Similarity Searching Using Reduced Graphs†

Valerie J. Gillet,*,‡ Peter Willett,‡ and John Bradshaw§,#

Department of Information Studies and Krebs Institute for Biomolecular Research, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom, and GlaxoWellcome Research and Development Limited, Gunnels Wood Road, Stevenage, SG1 2NY, United Kingdom

Received August 30, 2002

Reduced graphs provide summary representations of chemical structures. In this work, the effectiveness of reduced graphs for similarity searching is investigated. Different types of reduced graphs are introduced that aim to summarize features of structures that have the potential to form interactions with receptors while retaining the topology between the features. Similarity searches have been carried out across a variety of different activity classes. The effectiveness of the reduced graphs at retrieving compounds with the same activity as known target compounds is compared with searching using Daylight fingerprints. The reduced graphs are shown to be effective for similarity searching and to retrieve more diverse active compounds than those found using Daylight fingerprints; they thus represent a complementary similarity searching tool.

INTRODUCTION

Data mining is becoming an increasingly important technique in drug discovery due to the enormous increase in data about compounds and their activities that is arising from combinatorial chemistry and high-throughput screening programs. Thus, there is a great deal of interest in the development of techniques that are able to identify bioactive molecules from within large data sets. There are a number of potential applications for these techniques, for example, to direct compound acquisition, to assist in the design of combinatorial libraries, and to select compounds for screening. The use of computational techniques to search chemical databases has become known as virtual screening,[1,2] and commonly used techniques include substructure searching; similarity searching; docking; and quantitative structure activity relationships.[3]

Similarity methods have been used for many years and are typically applied early in the drug discovery process when little is known about the biological target. The methods require that molecules are represented by numerical descriptors together with a similarity coefficient that quantifies the degree of resemblance based on the descriptors.[4] Many different descriptors have been used including whole molecule properties; descriptors calculated from 2D representations of molecules; and descriptors that are calculated from the 3D properties of molecules.[2−4] Examples of whole molecule properties include the physicochemical properties molecular weight, molar refractivity, and logP. The most commonly used 2D descriptors are fingerprints that record the presence or absence of molecular fragments within a molecule. Examples of 3D descriptors include 3- and 4-point pharmacophores where a pharmacophore is defined as the spatial arrangement of functional groups required for binding,[5,6] thus, pharmacophore descriptors emphasize the relative positions of atoms or groups of atoms within a molecule that can form hydrogen bonds and hydrophobic interactions.

In recent studies, the ability of 3D descriptors to separate active and inactive compounds has been compared with a variety of 2D descriptors.[7−11] Despite the fact that drug-receptor binding is a 3D event governed by the spatial arrangement of functional groups with different binding properties, the studies have found 2D fingerprints to be more effective. The major difficulty associated with 3D descriptors, especially for large data sets, is the problem of handling conformational flexibility, and it is likely to be the inability to handle this effectively that has resulted in the poor performance of the descriptors.

The success of 2D fingerprints for similarity searching is surprising since they were originally developed for substructure searching. However, they are particularly effective at identifying close analogues and are less effective at finding compounds that share the same activity but that are based on different lead series. Some of their limitations in this respect have been identified by Flower,[12] for example, fingerprints usually distinguish atoms on the basis of element and therefore physicochemcial equivalencies between different element types are not perceived. Thus, Flower suggests that the problems with bit-strings may arise from the features that are encoded within them rather than with the bit-strings themselves. The fact that characterizing atoms by element types can be too specific for similarity searching has also been recognized by Kearlsey et al.[13] who modified the atom-pairs and topological torsion descriptors described by Cahart et al.[14] and Nilakantan et al.[15] to encode physicochemical properties of atoms.

Recent studies have recognized the complementarity of 2D and 3D descriptors for similarity searching,[10,11] whereby each method tends to retrieve a different set of active compounds. Thus, there can be value in applying more than one similarity method to a given problem. Furthermore,

* Corresponding author e-mail: v.gillet@sheffield.ac.uk.

† First presented at the Fifth International Conference on Chemical Structures, Noordwijkerhout, The Netherlands, June 1999.

‡ University of Sheffield.

§ GlaxoWellcome Research and Development Limited.

Current address: Daylight Chemical Information Systems Inc., Sheraton House, Castle Park, Cambridge, CB3 0AX UK.

10.1021/ci025592e CCC: $25.00
Published on Web 02/12/2003

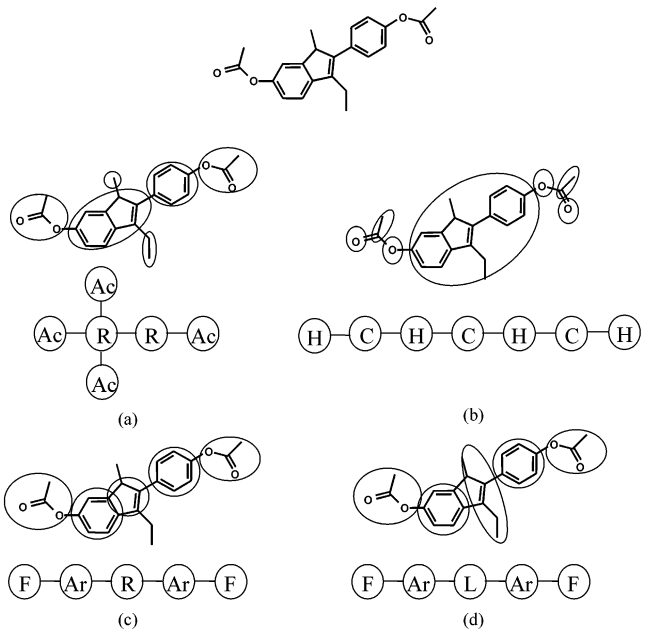

Figure 1. Examples of different types of reduced graphs that can be generated for the structure shown: (a) nodes correspond to ring systems (R) and connected acyclic components (Ac); (b) nodes correspond to connected carbon components (C) and connected heteroatom components (H); (c) nodes correspond to aromatic rings (Ar), aliphatic rings (R), and functional groups (F); (d) nodes correspond to aromatic rings (Ar), functional groups (F), and linking groups (L).

Sheridan and Kearsley[3] have observed that the effectiveness of any one method can vary greatly from one type of activity to another in a way which is not easy to predict and thus advocate the use of several search methods where possible.

In this paper we explore the potential of *reduced graphs* in the context of similarity searching. The representation of a chemical structure as a topological graph has long been recognized and forms the basis of current structure and substructure search systems.[16] Typically the nodes of the graph represent the atoms of the chemical structure and the edges in the graph represent its bonds. We refer to this type of graph as a *chemical graph*. In a reduced graph, each node represents a group of connected atoms, and an edge exists between two nodes if there is a bond in the original structure between an atom contained within one node and an atom in the other node. Thus, reduced graphs summarize the features of molecules while retaining the topology between the features. They allow the generalization of molecules in such a way that it is the properties of atoms or groups of atoms that are identified along with their topology (i.e., the way in which they are interconnected). Thus, they offer the potential of finding similarities between compounds that belong to different lead series. Many different types of graph reduction are possible, for example, Figure 1 shows a chemical graph together with four different reduced graph representations.

Reduced graphs were originally developed for structure and substructure searching of generic chemical structures.[17] Generic or Markush structures provide a compact representation of a set of specific structures with some common structural feature. They occur most frequently in chemical patent specifications where an entire class of compounds is protected. The complexities of generic structures give rise to a low screenout rate when using conventional fragment screening, and thus novel search methods are required compared to searching in databases of specific substances. Consequently, reduced graph representations of generic structures were developed to provide an additional level of search over the traditional two stage search in use for specific structures, i.e., fragment screening followed by atom-by-atom searching.

More recently reduced graphs have been used for similarity searching. Takahashi et al.[18] describe an approach to the identification of molecular similarity using reduced graphs. The nodes in their reduced graphs are defined with the aid of a dictionary of fragments, and there can be a one-to-many mapping between atoms and nodes, i.e., an atom may belong to more than one node simultaneously. The edges in the reduced graph are weighted by the distance between the nodes. The reduced graph representations of two different molecules are then compared using clique detection to find the docking graph, i.e., a mapping between the nodes and edges in one reduced graph with the nodes and edges in the other. 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. To our knowledge this method was never applied at the database level.

Fisanick et al.[19] describe a 2D similarity search system based on reduced graphs that has been developed at CAS. They use reduced graphs to filter the output from a conventional bit-string similarity search. The query structure is described in generic terms using node descriptors and a graph matching procedure is used to eliminate those hits that do not match at the reduced graph level.

Rarey and Dixon[20,21] have described an approach to similarity searching known as Feature Trees. Their method is conceptually similar to the reduced graphs described here; however, differences lie in the way the trees are defined and also in the way in which two Feature Trees are compared.

In the present study, our aim is to determine a level of graph reduction that is appropriate for identifying compounds with the same bioactivity. Ideally, the features of the compounds should be generalized so that compounds with equivalent bioactivity are identified as similar and compounds with different bioactivity are identified as dissimilar. Accordingly, the graph reduction will focus on structural features that are thought to be relevant to drug-receptor binding and the topology between the features will be retained. Thus we are attempting to describe structures as topological pharmacophores rather than the more usual topographical or 3D pharmacophores. The methods overcome some of the limitations that arise from the use of 2D fingerprinting methods while at the same time avoid the need to consider 3D conformational issues. The graph reduction process is algorithmic, that is, it does not require a dictionary of fragments and is rapid enough that it can be applied to large data sets.

METHODS

Types of Reduced Graph. As already mentioned, the aim of the reduced graphs is to emphasize potential binding groups and their relative positions within a structure. Consequently, we have chosen to concentrate on ring systems

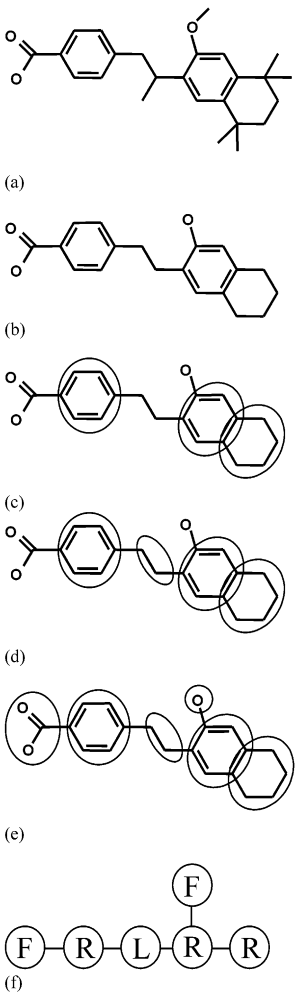

Figure 2. Steps involved in generating a Ring/Feature reduced graph. (a) The chemical structure. (b) Hydrogen-bonding atoms are identified and terminal non-hydrogen-bonding atoms are removed, recursively. (c) The smallest set of smallest rings is identified and each ring reduced to a Ring Node. (d) The connected isolating carbons are reduced to Link Nodes. (e) The remaining connected acyclic parts form Feature Nodes. (f). The nodes are linked by edges to form an R/F reduced graph.

and hydrogen-bonding groups and have focused on two different types of reduced graph, namely, Ring/Feature (R/F) reduced graphs and Aromatic/Feature (Ar/F) reduced graphs.

In R/F reduced graphs, three different node types are possible: Ring Nodes; Feature Nodes; and Link Nodes. A Ring Node is defined by the smallest set of smallest rings as defined in the Daylight Chemical Information System,[22] with each individual ring being reduced to a Ring Node. Link Nodes are determined using the concept of an *isolating carbon*,[22] which is defined as a carbon atom that is not doubly- or triply-bonded to a heteroatom. Connected isolating carbons form Link Nodes. The remaining connected acyclic regions that contain hydrogen-bonding atoms become Feature Nodes, and there is a many-to-one correspondence between atoms in the chemical graph and Feature Nodes in the reduced graph with each atom being assigned to one Feature Node. Terminal acyclic groups that do not contain hydrogen-bonding atoms are removed recursively.

The steps involved in generating an R/F reduced graph are summarized in Figure 2. All programming was done using the Daylight Toolkit.[22] The first step is to identify all

Table 1. SMARTS Definitions for Substructural Features

feature	SMARTS
donor	[!#6;!H0]
acceptor	[$([!#6;+0]);!$([F,Cl,Br,I]);!$([o,s,nX3]);!$([Nv5,Pv5,Sv4,Sv6])]

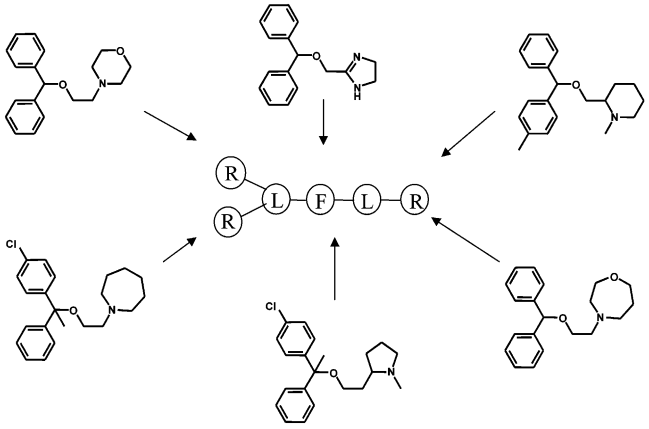

Figure 3. Examples of chemical structures that reduce to the same R/F reduced graph.

atoms in the chemical graph that can act as hydrogen bonding atoms (donors, acceptors, and atoms with the ability to act as both donors and acceptors). This is done using SMARTS definitions of donors and acceptors that are read in from an input file. The definitions used in the work described here are shown in Table 1; however, it should be noted that it is easy to change the definitions simply by changing the definitions in the input file. Heteroatoms that remain in the chemical graph but that are not identified as hydrogen bonding are treated as carbon atoms. The next step is to recursively delete all singly connected non-hydrogen bonding atoms. This is done since the R/F reduced graphs are designed to summarize rings and hydrogen bonding features, only. No account is taken of acyclic hydrophobic features. Next, the rings are identified and reduced to Ring Nodes. Then, all connected isolating carbon atoms are reduced to Link Nodes and the remaining connected acyclic parts of the graph become Feature Nodes. The final step is to form edges between the nodes in the reduced graph that correspond to bonds in the original chemical graph. Figure 3 shows a series of chemical graphs that reduce to the same R/F reduced graph.

In Ar/F reduced graphs, three different node types are possible: Aromatic Nodes; Feature Nodes; and Link Nodes. Acyclic rings are not represented explicitly but are treated as if they are acyclic groups so that they can become incorporated within Feature Nodes or Link Nodes or are eliminated completely if they are terminal and do not contain any hydrogen-bonding atoms. This is achieved by preprocessing the chemical graph to open acyclic rings by removing two-connected non-hydrogen bonding aliphatic ring atoms. In some cases this can lead to disconnected components in which case the atoms in the shortest path are reinserted. Where there is more than one shortest path, one of the paths is chosen arbitrarily. (Note that this method fails on molecules that have aliphatic rings that do not contain any two-connected atoms and therefore cannot be broken; however, such rings occur infrequently.) Opening the acyclic rings may result in new terminal non-hydrogen-bonding atoms which are then deleted. The steps involved in

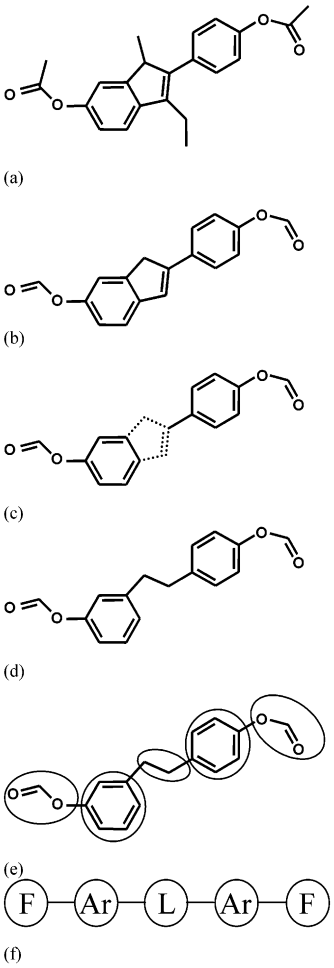

Figure 4. Steps involved in generating an Aromatic/Feature reduced graph. (a) The chemical structure. (b) Hydrogen-bonding atoms are identified and terminal non-hydrogen-bonding atoms are removed, recursively. (c) Aliphatic rings are broken by removing 2-connected atoms, any newly created terminal non-hydrogen-bonding atoms are removed, recursively. (d) Disconnected fragments are reconnected by reinserting a shortest path. (e) Aromatic Nodes; Link Nodes; and Feature Nodes are identified.

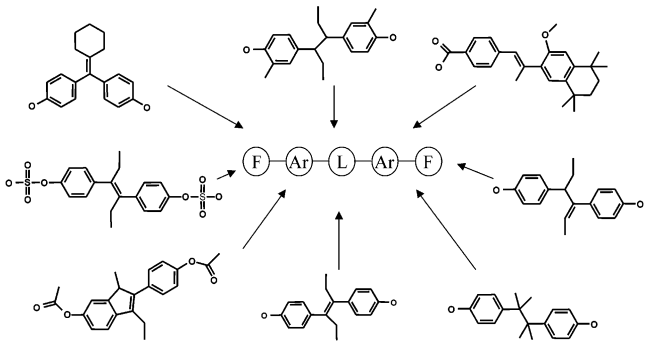

Figure 5. Example chemical structures that reduce to the same Ar/F reduced graph.

generating Ar/F reduced graphs are illustrated in Figure 4. Figure 5 shows a number of structures that reduce to the same Ar/F reduced graph.

Labeling Nodes. The information content of the reduced graph nodes can be varied and a hierarchy of levels of description is possible. In the current work, we have used a four level hierarchy as shown in Figure 6. The hierarchy is based on Ring Nodes where at level 1, no additional label is included and just the node types themselves are considered;

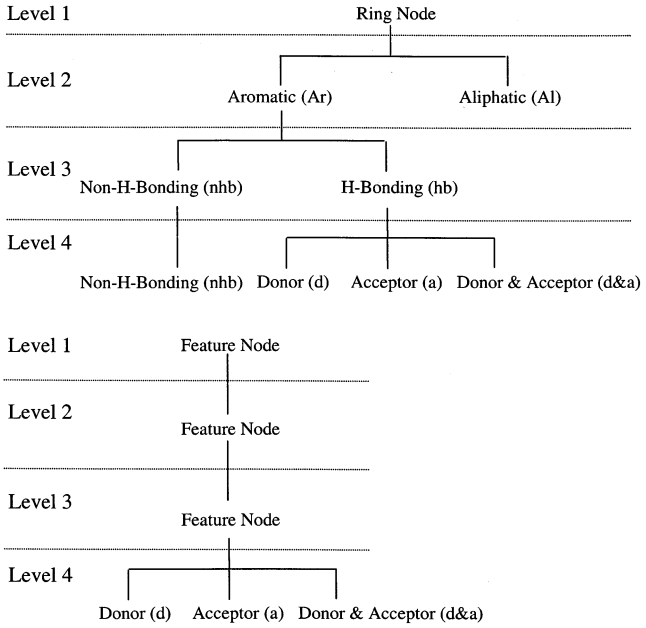

Figure 6. The node labeling scheme.

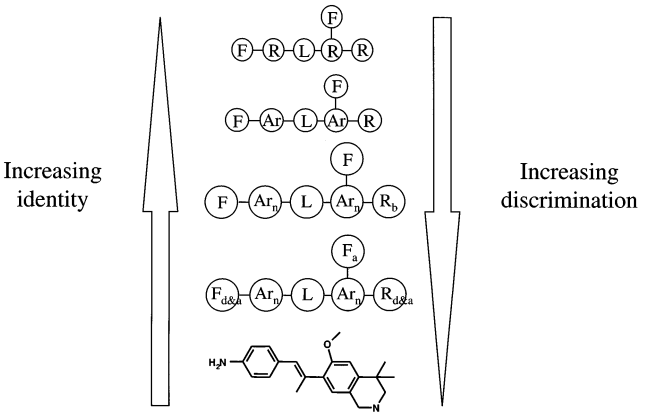

Figure 7. A hierarchy of reduced graphs exists.

at level 2 the nodes are differentiated as aromatic or aliphatic ring nodes; at level 3 they are further differentiated as hydrogen-bonding (i.e. containing at least one hydrogen-bonding atom) or non-hydrogen-bonding; and at level 4, the most detailed level, the nodes are differentiated as hydrogen bond donor, hydrogen bond acceptor, or both donor and acceptor. Feature Nodes are by definition hydrogen-bonding, thus, they are treated the same at levels 1, 2, and 3. At level 4, they are differentiated as donor, acceptor, or donor and acceptor. Aromatic Nodes are by definition aromatic, and thus they are treated identically at levels 1 and 2. Link Nodes are treated identically at all levels of the hierarchy, since by definition, they do not contain hydrogen-bonding atoms.

The combination of two types of reduced graphs (R/F and Ar/F) and four different labeling schemes leads to seven distinct representations. These are referred to as R/F(4) (where the number in brackets refers to the level of node labeling); R/F(3); R/F(2); R/F(1); Ar/F(4); Ar/F(3) and Ar/F(2). (Ar/F(1) is identical to Ar/F(2)).

The different levels of node labeling give rise to a hierarchy of reduced graphs as shown in Figure 7, where a reduced graph higher up in the hierarchy can be considered to be less discriminating than one lower in the hierarchy, for example, R/F(1) is less discriminating than R/F(2) which

Table 2. Activity Classes Extracted from WDI Are Shown

activity class	no. actives
narcotics	121
antihistamines	162
tranquilizers	102
dopaminergics	102
angiotensin-antagonists	80
estrogens	182
parasympathomimetics	76

is less discriminating than R/F(3), etc. Another way of viewing the hierarchy is that at the most discriminating level every structure in a database has a unique chemical graph (ignoring stereochemistry, etc.). As we ascend the hierarchy of levels of reduction then grouping of structures occurs with a many-to-one relationship between structure and reduced graph, i.e., there are fewer unique reduced graphs than structures.

Comparing Reduced Graphs. The effectiveness of the reduced graphs is investigated by performing similarity searches using different classes of active compounds. Thus, a means of performing pairwise similarity comparisons of the reduced graphs is required. This is achieved by converting the reduced graphs into pseudo-SMILES representations by mapping each node type to a different heavy element type and by mapping the edges in the reduced graph to single bonds. Fingerprints can then be generated from the pseudo-SMILES using the Daylight toolkit and can be used in similarity searches in the same way as Daylight fingerprints that are generated directly from the chemical graph.

EXPERIMENTAL RESULTS

The effectiveness of reduced graphs at identifying bioactivity classes was investigated using data derived from the World Drugs Index.[23] Initially, the WDI was preprocessed so that parent structures only were included and charges were neutralized where possible to allow for a consistent definition of hydrogen bond donors and acceptors. Activity classes were extracted using keywords as described by Kearlsey et al.[13] who used the following criteria for selection: (1) the majority of the actives in the therapeutic area of the probe should work by the same mechanism as the probe (but given the limitations of the database there is no way to ensure that all actives work by the same mechanism: thus some actives are false actives in this respect); (2) there should be at least 50 actives to ensure reasonable statistics. A random sample of 3000 of the remaining WDI compounds was then added to each of the activity classes and the compounds were labeled as inactives. It should be noted here that since every compound has not been tested in every activity class there may be false negatives in the inactive sample.

The seven activity classes we used are listed in Table 2 which also shows the number of actives in each class.

Reduced graphs of each type and at each level of node labeling were generated for each of the data sets and converted first to pseudo-SMILES and then to fingerprints. The relative effectiveness of the reduced graphs and Daylight fingerprints at identifying molecules with the same activity was measured by performing similarity searches and calculating enrichment factors as described below.

An active compound was used as the target compound in a similarity search, and a Target Enrichment Factor, shown as E_T below, was calculated from the nearest neighbors list produced. The Target Enrichment Factor is defined as the number of compounds in the top 300 nearest neighbors of the target compound that share the same activity as the target divided by the number that would be expected to be found if the actives were distributed evenly throughout the entire list. Thus

$$E_T = \frac{n_a}{\left(\frac{N_A}{N} \times 300\right)}$$

where n_a is the number of actives in the top 300 positions of the ranked list; N_A is the number of actives in the data set; and N is the total number of compounds in the data set.

The Average Enrichment Factor, E_A, was then calculated by taking each active compound as target in turn and averaging the Target Enrichment Factors as shown:

$$E_A = \frac{\sum_{T \in A} E_T}{N_A}$$

RESULTS

Table 3 shows the average number of actives retrieved for each of the seven different reduced graphs for each activity class. The best result is italicized. It can be seen that the Ar/F reduced graphs perform better for four of the activity classes (estrogens, dopamines, antihistamines, and narcotics) and that the R/F reduced graphs perform better for the other three activity classes (angiotension antagonists, parasympathomimetics, and tranquilizers). As far as the levels of discrimination are concerned, level 1 and level 2 do not appear to be sufficiently discriminating, and no clear pattern has emerged between level 3 and level 4 with the relative performance depending on the data set.

The results are shown as Average Enrichment Factors in Table 4 and are compared with the results found using conventional Daylight fingerprints. Again the best performing method is italicized. It can be seen that Daylight gives the best performance in six out of the seven data sets. The only data set where reduced graphs show better performance is the antihistamines.

As described in the Introduction, Daylight fingerprints are well suited to finding close analogues (so called "me-too" compounds), whereas reduced graphs offer the potential to be able to identify more diverse compounds that share the same activity, for example, that may have different carbon skeletons or different element types. Simply calculating enrichment factors fails to take into account the actual compounds retrieved by each method. Hence, the actives occurring in the near neighbor lists are compared for the different search methods. For each activity class, the nearest neighbor lists generated using the best performing R/F reduced graph and the best performing Ar/F reduced graph were each compared with the nearest neighbor lists generated using Daylight fingerprints. For each active compound, the number of actives retrieved by the reduced graph method that were not retrieved using Daylight fingerprints (unique actives) was recorded. The number of unique actives was then averaged over all targets in an activity class.

Table 3. Number of Actives Retrieved Averaged over All Actives in Each Activity Class[a]

	R/F(4)	R/F(3)	R/F(2)	R/F(1)	Ar/F(4)	Ar/F(3)	Ar/F(2)
angiotensins	*33.85 (17.34)*	32.74 (17.5)	33.58 (18.43)	28.09 (14.06)	32.50 (15.18)	23.05 (9.49)	20.79 (9.98)
estrogens	87.66 (39.68)	95.49 (37.99)	88.03 (38.96)	68.53 (30.80)	75.62 (28.56)	*99.67 (29.01)*	91.15 (29.83)
dopamines	32.71 (16.77)	31.97 (16.81)	34.80 (15.76)	22.71 (10.38)	*40.25 (14.35)*	39.34 (14.12)	36.59 (12.56)
antihistamines	43.61 (23.40)	39.00 (18.82)	37.92 (17.61)	33.17 (18.24)	50.45 (26.41)	40.07 (20.77)	43.27 (23.37)
narcotics	28.98 (11.03)	28.92 (11.35)	25.98 (11.30)	22.43 (12.06)	37.05 (13.68)	*45.84 (23.15)*	42.29 (19.13)
para-sym	15.64 (9.11)	*16.12 (9.81)*	14.41 (8.31)	14.41 (8.38)	15.26 (10.34)	14.96 (9.89)	15.39 (9.49)
tranquilizers	18.71 (10.11)	*20.79 (13.38)*	19.60 (13.03)	17.53 (10.37)	16.61 (6.81)	14.17 (5.22)	13.26 (4.64)

[a] Standard deviations are given in brackets.

Table 4. Average Enrichment Factors (E_A) for Different Activity Classes Using Conventional Daylight Fingerprints Derived from the Chemical Graphs (CG) and for Reduced Graphs

	R/F(4)	R/F(3)	R/F(2)	R/F(1)	Ar/F(2)	Ar/F(3)	Ar/F(4)	D
angiotensins	4.34	4.20	4.31	3.61	2.57	2.85	4.02	*5.10*
estrogens	5.11	5.56	5.13	3.99	5.15	5.63	4.27	*6.29*
dopamines	3.32	3.24	3.53	2.30	3.68	3.96	4.05	*4.18*
antihistamines	2.84	2.54	2.47	2.16	2.80	2.59	*3.26*	2.53
narcotics	2.49	2.49	2.23	1.93	3.57	3.86	3.12	*5.49*
para-sym	2.11	2.19	1.12	1.12	2.11	2.05	2.09	*2.78*
tranquilizers	1.90	2.11	1.99	1.78	1.33	1.43	1.67	*2.28*

Table 5. Average Number of Compounds Retrieved for Each Activity Class that Were Unique to the Reduced Graphs, i.e., Were Not Retrieved Using Daylight Fingerprints[a]

	R/F		Ar/F	
	unique hits	percent unique	unique hits	percent unique
angiotensins	6.23	18.4	6.90	20.4
estrogens	6.60	7.5	19.71	19.8
dopamines	8.60	26.3	12.9	32.0
antihistamines	22.40	51.4	26.68	52.9
narcotics	1.69	5.8	6.79	14.8
para-sym	7.01	44.8	6.67	41.4
tranquilizers	8.90	47.6	8.71	41.9

[a] The columns headed percent unique gives the unique hits as a percentage of the total compounds found using the reduced graph.

Results are shown in Table 5 as the average number of unique actives retrieved by the reduced graph method and as a percentage of the total number of actives retrieved by the reduced graph. The percentages are shown graphically in Figure 8a. For the R/F reduced graphs, the percentages vary from 5.8% for the narcotics to 51.4% for the antihistamines. Thus, for the antihistamines more than 50% of the actives retrieved using the reduced graph were not found using Daylight fingerprints. The percentages are higher for the Ar/F reduced graphs (except for the parasympathomimetics and tranquilizers) and range from 14.8% for the narcotics to 52.9% for the antihistamines, Figure 8b. Thus, even though the reduced graphs tend to give lower overall performance than the Daylight fingerprints, they are successful in identifying actives not found using Daylight and thus reduced graphs can be said to complement the Daylight fingerprints.

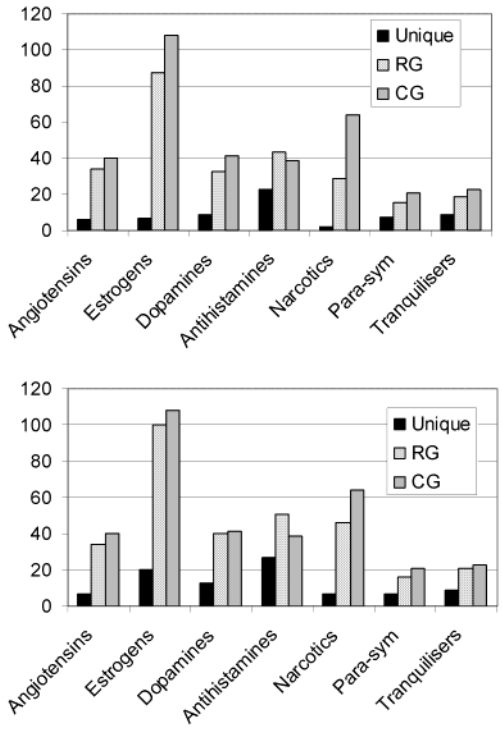

Figure 8. a. The average number of actives appearing in the top 300 nearest neighbor lists using Daylight fingerprints (grey); R/F(4) reduced graphs (hatched). The black bar shows the average number of actives found using the reduced graphs that are not found using Daylight fingerprints. b. The average number of actives appearing in the top 300 nearest neighbor lists using Daylight fingerprints (grey); Ar/F(4) reduced graphs (hatched). The black bar shows the average number of actives found using the reduced graphs that are not found using Daylight fingerprints.

DISCUSSION

The reduced graphs have been shown to be effective representations for searching for compounds which share the same activity. They give Average Enrichment Factors that are significantly better than random across a range of different data sets. When compared to Daylight fingerprints, the overall results show that reduced graphs give slightly worse enrichments; however, their strength lies in the fact that they are effective at identifying different actives to those found using Daylight fingerprints. Experiments have shown that up to 50% of the actives found using reduced graphs are unique to this representation.

The differences in the two approaches are emphasized using two example searches. Figure 9 shows some of the dopaminergic actives that are unique to the top 300 nearest neighbors when using reduced graphs and apomorphine as

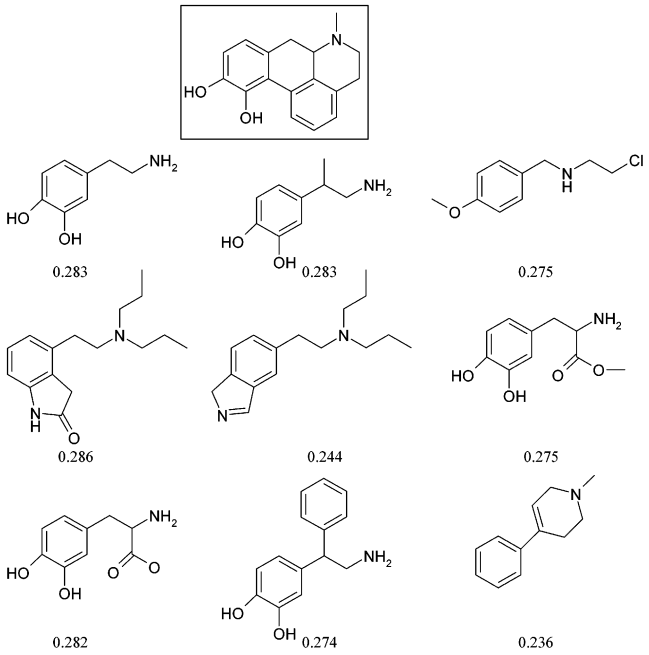

Figure 9. Compounds with dopaminergic activity that are found when searching for compounds similar to apomorphone using reduced graphs but that are missed when searching using Daylight fingerprints. The Tanimoto similarity using Daylight fingerprints of each compound to the target is shown.

the target compound. The Tanimoto similarities measured using conventional Daylight fingerprints are shown. It can be seen that the Daylight similarities are low due to differences in the carbon skeletons and in the exact characterization of the functional groups. These compounds all score highly when their reduced graphs are compared since these emphasize the topological relationships between the rings and the hydrogen-bonding properties of the functional groups.

Figure 10 shows compounds with estrogen activity that are unique to reduced graphs when diethylstilbestrol is used as target. The Tanimoto similarities calculated using Daylight fingerprints of each compound to the target are also shown. A similar rationale to that presented for the dopaminergics can be used to explain why these compounds occur much higher in the near neighbor list when using reduced graphs relative to Daylight fingerprints.

In addition, when the different reduced graphs are compared it is apparent that no one type of graph reduction is optimal for all data sets. For example, both the R/F reduced graphs and the Ar/F reduced graphs were shown to be effective at finding actives not found using Daylight fingerprints and given the differences between the two types of reduced graphs it is likely that they complement one another as well as each complementing Daylight fingerprints.

The complementary nature of the reduced graphs and Daylight fingerprints suggests that it should be possible to combine the approaches in order to improve on the results of using a single method.[24]

Here we focused on one particular node labeling scheme; however, many different node labelings are also possible, e.g., number of atoms contained with a node (ring sizes); and specifiying the number of atoms in a Link Node. Edge labeling is also possible, for example, an edge arising from fused rings could be labeled differently to an edge derived from an acyclic bond.

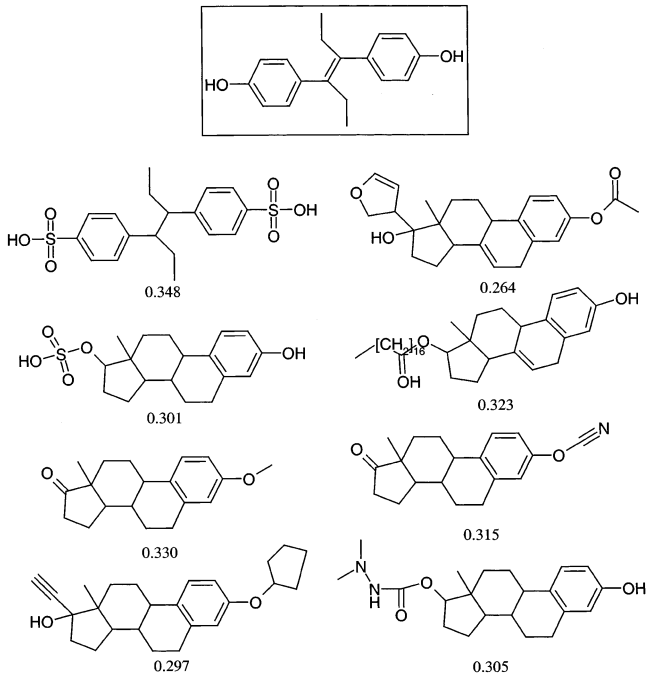

Figure 10. Compounds with estrogen activity that are found when searching for compounds similar to diethylstilbestrol using reduced graphs but that are missed when searching using Daylight fingerprints. The Tanimoto similarity using Daylight fingerprints of each compound to the target is shown.

Encoding the reduced graphs as pseudo-SMILES and then converting these to fingerprints provided a rapid way of investigating the potential of reduced graphs for similarity searching. However, the fingerprints are unlikely to be optimal as far as reduced graphs are concerned due to the different characteristics of the graphs relative to the chemical graphs that the fingerprints were designed to represent. For example, the average number of nodes in a reduced graph is much less than the average number of atoms in a chemical structure and hence the fingerprints generated from reduced graphs are much less dense. Second the connectivity of reduced graph nodes can be higher than that found in typical organic molecules, for example, a substituted ring will reduce to a single Ring Node with as many connecting nodes as there are substituents.

Furthermore, the Daylight fingerprint is a hashed fingerprint where there is a many-to-one correspondence between bits set in the fingerprint and a path in the chemical graph (or reduced graph). Collisions may occur where the set of bits associated with one path overlaps with the set of bits corresponding to another. Thus, it is not possible to map back from a bit in the fingerprint to a particular characteristic of the reduced graph, i.e., the fingerprint is not readily interpretable. The small size of the reduced graphs and the limited number of node types means that it is possible to explore other types of fingerprint representation.

The overall aim of this study was to investigate the potential of reduced graphs for similarity searching. While the potential has been demonstrated, several ways in which the method can be improved have been identified. Many of these suggestions are explored in the companion paper.[25]

ACKNOWLEDGMENT

We thank GlaxoWellcome Research and Development for funding and Daylight Chemical Information Systems Inc.

for software support. The Krebs Institute for Biomolecular Research is a designated Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

REFERENCES AND NOTES

(1) Böhm, H.-J.; Schneider, G., Eds.; *Virtual Screening for Bioactive Molecules*; Wiley-VCH: Weinheim, 2000.
(2) Bajorath, J. Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233−245.
(3) Sheridan, R. P.; Kearsley, S. K. Why Do We Need So Many Chemical Similarity Search Methods? *Drug Discov. Today.* **2002**, *7*, 903−911.
(4) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(5) *Pharmacophore Perception, Development, and Use in Drug Design*; Güner, O. F., Ed.; International University Line: La Jolla, CA, 2000.
(6) Beno, B. R.; Mason, J. S. The Design of Combinatorial Libraries Using Properties and 3D Pharmacophore Fingerprints. *Drug Discovery Today* **2001**, *6*, 251−258.
(7) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 572−584.
(8) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand−Receptor Binding. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 1−9.
(9) Matter, H. Selecting Optimally Diverse Compounds from Structural Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem*. **1997**, *40*, 1219−1229.
(10) Briem, H.; Kuntz, I. D. Molecular Similarity Based on Dock-Generated Fingerprints. *J. Med. Chem*. **1996**, *39*, 3401−3408.
(11) Schuffenhauer, A.; Gillet, V. J.; Willett, P. Similarity Searching in Files of Three-Dimensional Chemical Structures: Analysis of the BIOSTER Database Using Two-Dimensional Fingerprints and Molecular Field Descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 295−307.
(12) Flower, D. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 379−386.
(13) Kearsley, S. K.; Sallamack, S.; Fluder, E.; Andose, J. D.; Mosley, R. T. Sheridan, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.
(14) Cahart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Application. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(15) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan R. Topological Torsions: a New Molecular Descriptor for SAR Applications, Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.
(16) Barnard, J. M. Substructure Searching Methods: Old and New. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 532−538.
(17) Gillet, V. J.; Downs G. M.; Holliday, J. D.; Lynch, M. F.; Dethlefsen W. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 13. Reduced Graph Generation. *J. Chem. Inf. Comput. Sci*. **1991**, *31*, 260−270.
(18) 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.
(19) Fisanick, W.; Lipkus, A. H.; Rusinko, A., III Similarity Searching on CAS Registry Substances. 2. 2D Structural Similarity. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 130−140.
(20) Rarey, M.; Dixon, J. S. Feature Trees: A New Molecular Similarity Measure Based on Tree Matching. *J. Comput.-Aided Mol. Design*. **1998**, *12*, 471−490.
(21) Rarey, M.; Stahl, M. Similarity Searching in Large Combinatorial Chemistry Spaces. *J. Comput.-Aided Mol. Design*. **2001**, *15*, 497−520.
(22) Daylight Chemical Information Systems, Inc., Mission Viejo, CA, www.daylight.com.
(23) The World Drug Index is available from Derwent Information, 14 Great Queen St., London W2 5DF, UK.
(24) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of Molecular Similarity Measures Using Data Fusion. *Perspec. Drug Discov. Design* **2000**, *20*, 1−16.
(25) Barker, E.; Gardiner, E.; Gillet, V. J.; Kitts, P.; Morris, J. Further Developments of Reduced Graphs for Identifying Bioactive Compounds. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 346−356.

CI025592E