Further Development of Reduced Graphs for Identifying Bioactive Compounds

Edward J. Barker,[†] Eleanor J. Gardiner,[†] Valerie J. Gillet,[*,†] Paula Kitts,[‡] and Jeff Morris[‡]

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 30, 2002

Reduced graphs provide summary representations of chemical structures. Here, a variety of different types of reduced graphs are compared in similarity searches. The reduced graphs are found to give comparable performance to Daylight fingerprints in terms of the number of active compounds retrieved. However, no one type of reduced graph is found to be consistently superior across a variety of different data sets. Consequently, a representative set of reduced graphs was chosen and used together with Daylight fingerprints in data fusion experiments. The results show improved performance in 10 out of 11 data sets compared to using Daylight fingerprints alone. Finally, the potential of using reduced graphs to build SAR models is demonstrated using recursive partitioning. An SAR model consistent with a published model is found following just two splits in the decision tree.

INTRODUCTION

The development of combinatorial chemistry and high-throughput screening techniques has led to a massive increase in the number of compounds that can be synthesized and tested for activity. However, simply increasing the throughput of the synthesis and test cycle in itself does not necessarily lead to more high quality lead compounds. Thus, there is increasing interest in the use of computational tools, both for the analysis of high-throughput screening data in order to derive models for activity and in virtual screening techniques.[1] Many of the techniques that have been developed are based on the concept of molecular similarity.[2]

Similarity searching requires the definition of a chemistry space through the use of molecular descriptors together with a way of quantifying the degree of similarity of molecules within the space. Many different descriptors have been developed, ranging from whole molecule properties such as molecular weight and logP to descriptors that are derived from the 2D representation of molecules such as 2D fingerprints and 3D descriptors such as molecular volume or pharmacophoric keys.[3] One of the challenges in using similarity methods is in choosing the best descriptors for a particular task.

Comparative studies have been performed on the effectiveness of different descriptors at identifying molecules with similar activity:[4−6] these studies have shown that 2D descriptors, such as fragment-based bitstrings and hashed fingerprints, can be more effective than 3D descriptors. The 2D fingerprint methods have been shown to be very good at identifying structural analogues; however, they are less effective at identifying compounds that have similar activities but that do not share the same structural skeleton. They were originally developed for substructure searching, and some of their limitations for similarity searching have recently been identified.[7−9] Other comparative studies have shown that 3D descriptors can be highly selective;[10] however, effective handling of conformational flexibility remains a major difficulty in the use of 3D descriptors. Recently, 2D and 3D descriptors have been shown to be complementary, and improved results have been found by combining different descriptors.[11]

We explored the use of the reduced graph for similarity searching in the companion paper.[12] Reduced graphs are topological graphs that attempt to summarize the features of molecules that can result in drug-receptor binding while retaining the connections between the features. They provide a more generalized representation than 2D fingerprints such as Daylight and UNITY fingerprints, and whereas 2D fingerprints are very effective at finding close analogues, reduced graphs offer the potential to be able to find structures that have the same binding characteristics but that have different carbon skeletons and hence may belong to different lead series. Since reduced graphs are topological representations they avoid the need to generate 3D conformations.

The previous work demonstrated the potential of reduced graphs for similarity searching; however, many different types of graph reduction are possible. In this study, we explore a number of different types of reduced graphs and different ways of representing them in fingerprints. The effectiveness of the reduced graphs is investigated using similarity searching over a variety of different activity classes. We then use data fusion to see if improved performance can be achieved by combining different descriptors. Finally, reduced graphs are used as the descriptors in recursive partitioning in an attempt to build a structure−activity relationship (SAR).

METHODS

In the following, the different reduced graphs are presented first according to the different node definitions that are used to summarize the features of molecules and second by the

[*] Corresponding author e-mail: v.gillet@sheffield.ac.uk.
[†] University of Sheffield.
[‡] AstraZeneca.

10.1021/ci0255937 CCC: $25.00
Published on Web 02/12/2003

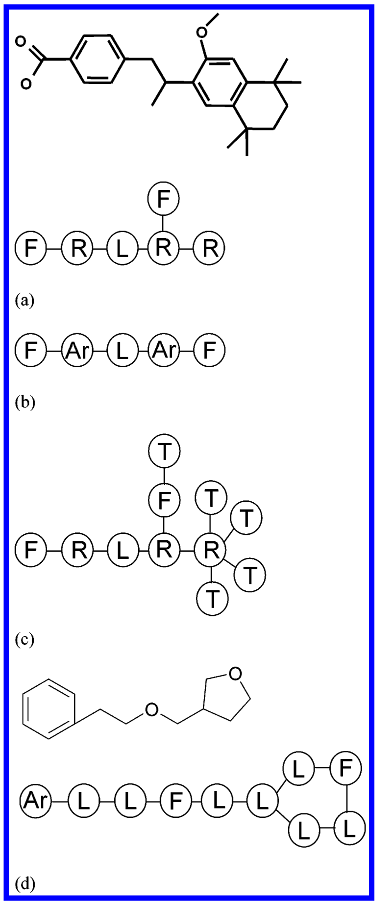


Figure 1. (a). An R/F reduced graph; (b) An Ar/F reduced graph; (c) An R/F/T reduced graph; and (d) an Ar/F/L reduced graph.

different fingerprints that are used to represent the reduced graphs for similarity searching.

Node Definitions. R/F Reduced Graphs. These reduced graphs were introduced in the previous paper,[12] and they provide a baseline against which to measure the performance of new descriptors. The nodes in a R/F reduced graph consist of Ring Nodes; Feature Nodes; and Link Nodes. A Ring Node represents a single ring as defined by the smallest set of smallest rings identified using the Daylight toolkit.[13] Terminal non-hydrogen-bonding atoms are removed from the chemical graph and Link Nodes are created from connected isolating carbons, where an isolating carbon is a carbon atom that is not doubly or triply bonded to heteroatoms.[13] Feature Nodes are created from the acyclic fragments that remain once the isolated carbons have been removed. An example R/F reduced graph is shown in Figure 1(a). In the previous work, the nodes in the reduced graph are further characterized at different levels of discrimination. Here we use levels 3 and 4. At level 3, Ring Nodes are characterized as aromatic or aliphatic and as hydrogen-bonding or non-hydrogen-bonding. No additional characterization is used for Link Nodes or Feature Nodes (which, by definition, are already hydrogen-bonding). An R/F reduced graph at level 3 is referred to as R/F(3). At level 4, hydrogen-bonding nodes are further characterized as donors, acceptors, or donors and acceptors. A reduced graph at this level of description is referred to as R/F(4).

Ar/F Reduced Graphs. These reduced graphs were also introduced in the previous paper.[12] The chemical graph is partitioned into Aromatic Ring Nodes; Feature Nodes; and Link Nodes with aliphatic rings being treated in the same way as acyclic parts of the graph, i.e., the chemical graph is preprocessed to first break aliphatic rings open by removing two-connected non-hydrogen-bonding atoms and then second to recursively remove terminal non-hydrogen-bonding atoms. The steps involved in identifying the nodes then proceed as before. Thus, Ring Nodes are now limited to aromatic rings and aliphatic rings become incorporated into Feature Nodes, Link Nodes, or they can be removed entirely if they are terminal and do not contain any hydrogen-bonding atoms. An example Ar/F reduced graph is shown in Figure 1(b). An Ar/F reduced graph at level 4 is referred to as Ar/F(4).

Table 1. Ring Systems Identified in the RS/F Reduced Graphs

number of rings	ring sizes
1	≠6
1	6
2	≤56
2	>56
3	≤566
3	>566
4	any
>5	any

R/F/T Reduced Graphs. The removal of terminal non-hydrogen bonding atoms, as done in the previous two reduced graph types, may result in the loss of hydrophobic features such as methyl groups that are often found in active drugs.[14] Therefore, the R/F/T reduced graphs were developed. These are similar to R/F reduced graphs; however, the terminal groups are retained in the chemical graph and are treated as Link Nodes. An example R/T/F reduced graph is shown in Figure 1(c).

RS/F Reduced Graphs. These reduced graphs are the same as the R/F reduced graphs; however, additional information on the sizes of the ring systems included within the chemical graph is appended to the various fingerprint representations, described below. An analysis of the ring systems contained within the ID ALERT database[15] was carried out to determine commonly occurring ring systems, and the presence of eight different sized ring systems is recorded, as shown in Table 1. The ring systems are further differentiated according to whether they contain hydrogen-bonding atoms. Thus an additional 16 bits are appended to the fingerprints: eight bits for the ring systems shown in Table 1 that do not contain hydrogen-bonding atoms and eight bits to represent the same sized ring systems that contain hydrogen bonding atoms.

Ar/F/L Reduced Graphs. Here aromatic rings are reduced to aromatic ring nodes, Feature nodes are identified as before, but now every isolating carbon that occurs in an aliphatic ring or an acyclic chain is treated as an individual Link Node. An example is shown in Figure 1(d). The result is a much less severe form of graph reduction. Aromatic and Feature nodes are further differentiated according to hydrogen bonding character.

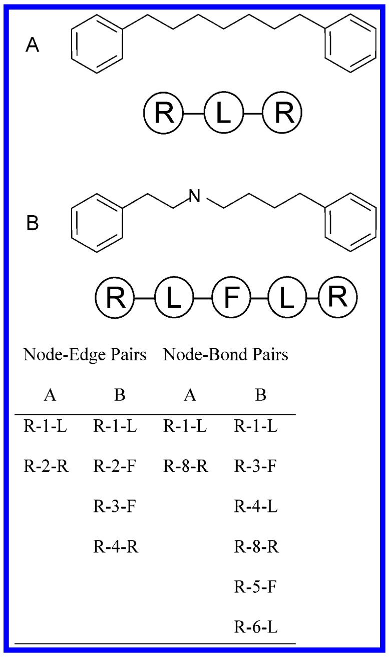
Fingerprint Representations. Hashed Fingerprints. In the previous work, the reduced graphs were translated into pseudo-SMILES representations by mapping each node type to a different heavy atom, and the pseudo-SMILES were used to generate hashed Daylight fingerprints. This approach provided a convenient, quick-to-implement, way of testing the general effectiveness of the reduced graph. This representation, however, is unlikely to be ideal for encoding reduced graphs since the characteristics of reduced graphs are different from chemical graphs. For example, reduced graphs consist of fewer nodes than the chemical graphs from which they are derived and hence their fingerprints are much sparser than those generated directly from the chemical graph. Furthermore, the hashing procedure used to generate Daylight fingerprints results in a one-to-many correspondence between a subgraph of the reduced graph and bits in the fingerprint and collisions may occur where the bits corresponding to one subgraph overlap with the bits corresponding to another. Thus, it is not possible to map from a bit in the fingerprint back 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.

Node-Edge Pairs. In this representation, the fingerprint represents node-pairs in the reduced graph (cf., atom-pairs as described by Cahart et al.[16]). A node-pair is defined as *node type* − *distance* − *node type* where distance is defined as the number of edges on the shortest path between the two nodes in the node-pair. The fingerprint is designed so that each bit represents a different node-pair, that is, there is a one-to-one correspondence between a bit in the fingerprint and a node-pair.

The hashed fingerprint method described above encodes three distinct categories of hydrogen bonding nodes: *acceptor*; *donor*; and *donor and acceptor*. The result is that a node characterized as *donor and acceptor* only matches nodes with exactly the same specification and will not match either a *donor* node or an *acceptor* node. Using the node-pair representation, however, a *donor and acceptor* node can be encoded by setting the corresponding *donor* bit to "on" and also setting the *acceptor* bit to "on". Thus, a *donor* node will set a subset of the bits set by a *donor and acceptor* node. This procedure is used in all fingerprint representations described hereon.

H-Node-Edge Pairs. A hologram version of the Node-Edge Pairs in which the number of times a node-pair occurs is recorded, rather than just its presence or absence.

Node-Bond Pairs. This representation is similar to the Node-Edge Pairs except that the distance between nodes is given by the number of **bonds** on the shortest path between the two closest atoms in the original chemical graph, where the atoms occur in different nodes. With Node-Edge Pairs, it is possible that small changes in chemical structure can result in a very different pattern of bits being set. Consider two rings separated by a long chain of carbon atoms as shown in structure A in Figure 2. The reduced graph representation consists of three nodes: R−L-R (a Ring Node connected to a Link Node connected to further a Ring Node). Replacing one of the carbon atoms with a hydrogen-bonding atom, structure B, will result in the reduced graph R-L-F-L-R (i.e., the acyclic part of the chemical graph now becomes two Link Nodes split by a Feature Node). Using the Node-Edge Pairs, the two rings are separated by two edges in structure A and four edges in structure B, thus different bits are set in the bit string representations. When distance is measured by the number of bonds in the original graph then, in both cases, the rings are separated by eight bonds and thus the same bit is set "on".



Node-Edge Pairs		Node-Bond Pairs	
A	B	A	B
R-1-L	R-1-L	R-1-L	R-1-L
R-2-R	R-2-F	R-8-R	R-3-F
	R-3-F		R-4-L
	R-4-R		R-8-R
			R-5-F
			R-6-L

Figure 2. Node-pairs resulting from the Node-Edge Pairs and the Node-Bond Pairs.

H-Node-Bond Pairs. A hologram version of the Node-Bond Pairs.

Node-Edge Paths. Reduced graphs are represented by paths of up to length three, i.e., singleton reduced graph nodes, neighboring pairs of nodes, and linear paths of three nodes. A dictionary is used to map all possible paths to the fingerprint so that each subgraph corresponds to a single bit, rather than the multiple bits used in the hashed approach described above. Thus, a one-to-one correspondence exists between the bits and the subgraphs of the reduced graph. One consequence of using a dictionary is that a large number of bits is required to represent all possible paths, hence the upper limit on the path length is three (compared with the default of seven for the hashed method). However, given the small number of nodes in a reduced graph relative to a chemical graph and the fact that each node typically represents several atoms, a sizable portion of a reduced graph can be represented by three connected nodes.

H-Node-Edge Paths. A hologram version of the above.

The combination of the seven different fingerprint representations and the different reduced graph types and node specifications results in a large number of possibilities to investigate. Table 2 summarizes the different combinations that are used in the following experiments and the names given to those combinations for later reference. The number in brackets refers to the level of node characterization used in the reduced graph as discussed earlier.

Similarity Searches. Eleven activity classes were extracted from the ID Alert database,[15] and a random subset of 100 compounds was selected for each class. The activity classes are angiotensin II receptor antagonists (ANG), calcium channel antagonists (CAL), HIV protease inhibitors (HIV), HMG coenzyme A reductase inhibitors (HMG), 5-lipoxygenase inhibitors (LIP), NMDA antagonists (NMDA), platelet activating factor antagonists (PAF), potassium channel inhibitors (POT), renin inhibitors (REN), reverse transcriptase inhibitors (REV), and squalene synthase inhibitors (SQUA). A random sample of 2000 of the remaining ID

Table 2. Summary of the Different Reduced Graph Types and Representation Methods Used

	reduced graph type					
representation	R/F(4)	Ar/F(4)	R/F(3)	R/F/T(4)	RS/F(4)	Ar/F/L(4)
hashed	RG1	RG2				
node-edge pairs	RG3	RG4		RG5	RG6	RG7
H-node-edge pairs	RG8				RG9	
node-bond pairs	RG10		RG11	RG12		
H-node-bond pairs	RG13					
node-edge paths	RG14					
H-node-edge paths	RG15					

Table 3. Results Found Using the Original R/F(4) and Ar/F(4) Reduced Graphs[a]

RG	ANG	CAL	HIV	HMG	LIP	NMDA	PAF	POT	REN	REV	SQUA
RG1	*5.67*	2.48	*2.98*	*2.49*	*2.21*	1.70	*1.79*	*3.37*	*5.48*	1.87	*1.67*
	(2.30)	(1.12)	*(1.22)*	*(1.13)*	*(0.60)*	(0.76)	*(0.81)*	*(1.57)*	*(2.47)*	(0.93)	*(0.60)*
RG2	4.77	*2.6*	2.75	2.17	2.09	*2.09*	1.77	2.88	4.14	*2.38*	1.48
	(2.41)	*(1.16)*	(1.31)	(0.90)	(0.87)	*(0.83)*	(0.84)	(1.30)	(2.18)	*(1.13)*	(0.51)

[a] The best performing method for each activity class is italicized.

Alert 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 class there may be false negatives in the inactive sample.

Similarity searches were carried out using each of the reduced graphs identified in Table 2 and conventional Daylight fingerprints calculated from the chemical graphs. For each data set, each active compound was used in turn as a target against which the similarity of the remaining structures was measured. Similarity was measured using the binary version of Tanimoto coefficient for the binary fingerprints and the set-theoretic version for the hologram fingerprints.[2] An enrichment factor, *EF*, was calculated as the number of actives, n_a, in the top 10% of the ranked hit list divided by the total number of actives, N_a, that would be expected in this decile if the actives were distributed at random.

$$EF = \frac{n_a}{N_a \times 0.1}$$

The enrichment factors were then averaged over all targets in a given activity class.

Data Fusion. Data fusion is a technique that is used to combine data from different sources in the hope of improving on results found using just one data source. The technique has been shown to be effective in combining the hit lists generated from different similarity searching methods with improved results over using a single method.[17−20] The technique is similar to consensus scoring which has been used to combine results from different docking methods.[21] Here, data fusion is used to combine the results from the different types of reduced graphs and Daylight fingerprints in an attempt to improve on results found using a single descriptor. The hit lists are combined using the SUM method of Ginn et al.[18] which is described here for the fusion of two hit lists. First, the rank position of a molecule in hit list 1, m_1, is summed with the rank position of the same molecule in hit list 2, m_2, to give a summed rank, m_{1+2}. This is repeated for each molecule. The summed ranks are then used to reorder the structures to give a new fused hit list. Finally, an enrichment factor is derived for the fused hit list. The method can be extended to any number of hit lists.

Recursive Partitioning. Recursive partitioning (RP) is a statistical technique that can be used to find patterns in large data sets by building what is known as a decision tree.[22] A data set is recursively divided into subsets depending on the value of a splitting criterion, such as the presence or absence of a feature. The feature is chosen so that the subsets are as diverse as possible according to some criterion such as biological response. RP has been used with binary 2D and 3D fingerprints to develop SAR rules from initial screening results by identifying the descriptors which lead to an active leaf node.[23,24] The SAR can then be used to select additional compounds for subsequent rounds of screening. A problem with using 2D fingerprints as descriptors is that each feature typically represents a very small substructure, hence Cho et al. developed the BFIRM (Binary Formal Inference-Based Recursive Modeling) approach[25] that uses multiple descriptors to decide on a split. More useful structural information is obtained since each split is due to a larger common substructure than when using a single split. A related approach has also been developed by Blower et al.[26] Here, RP is used with reduced graphs where the features used to decide on splits are Node-Edge Pairs, and thus each split can represent a large substructural fragment. We have used the CART program[27] which uses a modified *t*-test to decide on the splits in the decision tree.

RESULTS AND DISCUSSION

The results of the similarity searches performed on the ID Alert data sets are shown in Tables 3−8.

Table 3 presents the results for the two reduced graph types and the fingerprint representation described in the earlier work, that is, R/F(4) and Ar/F(4) using hashed fingerprints, represented by RG1 and RG2, respectively. The reduced graph that is most effective at identifying active compounds is italicized for each activity class. The results show that the R/F(4) reduced graph is more effective at identifying active compounds in eight of the 11 activity classes; however, it is outperformed by the Ar/F(4) reduced graph in the remaining three data sets. The fact that neither method consistently outperforms the other is in keeping with previous results.[12]

Table 4. Average Enrichment Factors, with Standard Deviations in Brackets, for Reduced Graph Type (R/F(4)) Represented by a Variety of Different Fingerprints[a]

RG	ANG	CAL	HIV	HMG	LIP	NMDA	PAF	POT	REN	REV	SQUA
RG1	5.67 (2.30)	2.48 (1.12)	2.98 (1.22)	2.49 (1.13)	2.21 (0.60)	1.70 (0.76)	1.79 (0.81)	*3.37* *(1.57)*	5.48 (2.47)	1.87 (0.93)	1.67 (0.60)
RG3	5.09 (2.41)	2.62 (1.06)	3.00 (0.95)	2.76 (1.06)	1.88 (0.76)	1.82 (0.71)	1.65 (0.84)	2.83 (1.22)	5.17 (2.49)	1.67 (0.7)	*2.08* *(0.63)*
RG8	5.66 (1.94)	2.25 (0.95)	2.7 (1.48)	2.42 (1.13)	2.43 (0.81)	*1.99* *(0.73)*	1.71 (0.74)	2.83 (1.06)	4.12 (2.48)	1.85 (0.59)	1.46 (0.47)
RG10	5.18 (2.48)	2.58 (1.12)	*3.25* *(1.15)*	*2.83* *(1.01)*	1.71 (0.64)	1.81 (0.68)	1.72 (0.91)	2.92 (1.3)	5.26 (2.54)	1.82 (0.7)	2.07 (0.7)
RG13	5.53 (2.16)	2.34 (1.04)	2.65 (1.41)	2.64 (1.14)	2.32 (0.81)	1.88 (0.72)	*1.83* *(0.85)*	2.74 (0.88)	4.5 (2.55)	1.88 (0.67)	1.6 (0.58)
RG14	5.6 (2.21)	*2.84* *(1.34)*	3.12 (0.94)	2.4 (1.09)	1.91 (0.54)	1.71 (0.71)	1.69 (0.86)	2.88 (1.46)	*5.59* *(2.33)*	1.86 (0.89)	1.94 (0.66)
RG15	*6.18* *(2.11)*	2.31 (1.15)	2.77 (1.51)	2.65 (1.21)	*2.5* *(0.79)*	1.7 (0.69)	1.71 (0.7)	2.82 (1.33)	4.69 (2.51)	*1.95* *(0.82)*	1.63 (0.65)

[a] See Table 2 and text for details.

Table 5. Average Enrichment Factors, with Standard Deviations in Brackets, of Different Reduced Graph Types Using the Same Fingerprint Representation[a]

RG	ANG	CAL	HIV	HMG	LIP	NMDA	PAF	POT	REN	REV	SQUA
RG3	*5.09* *(2.41)*	2.62 (1.06)	3.00 (0.95)	2.76 (1.06)	1.88 (0.76)	1.82 (0.71)	1.65 (0.84)	2.83 (1.22)	*5.17* *(2.49)*	1.67 (0.7)	2.08 (0.63)
RG4	4.45 (2.41)	2.35 (0.97)	2.83 (1.09)	2.6 (1.03)	*1.92* *(0.9)*	1.81 (0.72)	*1.93* *(1.05)*	2.97 (1.16)	4.17 (2.12)	1.54 (0.66)	1.69 (0.6)
RG5	4.91 (2.28)	*3.00* *(1.28)*	2.76 (0.86)	*2.99* *(0.98)*	1.84 (0.65)	1.79 (0.72)	1.68 (0.84)	*3.07* *(1.24)*	4.99 (2.42)	1.64 (0.66)	*2.13* *(0.6)*
RG6	5.12 (2.36)	2.59 (1.08)	3 (0.97)	2.87 (1.1)	*1.92* *(0.73)*	1.84 (0.71)	1.79 (0.76)	3.03 (1.14)	4.1 (2.47)	*2.04* *(0.62)*	1.49 (0.48)
RG7	4.13 (2.52)	2.85 (1.26)	*3.62* *(1.22)*	2.79 (1.22)	1.62 (0.73)	*1.9* *(0.81)*	1.87 (1.22)	2.63 (1.14)	4.2 (2.23)	1.97 (0.95)	1.96 (0.87)

[a] See Table 2 and text for details.

Table 6. Average Enrichment Factors, with Standard Deviations in Brackets, for Hologram Representations of Different Reduced Graph Types[a]

RG	ANG	CAL	HIV	HMG	LIP	NMDA	PAF	POT	REN	REV	SQUA
RG8	*5.66* *(1.94)*	2.25 (0.95)	*2.7* *(1.48)*	2.42 (1.13)	2.43 (0.81)	*1.99* *(0.73)*	1.71 (0.74)	2.83 (1.06)	*4.12* *(2.48)*	1.85 (0.59)	1.46 (0.47)
RG9	5.19 (1.86)	*2.29* *(0.97)*	2.68 (1.5)	*2.49* *(1.14)*	*2.45* *(0.75)*	1.94 (0.68)	*1.79* *(0.76)*	*3.03* *(1.14)*	4.1 (2.47)	*2.04* *(0.62)*	*1.49* *(0.48)*

[a] See Table 2 and text for details.

Table 4 presents the results for reduced graph R/F(4) represented by the different fingerprint encoding schemes and thus allows the effect of changing the fingerprint representation to be examined. The original hashed method, RG1, is top scoring for only one of the activity classes (POT), and the alternative representations outperform this in 10 out of the 11 data sets. However, no one fingerprint method stands out as offering the best performance across all the data sets. In fact, each of the fingerprint representation methods comes top for at least one of the data sets.

Table 5 presents the results for one fingerprint representation (Node-Edge Pairs) across a series of different reduced graph types. Again, the relative performance of the different reduced graphs varies across the different data sets.

Table 6 compares two different reduced graph types represented as H-Node-Edge Pairs, and Table 7 represents three different reduced graph types encoded as Node-Bond Pairs. Again no clear pattern emerges as to which reduced graph type and which representation method are best.

Table 8 compares the best reduced graph result for a given data set with the enrichment factor found using Daylight fingerprints. Here, it can be seen that the reduced graphs outperform Daylight fingerprints for six data sets, the performance is the same for one data set, and Daylight outperforms all of the reduced graphs for four of the data sets.

Previous experiments with reduced graphs showed that they were complementary to Daylight fingerprints, that is, they identify different actives to those found using Daylight fingerprints.[12] This suggests that it may be possible to improve on the performance of Daylight fingerprints by fusing the results found for different descriptors. Wang and Wang[28] have suggested that the performance enhancements that might be expected from data fusion level off after the inclusion of three or four separate rankings, thus, rather than test all combinations of descriptors that could be selected from the 15 reduced graphs, the descriptors were first clustered and a representative subset chosen.

Clustering the Reduced Graphs. The reduced graphs were clustered according to the similarity of the hit lists generated, that is, reduced graphs that rank a set of compounds in similar ways cluster together. Twenty target

Table 7. Average Enrichment Factors, with Standard Deviations in Brackets, for Node-Bond Pair Representations of Three Different Reduced Graph Types[a]

RG	ANG	CAL	HIV	HMG	LIP	NMDA	PAF	POT	REN	REV	SQUA
RG10	*5.18*	2.58	*3.25*	*2.83*	1.71	1.81	1.72	2.92	*5.26*	1.82	*2.07*
	(2.48)	(1.12)	*(1.15)*	*(1.01)*	(0.64)	(0.68)	(0.91)	(1.3)	*(2.54)*	(0.7)	*(0.7)*
RG11	4.45	2.19	2.83	2.40	*1.85*	1.73	*1.85*	2.91	4.09	1.54	1.64
	(2.41)	(1.10)	(1.10)	(1.19)	*(0.95)*	(0.80)	*(1.09)*	(1.21)	(2.18)	(0.65)	(0.65)
RG12	5.03	*3.01*	3.13	2.78	1.77	*1.91*	1.79	*3.2*	5.11	*1.88*	1.88
	(2.38)	*(1.32)*	(1.12)	(0.99)	(0.59)	*(0.69)*	(0.94)	*(1.36)*	(2.5)	*(0.69)*	(0.63)

[a] See Table 2 and text for details.

Table 8. Average Enrichment Factors, with Standard Deviations in Brackets, for the Best Performing Reduced Graph in Each Activity Class Is Compared with the Result Found Using Daylight Fingerprints

	ANG	CAL	HIV	HMG	LIP	NMDA	PAF	POT	REN	REV	SQUA
best RG	RG15	*RG12*	RG7	RG5	*RG15*	*RG2*	*RG4*	*RG1*	RG14	*RG2*	*RG5*
	6.18	*3.01*	3.62	2.99	*2.5*	*2.09*	*1.93*	*3.37*	5.59	*2.38*	*2.13*
	(2.11)	*(1.32)*	(1.22)	(0.98)	*(0.79)*	*(0.83)*	*(1.05)*	*(1.57)*	(2.33)	*(1.13)*	*(0.6)*
D	*6.49*	2.33	*4.66*	*3.56*	1.96	1.84	*1.93*	2.63	*5.84*	1.81	1.81
	(2.15)	(1.18)	*(1.25)*	*(1.33)*	(0.79)	(0.72)	*(0.85)*	(1.43)	*(1.95)*	(0.79)	(0.62)

Table 9. Descriptors Are Clustered into Six Distinct Clusters

cluster	descriptors
1	RG8, RG9, RG13, RG15
2	DL
3	RG7
4	RG2, RG4
5	RG11
6	RG1, RG3, RG5, RG6, RG10, RG12, RG14

compounds were selected at random from ID ALERT (regardless of activity class). Each target was used in turn in a similarity search against the rest of the database (∼11500 compounds) for each of the reduced graph types and for Daylight fingerprints and the top scoring 100 compounds was found for each descriptor. The similarity between two different descriptors was measured as the (normalized) number of compounds in common between the two hit lists using the following equation

$$S_{ij} = \frac{c_{ij}}{100}$$

where S_{ij} is the similarity between descriptor i and descriptor j and c_{ij} is the number of compounds common to the hit lists generated using i and j, respectively.

A pairwise similarity matrix was generated for all descriptors, and the descriptors were clustered based on the similarity matrices. This gave rise to one clustering per target. This process was repeated for each target, and a second level of clustering was performed in which the cluster memberships were compared across the different targets. The reduced graphs were reclustered on the basis of how often they clustered together for each target. This process was repeated using hit lists of size 100 and 400 and using three different clustering methods (single linkage, complete linkage, and group-average). The methodology is essentially the same as that used by Holliday et al.[19] in a study of the comparison of 22 similarity coefficients. Full details of our implementation are given in Barker.[29]

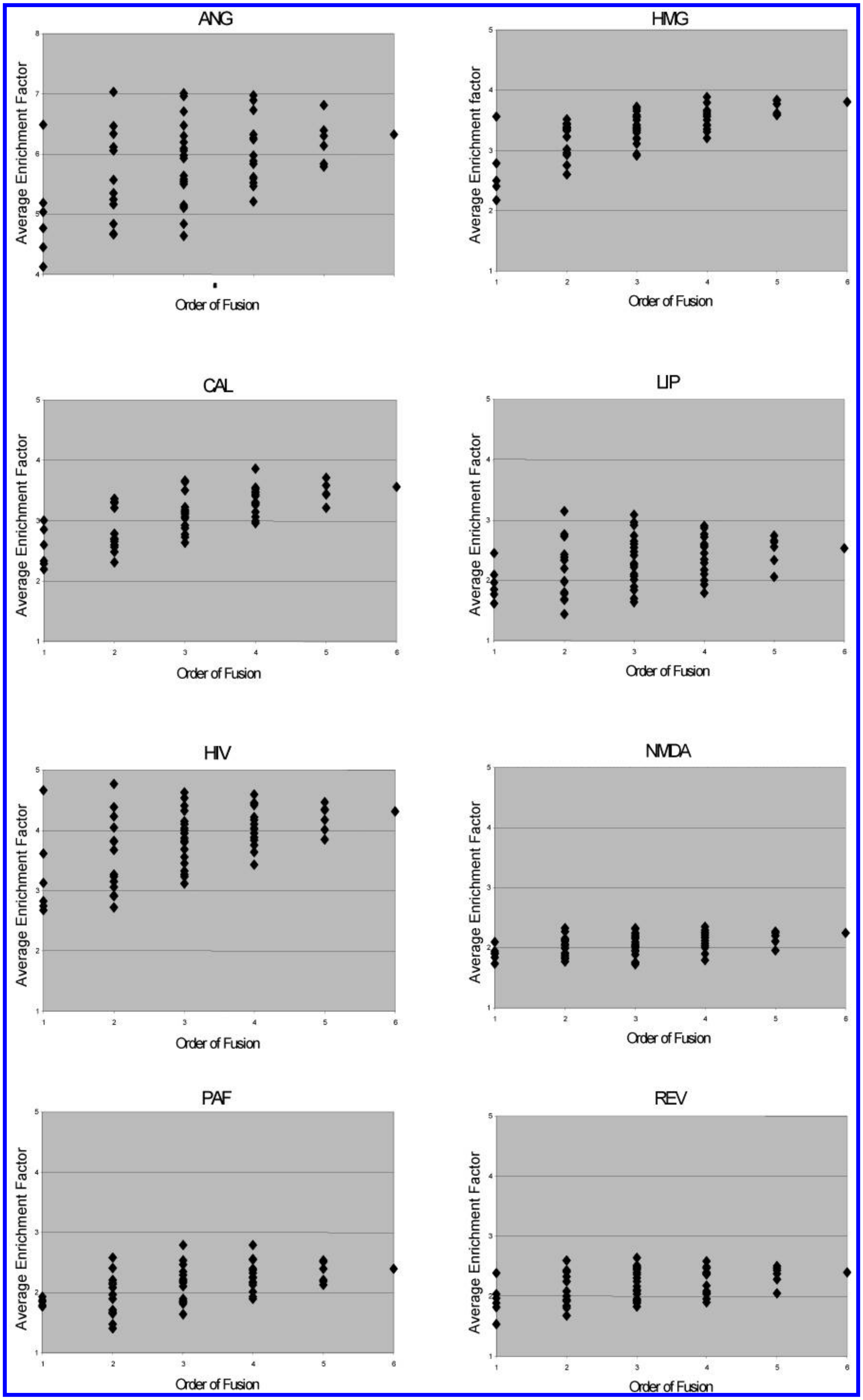
The resulting clusters are shown in Table 9 and are in general agreement with intuition. For example, one cluster corresponds to reduced graphs represented as holograms; the two Ar/F reduced graphs cluster together; and Daylight fingerprints form a singleton as do the Ar/F/L reduced graphs. The representative subset chosen for the data fusion experiments is shown in Table 10.

Table 10. One Descriptor Is Selected from Each Cluster To Form a Representative Subset of Descriptors

descriptors
RG9
DL
RG7
RG2
RG11
RG12

Data Fusion. Data fusion experiments were carried out using the five representative reduced graphs together with Daylight fingerprints. The graphs in Figure 3 show the results. The first column in each plot shows the enrichment factors found for the individual descriptors. The enrichment factors are also given in Table 11. In most cases the range of enrichments found varies considerably. In five of the data sets (ANG, HIV, HMG, PAF, and REN) the Daylight fingerprints score highest. In the other six data sets it is one of the reduced graph methods that scores highest, with the particular reduced graph varying from one data set to another. The second column in the plots in Figure 3 shows the results for all combinations of two descriptors. The third column shows the results for all combinations of three descriptors, and so on, up to all six descriptors.

It can be seen that in most cases the results improve when considering up to four descriptors, and the performance then tends to tail off as five and six descriptors are fused. This is consistent with Wang and Wang.[28] The results for fusing all combinations of four descriptors are shown in Table 12 with the best result for each data set italicized. Again it can be seen that different combinations work best for different data sets. For two of the data sets (SQUA and CAL) the best combinations do not include Daylight fingerprints. Two combinations each work best for three of the data sets. These combinations are D-RG2-RG9-RG12 and D-RG7-RG9-RG12. The percentage gains for each of these combinations over using Daylight fingerprints are given in Table 13. An enrichment factor of 6.98 is achieved for the ANG data set which means that on average 69.8 actives are found in the


ANG
Average Enrichment Factor
4
5
6
7
8
Order of Fusion
HMG
Average Enrichment factor
1
2
3
4
5
1
2
3
4
5
6
Order of Fusion
CAL
Average Enrichment Factor
1
2
3
4
5
1
2
3
4
5
6
Order of Fusion
LIP
Average Enrichment Factor
1
2
3
4
5
1
2
3
4
5
6
Order of Fusion
HIV
Average Enrichment Factor
1
2
3
4
5
1
2
3
4
5
6
Order of Fusion
NMDA
Average Enrichment Factor
1
2
3
4
5
1
2
3
4
5
6
Order of Fusion
PAF
Average Enrichment Factor
1
2
3
4
5
1
2
3
4
5
6
Order of Fusion
REV
Average Enrichment Factor
1
2
3
4
5
1
2
3
4
5
6
Order of Fusion

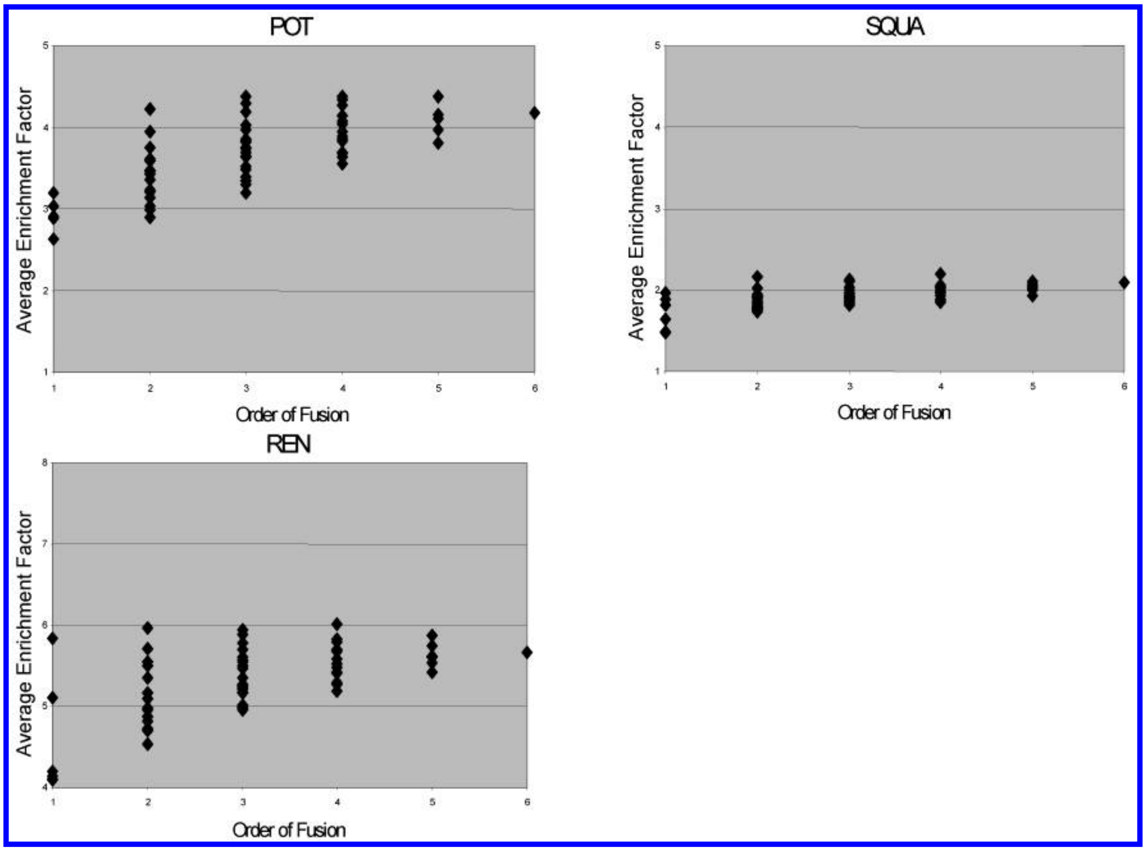


Figure 3. Average Enrichment Factors are plotted for all combinations from one to up to six fused descriptors.

Table 11. Average Enrichment Factors Are Shown for Each of the Representative Descriptors

	ANG	CAL	HIV	HMG	LIP	NMDA	PAF	POT	REN	REV	SQUA
D	*6.49*	2.33	*4.66*	*3.56*	1.96	1.84	*1.93*	2.63	*5.84*	1.81	1.81
RG2	4.77	2.6	2.75	2.17	2.09	*2.09*	1.77	2.88	4.14	*2.38*	1.48
RG7	4.13	2.85	3.62	2.79	1.62	1.9	1.87	2.63	4.2	1.97	*1.96*
RG9	5.19	2.29	2.68	2.49	*2.45*	1.94	1.79	3.03	4.1	2.04	1.49
RG11	4.45	2.19	2.83	2.40	1.85	1.73	1.85	2.91	4.09	1.54	1.64
RG12	5.03	*3.01*	3.13	2.78	1.77	1.91	1.79	*3.2*	5.11	1.88	1.88

Table 12. Average Enrichment Factors for All 15 Combinations of Four Hit Lists of the Six Descriptors for All 11 ID Alert Data Sets[a]

Comb.	ANG	CAL	HIV	HMG	LIP	NMDA	PAF	POT	REN	REV	SQUA
D-2−7−9	6.33	3.47	4.45	3.67	2.72	2.21	2.56	4.07	5.79	2.4	1.93
D-2−7−11	5.62	2.99	4.22	3.51	2	2	2.16	3.68	5.42	1.95	1.86
D-2−7−12	5.88	3.31	4.42	3.58	2.1	1.9	2.32	3.69	5.7	2.04	1.88
D-2−9−11	6.9	3.3	4.02	3.42	2.87	2.25	2.55	*4.37*	5.48	2.38	1.98
D-2−9−12	*6.98*	3.55	4.17	3.58	*2.9*	2.26	*2.79*	4.34	5.82	2.49	2.03
D-2−11−12	6.32	3.06	3.88	3.35	2.17	2.04	2.18	3.95	5.48	2.06	1.85
D-7−9−11	5.98	3.15	4.44	3.8	2.34	2.29	2.17	3.89	5.68	2.49	1.96
D-7−9−12	6.25	3.53	*4.59*	*3.89*	2.45	2.25	2.39	3.85	*6.01*	2.46	2.02
D-7−11−12	5.46	2.96	4.1	3.62	1.79	2.04	1.93	3.55	5.58	2.08	1.99
D-9−11−12	6.73	3.26	4.04	3.62	2.55	*2.35*	2.24	4.15	5.67	*2.58*	2.05
2−7−9−11	5.59	3.48	3.84	3.41	2.58	2.12	2.13	4.04	5.27	2.18	2.04
2−7−9−12	5.84	*3.86*	3.95	3.56	2.6	2.07	2.37	4.04	5.52	2.36	2.04
2−7−11−12	5.21	3.28	3.64	3.31	1.93	1.79	1.9	3.64	5.19	1.89	2.02
2−9−11−12	6.27	3.41	3.43	3.2	2.76	2.18	2.26	4.27	5.29	2.39	1.98
7−9−11−12	5.52	3.43	3.76	3.66	2.29	2.16	2.01	3.83	5.41	2.37	*2.19*

[a] The best performing combination is italicized.

top 210 positions of the ranked list compared with 65 found using Daylight alone (this represents an increase of 7.5% when using the fused descriptors); for CAL, on average 35 actives are found compared with 23 using Daylight alone (representing an increase of 52%); and for POT on average 43 actives are found compared with 26 found using Daylight alone (representing an increase of 65%). The combined descriptors give worse performance than Daylight alone for one of the data sets, HIV, where on average 41 actives are found instead of 47.

Thus, in 10 out of the 11 cases the combined result is better than the best individual descriptor with up to 65% more actives being retrieved. Thus, if it is not known which descriptor is the best indicator of activity for a data set (which

Table 13. Average Enrichment Factors Are Shown for Daylight Fingerprints and the Two Combinations of D-RG2-RG9-RG12 and D-RG7-RG9-RG12[a]

Comb.	ANG	CAL	HIV	HMG	LIP	NMDA	PAF	POT	REN	REV	SQUA
D	6.49	2.33	4.66	3.56	1.96	1.84	1.93	2.63	5.84	1.81	1.81
D-2−9−12	6.98	3.55	4.17	3.58	2.9	2.26	2.79	4.34	5.82	2.49	2.03
%	7.5	52.4	−10.5	0.56	47.9	22.83	44.6	65.0	−0.3	37.6	10.8
D-7−9−12	6.25	3.53	4.59	3.89	2.45	2.25	2.39	3.85	6.01	2.46	2.02
%	−3.7	51.5	−1.5	9.3	25.0	22.3	23.8	46.4	2.9	35.9	11.6

[a] The percentage gains are shown for the fused descriptors relative to using Daylight fingerprints alone. It can be seen that the fused descriptors are more effective in all but one case.

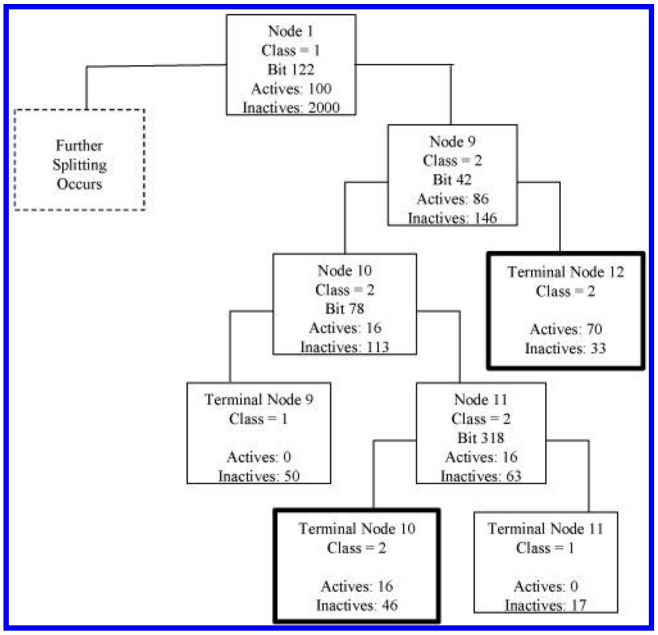


Figure 4. The decision tree grown using the ANG training set, terminal nodes classified as active are shown in bold.

is usually the case), then it appears that the most effective approach is to use a combination of descriptors and fuse the results.

Recursive Partitioning. Recursive partitioning was used to construct an SAR model for angiotensin II receptor antagonists based on reduced graph RG12. This reduced graph is encoded using the Node-Bond Pairs fingerprint where each bit in the fingerprint corresponds to the presence or absence of a particular Node-Bond Pair. The fingerprints are used to determine the splitting criteria in the decision tree and, since each bit is readily interpretable, the aim was to see if the combination of recursive partitioning and reduced graphs can be used to build a useful model of activity. The ANG activity class was chosen because a well-defined two-dimensional SAR model is available in the literature.[30,31]

The 100 actives and 2000 inactives were used to train the decision tree using 10-fold cross-validation. Figure 4 shows the resulting decision tree. It can be seen that Terminal Node 12 is highly enriched with respect to active compounds since it contains 70 (70%) of the actives together with only 33 (4.95%) of the inactives. Two Node-Bond Pair descriptors have been applied as splitting criteria to lead to this node. These are shown in Figure 5(a).

By combining these two features it is possible to suggest the substructure shown in Figure 5(b) as a model for activity. This substructure compares well with the 2D SAR model for activity of the more common angiotensin II receptor antagonists described by Bradbury et al.[31] and shown in Figure 6.

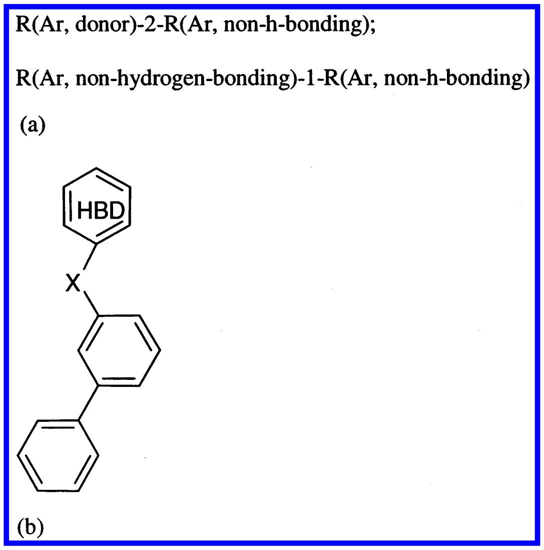

R(Ar, donor)-2-R(Ar, non-h-bonding);
R(Ar, non-hydrogen-bonding)-1-R(Ar, non-h-bonding)
(a)
HBD
X
(b)

Figure 5. (a) The two Node-Bond Pairs used as splitting criteria leading to the richest active node in the decision tree. The first Node-Bond Pair represents an aromatic heterocycle with hydrogen bond donor character separated by two bonds from a non-hydrogen-bonding aromatic ring. The second Node-Bond Pair represents two aromatic non-hydrogen-bonding rings separated by one bond. (b) A possible substructure formed by combining the two Node-Bond Pairs where X is any atom.

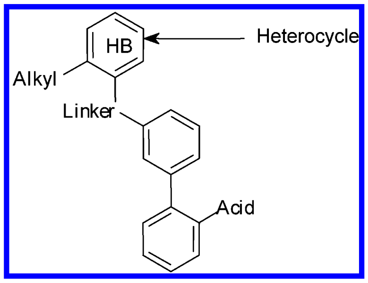


Figure 6. 2D SAR model for angiotensin II receptor antagonists (2-alkyl-4-(biphenylmethoxy)quinoline derivatives).

This experiment demonstrates the potential of using reduced graphs together with recursive partitioning in order to build SAR models. Once a suitable decision tree has been constructed it can be used to select compounds for screening. A preliminary experiment has been carried to test the validity of this approach as described below.

A decision tree was constructed from a data set consisting of 1580 AstraZeneca inhouse compounds of which 399 had shown activity in a kinase inhibition assay. The fingerprints used to construct the decision tree were type RG3. An in-house database of over half a million structures was then “dropped through” the decision tree. The compounds that fell into active nodes were extracted and further filtered by performing similarity searches against the 399 active targets. Compounds with a Tanimoto coefficient of 0.8 or greater against any of the active targets were retained (where the Tanimoto coefficient was calculated using the fingerprints generated from the reduced graphs). The resulting 2547 structures were then clustered using Daylight fingerprints and Jarvis-Patrick clustering, and 357 compounds (singletons and cluster centroids) were chosen for screening against the kinase assay.

The results of the assay revealed that 26 of the 357 compounds showed significant kinase inhibition activity. This represents a hit rate of 7.3% and is a significant improvement on the previous hit rates of 2.4% found in this assay using other compound selection methods.

CONCLUSIONS

This paper has focused on developing different types of reduced graphs and different fingerprint representation methods for use in similarity searching. Similarity searches have been performed over 11 different data sets. Improved results were found over earlier methods;[12] however, no consistent pattern has emerged as to which type of reduced graph and representation method is best overall. The reduced graphs were also compared with Daylight fingerprints: in some cases Daylight fingerprints gave the best performance and in others one of the reduced graph methods came top. Consequently, a representative set of reduced graphs was chosen and used together with Daylight fingerprints in data fusion experiments. Results are very encouraging with improved performance being found in 10 out of 11 data sets compared to using Daylight fingerprints alone. In some cases performance was improved by as much as 65%.

Reduced graphs have been shown to be effective in generating a model of activity for the angiotensin data set that is consistent with a published model. They offer several advantages over using Daylight fingerprints for this application, such as the following: the generalized nature of reduced graphs mean that typically fewer, more general, SAR models are generated; when using node-pair descriptors the meaning of each split is directly interpretable, thus it is relatively easy to build a model; and finally a small number of splits can lead to a fairly detailed model of activity. A preliminary experiment has been carried out using a reduced graph derived decision tree to select compounds for screening and has shown promising results. We are currently investigating the combination of recursive partitioning and reduced graphs applied to a wider variety of activity classes.

ACKNOWLEDGMENT

This work was funded by the Engineering and Physical Sciences Research Council and AstraZeneca. We thank Daylight Chemical Information Systems Inc. for software support and Current Drugs Limited for provision of the ID Alert database. The Krebs Institute for Biomolecular Research is a designated Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

REFERENCES AND NOTES

(1) *Virtual Screening for Bioactive Molecules*; Böhm, H.-J., Schneider, G., Eds.; Wiley-VCH: Weinheim, 2000.

(2) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(3) Bajorath, J. Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233−245.

(4) 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.

(5) 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.

(6) Matter, H. Selecting Optimally Diverse Compounds from Structure databases: A Validation Study of Two-Dimensional and Three-Dimensional Descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.

(7) Flower, D. R. On the Properties of Bit String Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.

(8) Godden, J. W.; Xue, L.; Bajorath, J. Combinatorial Preferences Affect Molecular Similarity/Diversity Calculations Using Binary Fingerprints and Tanimoto Coefficients. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 163−166.

(9) Dixon, S. L.; Koehler, R. T. The Hidden Component of Size in Two-Dimensional Fragment Descriptors: Side Effects on Sampling in Bioactive Libraries. *J. Med. Chem.* **1999**, *42*, 2887−2900.

(10) Pirard, B.; Pickett, S. D. Classification of Kinase Inhibitors Using BCUT Descriptors, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1431−1440.

(11) Schuffenhauer, A.; Gillet, V. J.; Willett, P. Similarity Searching in Files of Three-Dimensional Chemical Structures: Analysis of the BIOSTER Databases Using Two-Dimensional Fingerprints and Molecular Field Descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 295−307.

(12) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338−345.

(13) Daylight Chemical Information Systems Inc. Corporate Office 27401 Los Altos, Suite 360, Mission Viejo, CA 92691.

(14) Bemis, W.; Murcko, M. A. Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42*, 5095−5099.

(15) The ID Alert database is available from Current Drugs Ltd., Middlesex House, 34-42 Cleveland St., London, W1T 4LB.

(16) Cahart, R. E.; Smith, D. H.; Venkataraghavan, Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Application. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(17) Ginn, C. M. R.; Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W. Similarity Searching in Files of Three-Dimensional Chemical Structures: Evaluation of the EVA Descriptor and Combination of Rankings Using Data Fusion. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 23−37.

(18) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of Molecular Similarity Measures Using Data Fusion. *Perspec. Drug Discov. Design* **2000**, *20*, 1−16.

(19) Holliday, J. D.; Hu, C. Y.; Willett, P. Grouping of Coefficients for the Calculation of Inter-Molecular Similarity and Dissimilarity Using 2D Fragment Bit-Strings. *Combin. Chem. High-Through. Screening* **2002**, *5*, 155−166.

(20) Whittle, M.; Willett, P.; Klaffe, W.; van Noort, P. Evaluation of Similarity Measures for Searching the Dictionary of Natural Products Database, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338−345.

(21) Charifson, P. S.; Corkery, J. J.; Murco, M. A.; Walters, W. P. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.

(22) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*; Wiley-Interscience: New York, 2001.

(23) Hawkins, D. M.; Young, S. S.; Rusinko, A. Analysis of a Large Structure−Activity Data Set Using Recursive Partitioning. *Quant. Struct.-Act. Relat*. **1997**, *16*, 296−302.

(24) Rusinko, A.; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.

(25) Cho, S. J.; Shen, C. F.; Hermsmeier, M. A. Binary Formal Inference-Based Recursive Modelling Using Multiple Atom and Physiochemical Property Class Pair and Torsion Descriptors as Decision Criteria. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 668−680.

(26) Blower, P.; Fligner, M.; Verducci, J.; Bjoraker, J. On Combining Recursive Partitioning and Simulated Annealing to Detect Groups of Biologically Active Compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 393−404.

(27) CART is available from Salford Systems, 8880 Rio San Diego Dr., Ste. 1045, San Diego, CA 92108.

(28) Wang, P.; Wang, S. How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422−1426.

(29) Barker E. Ph.D. Thesis, University of Sheffield. In Preparation.

(30) Masek, B. B.; Merchant, A.; Matthew, J. B. Molecular Shape Comparison of Angiotensin II Receptor Antagonists. *J. Med. Chem.* **1993**, *36*, 1230−1238.

(31) Bradbury, R. H.; Allot, C. P.; Dennis, M.; Fisher, E.; Major, J. S.; Masek, B. B.; Oldham, A. A.; Pearce, R. J.; Rankine, N.; Revill, J. M.; Roberts, D. A.; Russell, S. T. New Nonpeptide Angiotensin II Receptor Antagonists. 2. Synthesis, Biological Properties, and Structure−Activity Relationships of 2-alkyl-4-(biphenylmethoxy)quinoline Derivatives. *J. Med. Chem.* **1992**, *35*, 4027−4038.

CI0255937