.

I.C. Count the total number of unique atomic frameworks retrieved.

II. Repeat this procedure using 10 sets of 10 targets, to obtain the average number of unique atomic frameworks found by 10 actives.

RESULTS

Virtual Screening. The results of searching the MDDR for 30 actives averaged over 11 activity classes are shown in Table 3 and Figure 8. As expected, all methods perform much better than would random selection, which would give an enrichment factor of 1. MCIS−FC has the highest mean enrichment, 20.9 (Table 3). Fingerprints also perform well, both the reduced graph fingerprints and conventional Daylight fingerprints giving mean enrichments of more than 19.0. MCIS−FC reduced graph matching gives the highest enrichment for 5 of the 11 activity classes; Daylight fingerprints perform best in four activity classes, and reduced graph fingerprints and MCES−FC are each top in one activity class. MCES−TC gives the lowest enrichment for 10 of the 11 activity classes. The excellent performance of all methods for the renin inhibitor class can be attributed to the homogeneous nature and high molecular weight of this activity class, which has the highest mean pairwise similarity of all the classes (Table 2).

Timings for all methods for the protein kinase C activity class are given in Table 4. Although, as expected, the two fingerprint methods proved fastest, reduced graph matching using MCIS−FC certainly proved fast enough for use in the large data set considered in these experiments, and MCES−TC also performed acceptably. These timings are representative of the performance for all activity classes for reduced graph matching using reduced graph fingerprints, MCIS−FC and MCES−TC. However, considerable variability was seen in the performance of MCES−FC, which took between 6 and 20 min for 9 of the 11 activity classes, 3 h 10 min for the HIV-1 protease inhibitors, and more than 24 h for the renin inhibitors. (See the Discussion section for further comments.)

Finding Unique Scaffolds. Table 5 gives the mean number of unique atomic frameworks found by 10 actives, averaged over all 11 activity classes for each similarity method. In the case of MCES−FC, its performance is compared with the other methods for those nine activity classes in which the time taken was acceptable. MCIS−FC is again the best performing method. Daylight fingerprints are second best, and MCES−TC is again the least effective method. MCES−FC performs reasonably well for those activity classes in which it was tested.

Figure 9 gives a comparison between the two best-performing methods, MCIS−FC versus Daylight fingerprints. It is clear that the results depend on the activity class, with

Table 3. Mean Enrichment Factors

method	Daylight fingerprints	RG fingerprints[a]	MCIS−FC	MCES−TC	MCES−FC
mean enrichment	19.4	19.9	20.9	16.4	18.2

[a] RG = reduced graph.

Table 4. Time to Search MDDR for 30 Protein Kinase C Inhibitors

method	Daylight fingerprints	RG fingerprints[a]	MCIS−FC	MCES−TC	MCES−FC
time[b]	4.50	3.20	7.07	12.08	11.20

[a] RG = reduced graph. [b] Time is given in minutes and seconds on a Dell Linux PC with a Xeon 2.8 GHz processor.

Table 5. Mean Number of Unique Atomic Frameworks Retrieved by 10 Actives

method	Daylight fingerprints	RG fingerprints[a]	MCIS−FC	MCES−TC	MCES−FC[b]
unique AF	153.5	150.8	158.0	133.6	148.6

[a] RG = reduced graph. [b] Averaged only over nine activity classes.

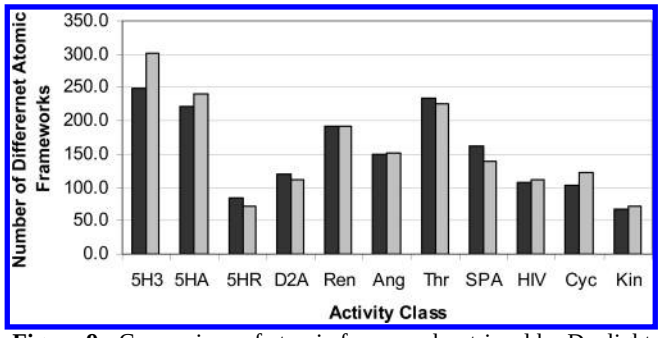

Figure 9. Comparison of atomic framework retrieval by Daylight fingerprints and MCIS−FC. Black bars, Daylight fingerprints; light gray bars, MCIS−FC.

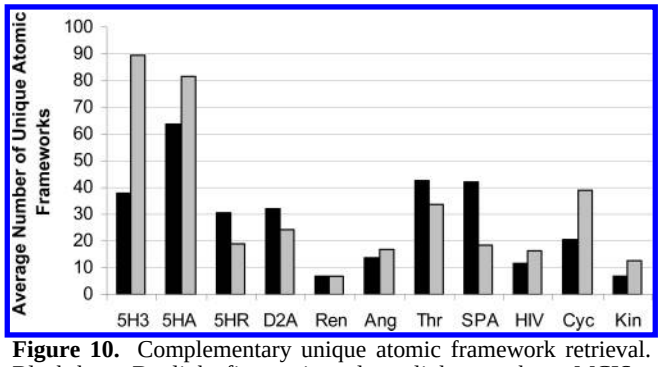

Figure 10. Complementary unique atomic framework retrieval. Black bars, Daylight fingerprints alone; light gray bars, MCIS−FC alone.

reduced graphs finding more unique atomic frameworks in six classes while Daylight fingerprints perform best in four of the classes.

Previous work has shown that reduced graph fingerprints and Daylight fingerprints can be complementary retrieval methods.[17] Here, we find reduced graph matching using the MCIS and Daylight fingerprints to be complementary in terms of the diversity of the atomic frameworks retrieved. On average, over all 11 activity classes, MCIS−FC found 33 atomic frameworks not found using Daylight fingerprints while Daylight fingerprints found 28 atomic frameworks not found by MCIS−FC. Figure 10 compares the number of unique atomic frameworks found by one method and not the other for each activity class. Figure 10 clearly shows that the two methods can be complementary. There are classes such as the 5HT3 antagonists (labeled 5H3), where MCIS−FC finds many more frameworks not found by Daylight fingerprints, but there are also classes such as the 5HT1 antagonists (labeled 5HA) where each method finds many frameworks not found by the other. In the case of the renin inhibitors (labeled REN), very high recall by both methods means that there is little scope for either retrieval method to find unique atomic frameworks—most frameworks are found by both methods.

DISCUSSION

Reduced Graph Matching. Reduced graph matching using the Bron−Kerbosch clique-detection algorithm to find the MCIS has been shown to be a very effective searching method when searching for molecules sharing the same biological activity. In previous work, when reduced graphs were compared using fingerprints, their performance was slightly worse than that of Daylight fingerprints.[17,18] In this work, MCIS−FC has outperformed all the other reduced graph methods considered and also performed better than conventional Daylight fingerprints, giving the highest enrichment in 7 of 11 activity classes. This demonstrates that, when the reduced graphs are matched as graphs, rather than as fingerprints, this matching of their topology captures relationships between features which are lost in a fingerprint representation.

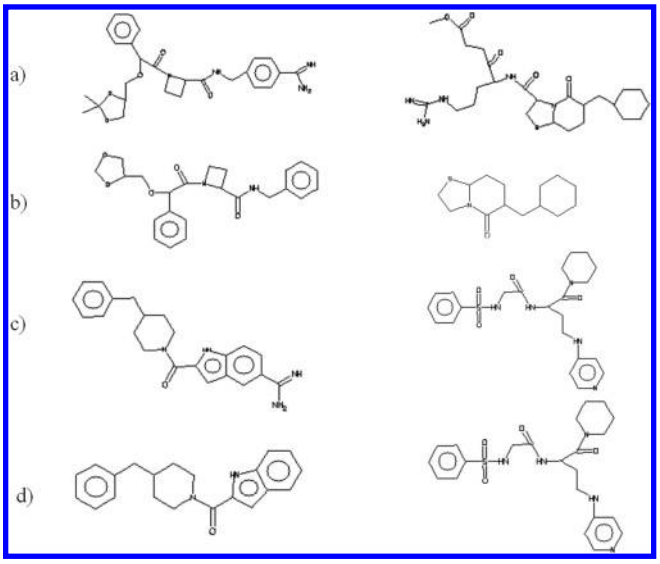

Figure 11. Thrombin scaffold-hopping example. (a) Two thrombin targets; (b) their atomic frameworks; (c) two thrombin actives retrieved by MCIS−FC but not by Daylight fingerprints; (d) their atomic frameworks.

Additionally, more unique atomic frameworks were found by MCIS−FC than by any other method. This demonstrates that reduced graph matching is finding molecules which are less structurally similar to the target than those found by conventional fingerprints, and yet which share the target activity. Thus, reduced graph matching can find alternative scaffolds, one of the main goals of similarity searching. Figure 11 shows two thrombin inhibitor targets (Figure 11a). Their atomic frameworks are quite different (Figure 11b). Two of the actives retrieved by MCIS−FC but not by Daylight fingerprints are shown in Figure 11c, and their atomic frameworks are given in Figure 11d. These two atomic frameworks differ greatly, both from each other and from the target atomic frameworks.

Performance of the MCES Formulation. One possible explanation for the relatively poor retrieval performance of the MCES formulation is that a reduced graph encapsulates a lot of molecular information, with each node representing multiple atoms. The line graph representation, in which a node incorporates two reduced graph nodes plus the edge between them, then condenses even more information into a single node. Two line graph nodes may fail to match even though they do contain matching reduced graph nodes. A failed match, thus, means that some partial match information is lost. Figure 12 shows an extreme example of this. Two simple molecules which differ by a single atom (Figure 12a) are represented by reduced graphs (Figure 12b). Their line graphs (Figure 12c) have no nodes in common, so their MCES−TC similarity is 0. In contrast, MCIS−FC gives a similarity score of 0.5, which is intuitively much more reasonable. This result indicates that the MCES is probably not the best measure of the similarity between a pair of

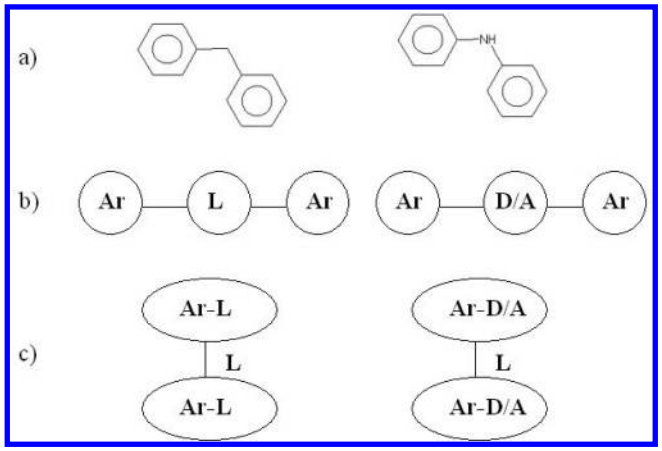

Figure 12. Poor performance of MCES−TC. (a) Molecules; (b) reduced graphs; (c) line graphs.

reduced graphs, despite its demonstrable utility in the comparison of chemical graphs.

The MCES−FC formulation of the problem gives much better enrichment than the TC version, but for some activity classes, particularly the HIV-1 protease inhibitors and the renin inhibitors, this improved effectiveness comes at the expense of decreased efficiency. The reason is that the RASCAL algorithm, using the MCES formulation of the problem, was developed in the context of chemical graphs. The transformation of chemical graphs into line graphs is advantageous since molecules usually contain approximately similar numbers of bonds and atoms. The modular product of the line graphs will then usually not contain many more vertices than a correspondence graph formed directly from the molecules and so is similarly sparsely connected. However, when a reduced graph with N nodes is fully connected, its line graph has $N(N-1)/2$ nodes. Then, for a pair of molecules which are very similar, such as the renin inhibitors, the modular product graph will have a very large number of nodes and edges. For example, the mean number of nodes for all modular product graphs created from pairs of line graphs of the renin inhibitor activity class is 121, while their correspondence graphs (used in the MCIS calculations) have, on average, 57 nodes. In contrast, for the 5HT3 antagonist activity class, which MCES−FC completed in 6 min, the average number of nodes of a modular product graph was 9.4, while the average for a correspondence graph was 14.1. This is a clear illustration of the well-known fact that the performance of a graph-matching algorithm depends on the nature of the graphs being matched. Thus, using the modular product of line graphs is an effective tactic if the graphs are sparse but becomes less so as the graph density increases.

Although MCIS−FC also operates on fully connected graphs, it does not suffer from the combinatorial explosion in nodes associated with the line graph formulation just described. Thus, although an MCIS−FC search for 30 targets in the large MDDR data set took longer than similarity searching using conventional Daylight fingerprints, the time taken was perfectly acceptable. This is due to the relatively small size of the reduced graphs in comparison to the original chemical graphs.

In summary, we have shown that graph matching, using a reduced graph representation of small molecules, is an effective and efficient similarity method, outperforming Daylight fingerprints. Use of the Bron−Kerbosch clique-detection algorithm to find the MCIS between pairs of fully connected reduced graphs was the best graph-matching method of those investigated. The hits found using this method were more diverse than those found by other methods, clearly demonstrating that reduced graphs offer a powerful method for finding novel scaffolds.

ACKNOWLEDGMENT

We thank Jerome Hert for identification of the unique atomic frameworks. This work was funded by the Engineering and Physical Sciences Research Council and AstraZeneca. We thank Daylight Chemical Information Systems Inc. for software support.

REFERENCES AND NOTES

(1) Lajiness, M. S. Dissimilarity-based compound selection techniques. *Perspect. Drug Discovery Des.* **1997**, *7*−*8*, 65−84.
(2) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.
(3) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(4) Bohm, H.-J.; Flohr, M.; Stahl, M. Scaffold Hopping. *Drug Discovery Today: Technol.* **2004**, *1*, 217−224.
(5) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-hopping" by topological pharmacophore search: A contribution to virtual screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894−2896.
(6) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903−911.
(7) Wild, D. J.; Willett, P. Similarity searching in files of three-dimensional chemical structures. Alignment of molecular electrostatic potential fields with a genetic algorithm. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 159−167.
(8) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*, 1653−1666.
(9) Schuffenhauer, A.; Popov, M.; Schopfer, U.; Acklin, P.; Stanek, J.; Jacoby, E. Molecular diversity management strategies for building and enhancement of diverse and focused lead discovery compound screening collections. *Comb. Chem. High Throughput Screening* **2004**, *7*, 771−781.
(10) Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein−protein interaction. *J. Med. Chem.* **2005**, *48*, 1489−1495.
(11) Bostrom, J. Reproducing the conformations of protein-bound ligands: A critical evaluation of several popular conformational searching tools. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 1137−1152.
(12) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical similarity using physiochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.
(13) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391−405.
(14) Gillet, V. J.; Downs, G. M.; Ling, A.; Lynch, M. F.; Venkataram, P.; Wood, J. V.; Dethlefsen, W. Computer-Storage and Retrieval of Generic Chemical Structures in Patents. 8. Reduced Chemical Graphs and Their Applications in Generic Chemical-Structure Retrieval. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 126−137.
(15) Harper, G.; Bravi, G. S.; Pickett, S. D.; Hussain, J.; Green, D. V. S. The reduced graph descriptor in virtual screening and data-driven clustering of high-throughput screening data. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2145−2156.
(16) Takahashi, Y.; Sukekawa, M.; Sasaki, S. Automatic Identification of Molecular Similarity Using Reduced-Graph Representation of Chemical-Structure. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 639−643.
(17) Barker, E. J.; Gardiner, E. J.; Gillet, V. J.; Kitts, P.; Morris, J. Further development of reduced graphs for identifying bioactive compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 346−356.
(18) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity searching using reduced graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338−345.
(19) Barker, E. J. *Chemical Similarity Searching Using Reduced Graphs*; University of Sheffield: Sheffield, U. K., 2004.
(20) Daylight Chemical Information Systems, Inc., Santa Fe, NM.
(21) Raymond, J. W.; Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 521−533.

(22) Raymond, J. W.; Gardiner, E. J.; Willett, P. RASCAL: Calculation of graph similarity using maximum common edge subgraphs. *Comput. J.* **2002**, *45*, 631−644.
(23) Bron, C.; Kerbosch, J. Finding All Cliques of an Undirected Graph [H]. *Commun. ACM* **1973**, *16*, 575−577.
(24) Raymond, J. W.; Willett, P. Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 59−71.
(25) Davis, A. M.; Keeling, D. J.; Steele, J.; Tomkinson, N. P.; Tinker, A. C. Components of successful lead generation. *Curr. Top. Med. Chem.* **2005**, *5*, 421−439.
(26) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256−3266.
(27) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): Evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708−1718.
(28) Tripos Inc., 1699 South Hanley Rd., St. Louis, Missouri, 63144.
(29) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

CI050347R