

Representing Clusters Using a Maximum Common Edge Substructure Algorithm Applied to Reduced Graphs and Molecular Graphs

Eleanor J. Gardiner,^{*,†} Valerie J. Gillet,[†] Peter Willett,[†] and David A. Cosgrove[‡]

Department of Information Studies, University of Sheffield, 211 Portobello St., Regent Court, Sheffield, United Kingdom, and AstraZeneca, Mereside, Alderley Park, Macclesfield, Cheshire, United Kingdom

Received October 16, 2006

Chemical databases are routinely clustered, with the aim of grouping molecules which share similar structural features. Ideally, medicinal chemists are then able to browse a few representatives of the cluster in order to interpret the shared activity of the cluster members. However, when molecules are clustered using fingerprints, it may be difficult to decipher the structural commonalities which are present. Here, we seek to represent a cluster by means of a maximum common substructure based on the shared functionality of the cluster members. Previously, we have used reduced graphs, where each node corresponds to a generalized functional group, as topological molecular descriptors for virtual screening. In this work, we precluster a database using any clustering method. We then represent the molecules in a cluster as reduced graphs. By repeated application of a maximum common edge substructure (MCES) algorithm, we obtain one or more reduced graph cluster representatives. The sparsity of the reduced graphs means that the MCES calculations can be performed in real time. The reduced graph cluster representatives are readily interpretable in terms of functional activity and can be mapped directly back to the molecules to which they correspond, giving the chemist a rapid means of assessing potential activities contained within the cluster. Clusters of interest are then subject to a detailed R-group analysis using the same iterated MCES algorithm applied to the molecular graphs.

INTRODUCTION

Chemical databases are routinely clustered, with the aim of grouping molecules which share similar structural features. However, the large data sets which result, for example, from high-throughput screening (HTS), produce many clusters, often containing hundreds of compounds. Ideally, medicinal chemists are then able to browse a few representatives of each cluster in order to interpret the shared activity of its members. This raises the question of how to select such representatives. Recent research has shown that medicinal chemists are inconsistent in their reviewing of molecules,¹ and an automated selection of cluster representatives might help to reduce this inconsistency. An obvious candidate for the cluster representative is the molecule with the highest mean similarity to all other compounds within the cluster. However, if the initial clustering is performed using noninterpretable descriptors such as molecular fingerprints, it may be difficult to decipher the structural commonalities which are present in this compound.

Clustering using topological features and other related problems has been the subject of several recent investigations. Stahl et al. describe a clustering method which uses the RASCAL algorithm,² incorporated into a hierarchical clustering method. Its aim is to maximize common substructural elements which are not necessarily connected in a single fragment.³ HierS is another hierarchical clustering algorithm which uses topological fragments, in this case, ring combina-

tions.⁴ Structural unit analysis attempts to find sets of rules for the connection of functional groups which represent molecules with high activity in HTS data. Finding all molecules corresponding to a rule then generates a cluster.⁵ The work most similar to that which we propose is by Stahl and Mauser.⁶ They first use a sphere exclusion algorithm⁷ to cluster a database represented as Daylight fingerprints. An iterative application of a maximum common subgraph (MCS) algorithm extracts a substructure common to all members of a cluster. Next, clusters containing a common substructural core are merged. A singleton is added to a cluster if it is sufficiently similar to the cluster's common substructural core. Thus, the main purpose of the MCS algorithm is to improve the clusters generated by initial clustering. The recent work of Raymond and Kibbey is also related to the cluster representation we describe.⁸ They map portions of molecules onto predefined structural templates using maximum common edge subgraph (MCES) and other graph-theoretic algorithms, such that the mapped substructures are maximally similar to the template substructures. The results are displayed in a hierarchic manner.

Here, we seek to represent a cluster by means of a maximum common substructure based on the shared functionality of the cluster members. We have previously described the use of reduced graphs (RGs) for the representation and searching of small molecules,^{9–13} as have other groups.^{14–16} In a RG, groups of atoms are replaced by a single node with the aim of capturing the functionality of the atoms. The nodes may be connected in a graph,^{13,14,16} stored as a fingerprint,^{9,10} or represented as a pseudo-SMILES string.¹¹

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

[†] University of Sheffield.

[‡] AstraZeneca.

In this work, we aim to exploit the generality of RGs as molecular descriptors in order to represent molecules which may have functional units rather than atoms in common. RGs are also smaller than chemical graphs and so are suited to interactively generating cluster representatives. We first cluster a database using any clustering method. Next, we represent all molecules in a cluster as RGs. An iterative application of a maximum common edge subgraph (MCES) algorithm finds one or more RGs which represent the common functionality of the compounds in the cluster. This summarizing of the cluster contents allows the medicinal chemist rapidly to reject clusters which are of no interest. For example, when examining compounds active against a kinase assay, the chemist might want rapidly to identify and discard clusters containing bis-anilinopyrimidines or anilinoquinazolines as not being novel. A cluster which seems more promising is then subjected to an R-group analysis using a very similar MCES algorithm applied to the all-atom structures of its compounds.

The remainder of the paper proceeds as follows: In the next section, we describe the graph reduction, followed by the cluster representation algorithm which is applied to a worked example. In the Results section, we apply the method to the representation of the MDL Drug Data Report (MDDR) database,¹⁷ illustrating the successes and drawbacks of the method in particular clusters. Finally, we apply the same MCES algorithm in an R-group analysis of some of the clusters to demonstrate the utility of the process as a whole.

METHODS

Reduced Graph Definitions. RG nodes are defined via user-defined SMARTS definitions. First, acid and base nodes are identified. These node types take precedence and are determined according to the SMARTS rules. The molecule is then partitioned into cyclic and acyclic fragments with rings identified using the Figueras ring perception algorithm.¹⁸ Cycles are defined as either aromatic or aliphatic. Within this definition, hydrogen-bonding capability is assigned as follows: no hydrogen-bonding characteristics, hydrogen-bond donor, hydrogen-bond acceptor, or hydrogen-bond donor and acceptor. Acyclic fragments also have their hydrogen-bonding capability assigned. Acyclic non-hydrogen bonding nodes are referred to as linkers. Note that terminal nonfeature nodes are also referred to as linkers. The 14 RG node types are detailed in Table 1—where possible, they are given an atomic code which is intuitively identifiable with the node characteristics. (The use of atom codes is convenient for programming.) Thus, for example, linker nodes have code Li and acid nodes have code Ac. Some example reduced graphs are shown in Figure 1. In the work described here, the SMARTS definitions defining acids, bases, and so forth are provided by AstraZeneca. The RG generation program was written in C++ incorporating routines from the OEchem toolkit.¹⁹

We have previously used a number of other levels of graph reduction.^{9–11} The levels form a hierarchy, with level 4 (illustrated in Figure 1) as the most discriminating and level 1 the most general. At level 1, nodes are one of three types: ring, acyclic hydrogen-bonding feature, or linker. Level 2 is the same as level 1 except that rings are classified as either aromatic or aliphatic. At level 3, rings are one of the

Table 1. Reduced Graph Nodes^a

node description	node code	frequency
acyclic inert node, linker	Li	209
acyclic feature node, acceptor only	Ga	50
acyclic feature node, donor only	Gd	8.2
acyclic feature node, both donor and acceptor	Ge	56
aromatic ring, hydrogen-bond acceptor	Na	6.4
aromatic ring, hydrogen-bond donor	Nd	33
aromatic ring, both donor and acceptor	Ne	4.4
aromatic ring, no hydrogen bonding	No	109
aliphatic ring, hydrogen-bond acceptor	Ca	11
aliphatic ring, hydrogen-bond donor	Cd	3.2
aliphatic ring, both donor and acceptor	Ce	1
aliphatic ring, no hydrogen bonding	Co	38
acid feature	Ac	12
base feature	Ba	31

^a Frequency is the normalized frequency of occurrence relative to that of the least common node type (Ce = aliphatic ring donor/acceptor) in the cleaned MDDR.

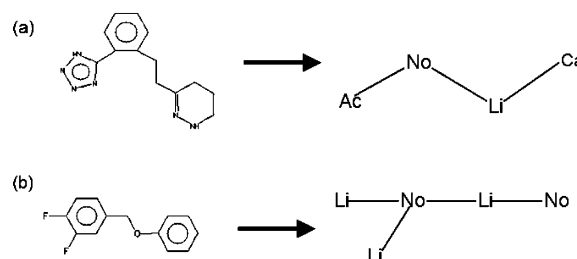


Figure 1. Example reduced graphs. Ac = acid feature, Ca = aliphatic ring donor, Li = linker, No = aromatic ring—no hydrogen bonding. (a) Acids take precedence over rings. (b) Terminal nonfeature nodes are described as linkers. Linker nodes may include heteroatoms if they have no hydrogen-bonding character.

following: aliphatic hydrogen-bonding, aliphatic non-hydrogen-bonding, aromatic hydrogen-bonding, or aromatic non-hydrogen bonding. Level 4 is the level at which different types of hydrogen bonding (donor/acceptor) and acids and bases are introduced. At any level, terminal nonfeature nodes (linkers) can be included or excluded.

Table 1 also shows the distribution of RG node types in the MDDR using level 4 graph reduction. As might be expected, linker nodes (Li) are the most common, occurring twice as often as the next most frequent type, which is the non-hydrogen-bonding aromatic ring node (No). Particularly infrequent node types are aliphatic rings with any kind of hydrogen-bonding capability (Ca, Cd, and Ce). Clearly, the occurrence of a particular node type will depend on its SMARTS definition and also on the particular set of molecules in the collection.

Initial studies considered the effect of level of reduction on the number of distinct RGs that might be found in a cluster, that is, if when clustered, all molecules in a cluster might have the same reduced graph. We investigated this hypothesis using the MDDR. The MDDR was first cleaned using a set of filters from AstraZeneca.²⁰ The 61 902-compound cleaned MDDR was clustered using Daylight fingerprints²¹ and a sphere exclusion algorithm⁷ with a Tanimoto cutoff of 0.7, giving 14 901 clusters, of which ~6000 were singletons. RGs were generated for all molecules in each of 8316 nonsingleton clusters, and the number of unique RGs per cluster was counted. On average, there were 6.65 compounds per nonsingleton cluster. The four

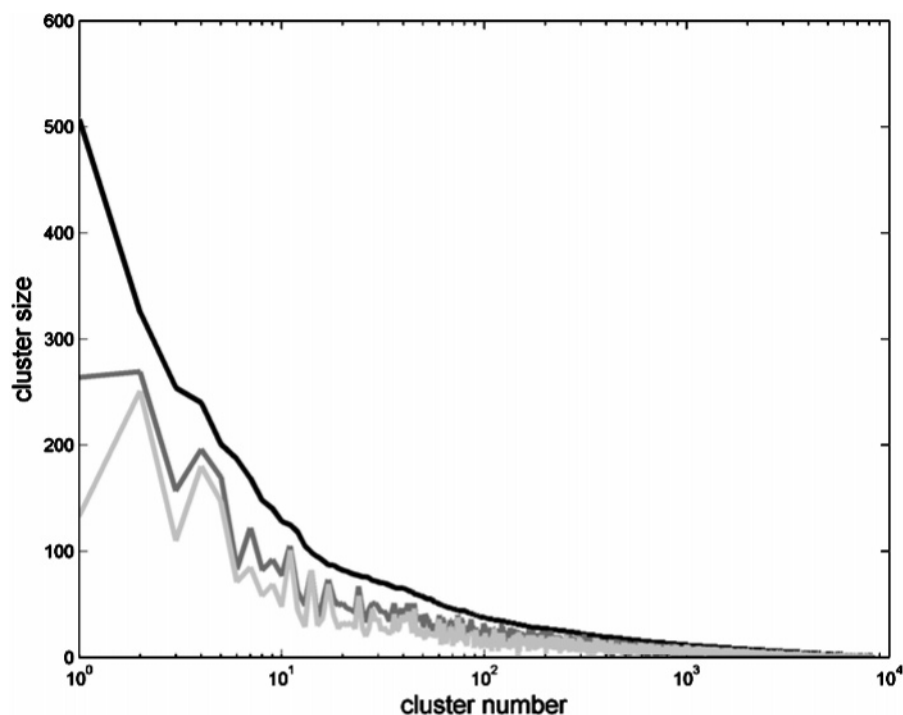


Figure 2. Relationship between the number of molecules per cluster and the number of RGs per cluster. Black line = molecules, dark gray line = RGs with terminal linkers, light gray line = RGs without terminal linkers. This figure was drawn in Matlab 7.1.²²

levels of graph reduction were used, each with and without keeping terminal linker nodes, giving eight different reductions. The results are summarized in Table 2.

It is clear from Table 2 that the more discriminating the level of graph reduction, the more reduced graphs are found, on average, within a cluster. Our normal method of graph reduction, level 4 including terminal linkers, is the least general of the graph representations considered and, as might be expected, resulted in the most unique RGs and the fewest number of clusters represented by a single unique reduced graph. The greatest reduction was achieved using level 1 excluding terminal linkers. However, even this level of graph reduction only resulted in 28% of clusters having a single unique RG representative. Since our experience with reduced graphs indicates that level 4 with terminal linker nodes is the RG representation which performed best in similarity searching and virtual screening,^{9,13} it was decided to further investigate cluster representation at this level of reduction. However, it is clear from Table 2 that the absence of terminal linker nodes at any level of graph reduction increases the ability of RGs to represent clusters. We therefore also investigated level 4 reduction excluding terminal linker nodes.

Figure 2 shows the relationship between the number of molecules per cluster and the number of RGs per cluster. The clusters are numbered in decreasing order of size so that cluster 1 has the most molecules. The *x* axis is plotted on a log scale, which clearly shows that, while the number of RGs generally falls with decreasing cluster size, the relationship is certainly not linear. Thus, the number of RGs per cluster depends on the cluster composition as well as its size.

Cluster Representation Algorithm. We used MCES calculations in a manner very similar to that described by Stahl and Mauser for chemical graphs⁶ in order to find a reduced graph MCES cluster representative. We first pre-cluster a database using some non-RG method. Next, for a given cluster, considering only those unique RGs present within the cluster, we find a RG MCES for as many RGs as possible, given certain user-defined constraints such as the minimum number of edges in the MCES and the minimum RASCAL similarity (minsim) between any two pairs of RGs represented by the MCES. The RASCAL score between two

Table 2. Counting Reduced Graphs within a Cluster in the MDDR

RG level	include terminal linkers			exclude terminal linkers		
	total number of unique RGs	mean number of RGs per cluster	clusters with only one RG	total number of unique RGs	mean number of RGs per cluster	clusters with only one RG
4	38 126 ^a	4.6	887	30 119	3.6	1742
3	36 774	4.4	946	26 973	3.2	2066
2	35 652	4.3	1022	25 318	3.0	2318
1	35 214	4.2	1042	24 765	3.0	2356

^a "Normal" graph reduction as described above is level 4 with terminal linker nodes included. Level 1 is the most general, level 4 the most detailed. There are 55 315 compounds in 8316 clusters giving an average of 6.65 molecules per cluster.

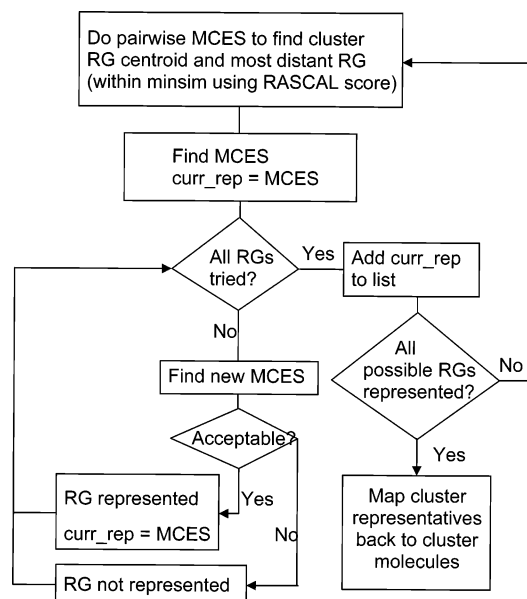


Figure 3. Clustering algorithm. Minisim is a user-defined minimum similarity. The centroid is the molecule with the most RG neighbors within minsim.

graphs, G_A and G_B , is a measure of the size of their MCES, G_{AB} , and is given by

RASCAL score =

$$\frac{[|V(G_{AB})| + |E(G_{AB})|]^2}{[|V(G_A)| + |E(G_A)|][|V(G_B)| + |E(G_B)|]} \quad (1)$$

where $V(G)$ is the number of vertices in graph G and $E(G)$ is the number of edges.

We find a RG MCES by first performing MCES calculations between all pairs of RGs in order to create a near-neighbors list. The RG with the most neighbors within minsim we term the *centroid*. We then determine its furthest neighbor (within minsim) and find the MCES between these two RGs. This MCES is the current cluster representative. We then compare all remaining RGs with the current representative, in input order, finding a new MCES whenever one RG is not a subgraph of the other. The current representative is replaced, providing that the new MCES meets the minimum similarity and minimum number of edges requirements. If these constraints are not met, then the RG under consideration cannot be represented by the current cluster member. When all RGs have been considered, we have found a cluster representative. We then iterate this procedure for any unrepresented RGs within the cluster, resulting in a small number of RG cluster representatives. Finally, we map back from the cluster representatives to all the molecules present in the cluster. A flowchart of the algorithm is shown in Figure 3. The MCES calculations were performed using the RASCAL algorithm.^{23,24} In their MCS calculations, Stahl and Mauser used the graph matching and clique-detection routines from the OEChem library.¹⁹

We have previously studied several different types of RG with both MCES and maximum common induced subgraph (MCIS) graph matching,¹³ and of those, we have chosen here to use the topologically connected RG MCES. In the *topologically connected* (TC) RG, nodes are only joined if

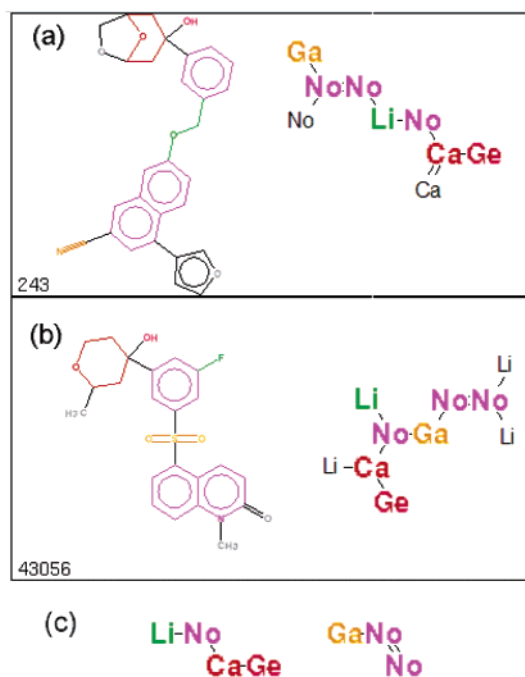


Figure 4. MCIS versus MCES. Ca = aliphatic ring acceptor, Ga = acyclic feature—acceptor only, Ge = acyclic feature—both donor and acceptor, Li = linker, No = aromatic ring—no hydrogen bonding. (a and b) 5-Lipoxygenase inhibitors with their RGs. The MCES atoms in each molecule are colored by the RG node type to which they belong. (c) The RG representation of the MCES, showing that the MCES falls into two components. The MCIS is just the Li–No–Ca–Ge component since the through-bond distance between the two fragments is not the same.

there is a bond between a pair of atoms, one from each node, in the chemical graph. In this case, the edge is labeled with the bond type (single, or, in the case of two nodes representing a pair of fused rings, fused). TC MCES graph matching has important advantages. It was consistently fast, which is important for real-time cluster representation, and equally importantly, a TC MCES is interpretable. By this, we mean that there is a straightforward mapping from the MCES to a subset of the nodes and edges of the RG since each edge in the MCES represents an edge in the RG. A further point to take into consideration is that, in the current application, we are performing MCS calculations between pairs of RGs which are known to be similar since they each represent molecules which have previously been assigned to the same cluster. In such cases, we have observed that the MCES is frequently larger than the MCIS, since the MCES may be composed of disconnected fragments whereas the MCIS is connected. This is illustrated in Figure 4 in which two 5-lipoxygenase inhibitors from the ID Alert database²⁵ are colored according to the RG nodes of their TC MCES. The MCES has seven nodes, in two components, comprising Li, No, Ca, and Ge in the first component and Ga, No, and No in the second. In contrast, the fully connected MCIS comprises just the single component with nodes Li, No, Ca, and Ge.

In previous RG experiments, we found the best results were obtained when joint hydrogen-bond donor/acceptor nodes were allowed to match either hydrogen-bond donors or hydrogen-bond acceptors.⁹ We implemented this relaxed matching as the default for cluster representation but allowed

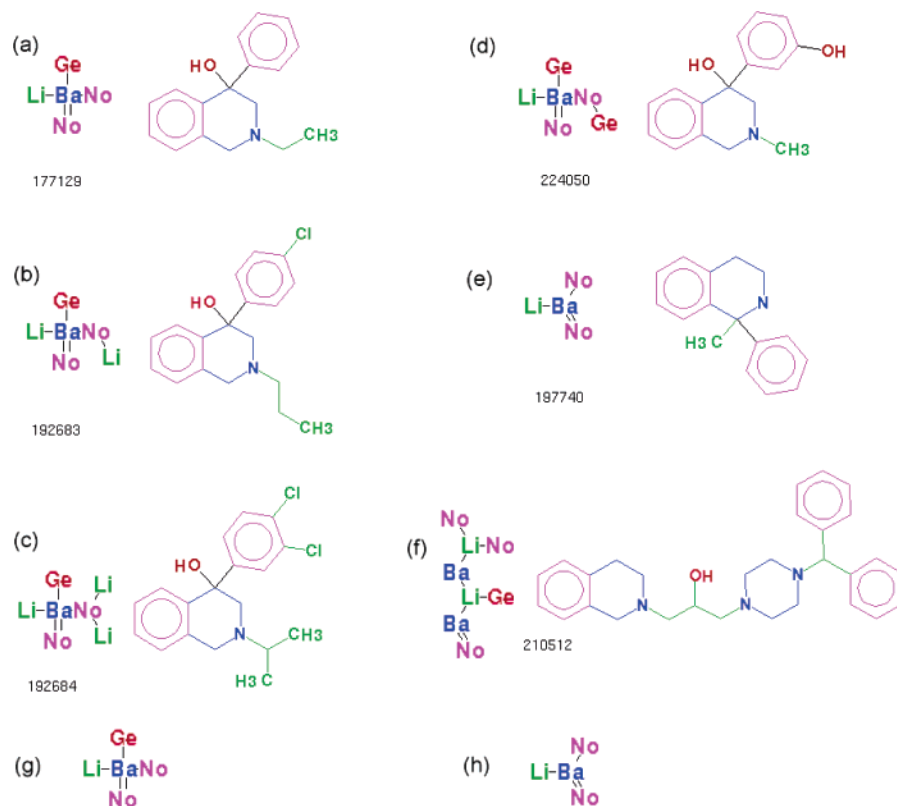


Figure 5. Finding a cluster representative for cluster 904. Ba = base feature, Ge = acyclic feature—both donor and acceptor, Li = linker, No = aromatic ring—no hydrogen bonding. (a–f) The unique RGs, each shown highlighted with the first molecule of the cluster which has that RG. The atoms are colored by RG node color. (g) The RG MCES of the cluster centroid, RG (a), and its most distant neighbor, RG (c). (h) The cluster representative.

this to be overridden by user-defined matching (or non-matching).

A further aim of the work is to map back from a cluster representative, which can be viewed as a RG SMILES string, to the molecules which it represents. We wish to isolate features common to all molecules and, in particular, to ascertain which RG nodes do, in fact, represent exactly the same set of atoms in each molecule. We therefore slightly modified the internal representation of a RG so that each node could be linked to its corresponding real atoms in the original molecule.

Experimental Details. We performed all experiments in the MDDR database cleaned as described above. Our cluster representation method is designed to be used in conjunction with any nonhierarchical clustering algorithm. In these experiments, we have used sphere exclusion⁷ with Daylight fingerprints²¹ as this represents a typical cluster method in use at AstraZeneca and widely elsewhere within the pharmaceutical industry. We applied a Tanimoto cutoff of 0.7. We chose this cutoff since it is used at AstraZeneca and gives clusters of a granularity appropriate for RG cluster representation. Very tight clusters have less need of representation since they are composed of very similar molecules, while loose clusters contain more, and more diverse molecules, and therefore are likely to have very small cluster representatives. The MCES and cluster representation programs were written in C++ and included routines from the OEChem and Ogham toolkits¹⁹ and Qt 4.1 from Trolltech.²⁶

The process of deriving a representative for cluster 904 using level 4 graph reduction with terminal linkers is

illustrated in Figure 5 with minsim set to 0.5 and minimum edges to 3. Cluster 904 contains 17 molecules but only six unique RGs. These RGs are shown in Figure 5a–f. Beside each RG is the first cluster molecule represented by that RG. The atoms of each molecule are colored according to the RG node type to which they belong. We do not normally color bonds which join two atoms which are in RG nodes of different types. The exception is linker nodes. Since most linker atoms are carbons (which are not routinely labeled), we color the bonds to any linker atom in the same color as that of the linker atoms (green) so that linkers can always be seen. RGs a–e each have the maximum (four) neighbors within a RASCAL score of 0.5, and so RG a is arbitrarily chosen as the cluster centroid. Its furthest neighbor is RG c, and their RG MCES is shown in Figure 5g. This RG is the current cluster representative, denoted as *curr_rep*. *curr_rep* is then compared to the remaining RGs in turn. *curr_rep* is a subgraph of both RGs b and d, and so the RG MCES does not change. When the iterated MCES process reaches RG e, a new (smaller) MCES is found, shown in Figure 5h. This MCES then becomes the new *curr_rep*. The similarity of *curr_rep* to RG f is less than minsim, and so *curr_rep* is the cluster representative, with RG f remaining unrepresented.

Figure 6 shows all 16 molecules represented by *curr_rep*. Atoms corresponding to the nodes of *curr_rep* are highlighted in the same colors as those of the RG nodes. It is clear that the single molecule omitted from the cluster representation (Figure 5f) is substantially different from the other molecules in the cluster.

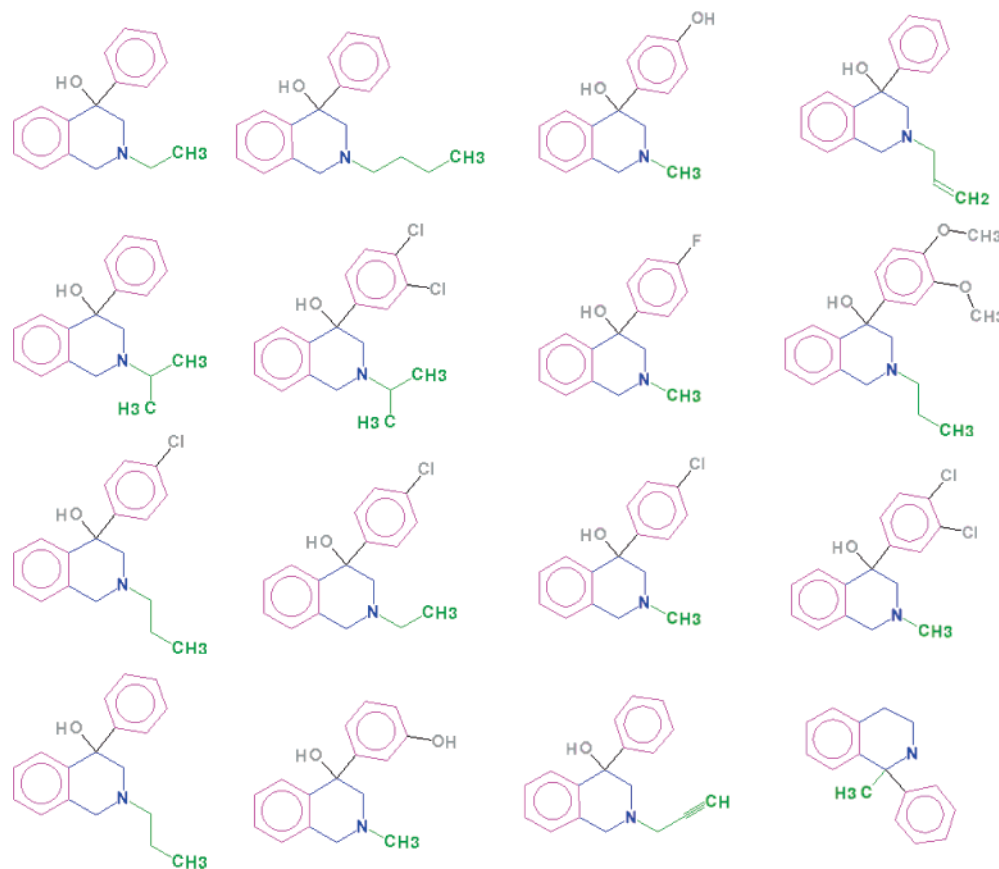


Figure 6. Mapping the cluster representative RG onto the cluster molecules. The 16 molecules represented by the RG of Figure 5h with atoms corresponding to RG nodes colored by node type. Gray atoms are not part of the cluster representative RG.

RESULTS AND DISCUSSION

Our tests were performed on the cleaned MDDR, clustered as previously described. The sphere exclusion clustering method produces the biggest clusters first. Since such clusters are both the most challenging for an MCES calculation and also those most in need of representation (to reduce the effort of the medicinal chemist), we retained the first (in the order created) 3000 clusters for cluster representation using RGs and MCES calculations. Of these, 229 were singletons; that is, they contained only a single molecule and so were removed. We converted each cluster into a set of reduced graphs, one for each molecule. This initial clustering and RG generation step took about 50 min on a 2.8 MHz Linux PC. Initial experiments revealed that a small number (eight) of 3000 clusters were composed of very large and very similar molecules, with a large number of reduced graph nodes. Tests showed that our MCES calculations timed out when processing these few clusters. As one of our goals is on-the-fly cluster representation, we discarded these clusters, leaving 2763 clusters containing 34 900 compounds. We used a minimum similarity RASCAL score (minsim) of 0.5 and required the RG cluster representative to have at least two edges. Deriving cluster representatives for all 2763 clusters takes about 10 min on a desktop 2.8 MHz Linux PC. (For comparison, Stahl and Mauser report a time of 3 h for the equivalent step in their algorithm for 50% more compounds.) The results are summarized in Table 3.

For level 4 graph reduction, with terminal linkers included, there are on average 1.58 cluster representatives per cluster.

Table 3. Cluster Representative Results with Terminal Linkers Included and Excluded^a

	terminal linkers included	terminal linkers excluded
mols represented	33336	29763
clusters considered	2771	2771
mean mols per rep (standard dev)	7.65 (12.08)	7.85 (11.56)
mean reps per cluster	1.58	1.49
clusters with single rep	1465	1557
unique cluster reps	3945	3104
total cluster reps	4358	3791
RG nodes per rep	6.55	5.55
clusters not represented	73	222
clusters with some mols not represented	139	157
mean mols not processed per cluster	0.20	0.35

^a mol = molecule and rep = cluster representative. A minimum of two edges was required in the MCES together with a minimum RASCAL similarity score of 0.5 for all molecules represented by the same cluster representative.

Over 95% of the molecules are represented; only 73 clusters are not represented. If no cluster member has neighbors within 0.5 or if any potential RG cluster representative has fewer than two edges, then the cluster cannot be represented using these constraints. The mean number of RG nodes per cluster representative is 6.5, which is pleasing, since most clusters are not represented by trivially small RGs.

When terminal linkers are not included, only 85% of the compounds are represented. This reflects the fact that this level of graph reduction gives smaller RGs, and so more potential cluster representatives fail the minimum-edge

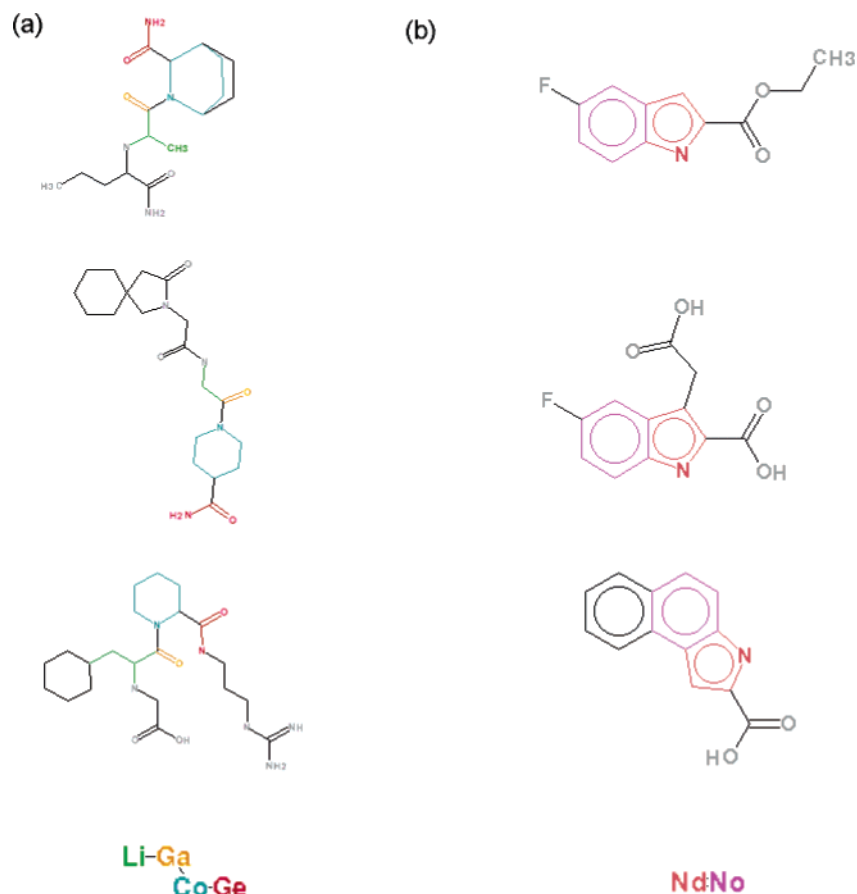


Figure 7. Clusters represented using less stringent constraints. Co = aliphatic ring—no hydrogen bonding, Ga = acyclic feature—acceptor only, Ge = acyclic feature—both donor and acceptor, Li = linker, Nd = aromatic ring donor, No = aromatic ring—no hydrogen bonding. (a) Large molecules with relatively small common core so minsim = 0.2. (b) Minimum edges = 1.

criterion. Although the RG nodes which are present tend to represent functional groups (since fewer of them are linkers), the fact that fewer compounds can be represented led us to reject this level of graph reduction and to proceed with the inclusion of terminal linker nodes.

On average, there are fewer than two cluster representatives per cluster, which shows that the initial clustering of the entire MDDR using Daylight fingerprints and sphere exclusion has worked very well for the majority of the larger clusters of which our 3000-cluster data set is mainly comprised. For example, the largest single cluster has 507 molecules but only 265 unique RGs (with terminal linker nodes). These 265 RGs are represented by only five cluster representatives, of which one represents 476 molecules, over 90% of the cluster members.

About 10% of the cluster representatives represent molecules in more than one cluster. In principle, therefore, we could use such common representatives as an aid in merging clusters. (This is the essence of the method of Stahl and Mauser.⁶) However, since about half of all the clusters have more than one representative, it is not necessarily straightforward to decide which clusters should be merged, and at present, cluster merging has not been implemented.

The clusters which are not represented illustrate the need to vary minsim, and occasionally the minimum number of edges, depending on the cluster being represented. The RASCAL similarity score takes account of the size of the graphs, and so two large compounds with a significant MCES

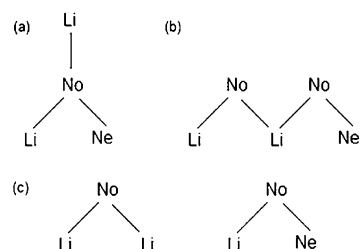


Figure 8. Molecules with more than one MCES. (a and b) Two hypothetical RGs. (c) Two possible MCESs.

may still have a similarity score less than 0.5. Sets of very small molecules are generally less likely to have a representative with three edges. Two clusters which are not represented when minsim = 0.5 and minimum edges is set to 3 are shown in Figure 7 along with their cluster representative, calculated using lower constraints.

Limitations. In the interest of speed, we choose only to find the first MCES in any MCES calculation. As a result, the cluster representative is not always the most appropriate representative. This will always be a problem (and is not confined to RG molecular representations). However, one possible amelioration would be to down-weight linker nodes in the MCES calculation since we would usually prefer the cluster representative to contain as many features as possible. This is illustrated by the two hypothetical RGs in Figure 8. There are two possible MCESs; one contains two linker nodes and the other, only one. A second limitation is that

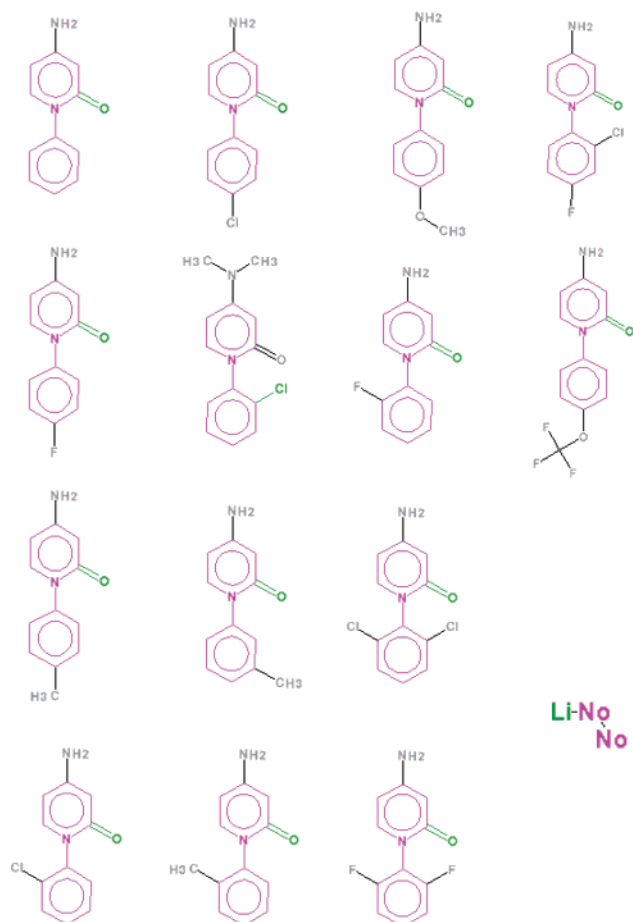


Figure 9. Limitation in representation of cluster 1215. The second molecule in the second row has a chlorine atom mapped as the linker. In all the other molecules, the oxygen has been chosen.

there is not always a unique mapping from the cluster representative back to the original molecules. This will always be a problem in some instances (for example, symmetric molecules) and, again, is not unique to RGs, but it could be resolved in an ad hoc manner in many cases, by using the information present in other cluster molecules. For example, in the case of a node which has only one candidate set of atoms for mapping in most molecules, but more than one in a few molecules, it is sensible to choose the atom set which is most similar to the case where there is no choice. This can be seen in Figure 9 where the representation of cluster 1215 is shown. The second molecule on the second row has a chlorine atom selected as the linker, whereas in all the other molecules, an oxygen is selected. Clearly, a more intelligent mapping would select the oxygen atom bonded to the other aromatic ring. These enhancements are intended for inclusion at a later stage.

The method as implemented is dependent on the order of the input molecules to some extent. Finding the RG centroid and the most distant neighbor as the initial step has the effect of reducing this dependency since, except in the case of ties, this step is order-independent. We investigated the effect of changing the order of the input molecules in the MDDR tests described above and found that approximately 75% of the clusters had exactly the same representatives, each representing the same numbers of molecules.

Representing Individual Clusters. The quality and detail of the cluster representation clearly depends on the cluster

composition and, as illustrated in Figure 7, the choice of good values for minsim and minimum edges. In most cases, we have found that minsim = 0.5 and minimum edges = 2 give acceptable results. Figure 10 illustrates the cluster representation of the 14 molecules of cluster 688 using these parameter values. It is immediately clear which are the fixed regions of the compounds, where the atoms or groups with a common function are, and where the variability occurs. Five- and six-membered aromatic rings are shown to be equivalent, at least as far as the RG representation is concerned.

Advantages of the RG Approach to Cluster Representation. *Finding Series.* One important use for the RG cluster representation is the fast division of individual clusters into two or more series. For example, cluster 1047 has 16 compounds, 12 of which are represented by one RG and the other four by a second RG (illustrated in Figure 11). It is clear that the two RGs have captured the differences between the two series.

Finding Other Similar Compounds. The cluster representative can also be used as a search pattern, to find additional molecules containing the same RG substructure. The pattern in Figure 11a was used to search the MDDR in a RASCAL similarity search with a minimum similarity of 0.5. A total of 380 compounds were retrieved, among which were the members of another cluster of 15 molecules, cluster 1146. It is clear that, although our initial clustering method has placed the molecules in Figures 11 and 12 in a separate cluster, they are indeed very similar. This technique can be used to combine clusters or possibly to add an outlier to its most similar cluster neighbor.

A simpler search can also be considered. Since the RG can be written as a SMILES string, it can also be used as a SMARTS search pattern, and all of the RG SMILES of a compound collection can be scanned extremely rapidly to find similar molecules (i.e., all compounds with an identical RG representation).

Speed. Using RGs rather than molecules is a great advantage when it comes to MCES calculations. All except very large clusters of large and very similar molecules can be processed interactively. Most clusters take less than 10 s on a desktop Linux PC. We implement the cluster representation of a group of clusters by using a script which repeatedly runs the program on clusters selected from a list. This means that the program processes a single cluster at a time, which maintains the interactivity of the procedure.

The main advantage of the RG cluster representation is the rapid summarization of the cluster compounds by at most a few RGs. This representation can aid the chemist in the rapid rejection of clusters which are of no interest. Having identified the more promising clusters, the RG approach is then too general, and so, for these clusters, a detailed R-group analysis of the compounds is normally performed.

R-Group Analysis. The purpose of an R-group analysis is normally to examine the structural variation of compounds within a cluster with a view to, for example, understanding any structure–activity relationship (SAR) for the compounds or suggesting changes that might improve the physical properties of a series. Generally, this commences with an identification of a common core for some or all of the compounds in the cluster, followed by the extraction of substituent tables for the points of variation within the

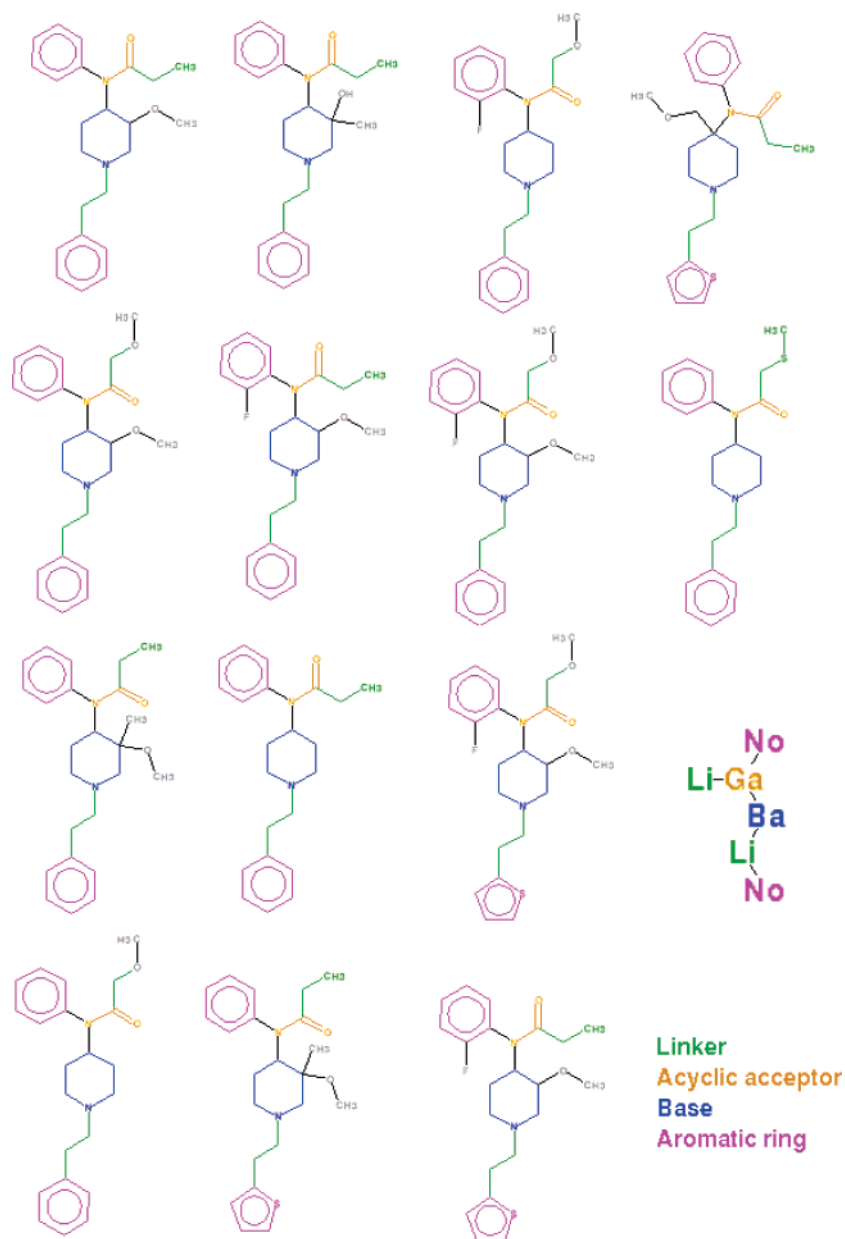


Figure 10. Typical cluster representation.

molecules. Historically at AstraZeneca, this core perception has been achieved by, among other methods, manual inspection of the structures in order to create a SMARTS pattern or ISIS query that is used in the substituent extraction process. In their MPX program, Kibbey and Calvet allow the user to draw a core with a 2D sketcher in order to perform such an analysis.²⁷ We have developed an automated method for discerning the core that uses a similar iterated MCES algorithm to that used for the RG cluster representation but works on the full molecular graphs. Because the program is considering all the atoms of each molecule, this procedure is relatively time-consuming and for clusters of more than a few tens of molecules does not run in interactive time. It is therefore run as a separate step on selected clusters only.

Having identified the core for the molecules in the cluster, a naïve R-group analysis is straightforward, involving extraction of the parts of the molecule emanating from each substituted core atom using a simple tree-searching algorithm. We use a xenon atom to mark the point of attachment of the

substituent to the core. In the case of cyclic substituents, an yttrium atom is used to mark the secondary attachment point.

This simple approach to building the R-group table is prone to ambiguities in the cases of local symmetry in the core, the most frequent example of which is substituted phenyl rings. In the structures in Figure 13, in which the common core is the toluene moiety, atom 3 of molecule A can map onto either atom 3 or atom 8 of molecule B; in the absence of additional information, both mappings are equally valid. The choice of mapping has clear implications for the final R-group table—the first mapping will have the two chlorine atoms in the same column; the second will have them in different columns. We have developed a set of rules to attempt to resolve the ambiguity in the case of phenyl rings, which, while not foolproof, are adequate for most occasions. The phenyl ring with the most substitutions is chosen, and the other molecules are mapped onto it, one by one, using the following rules:

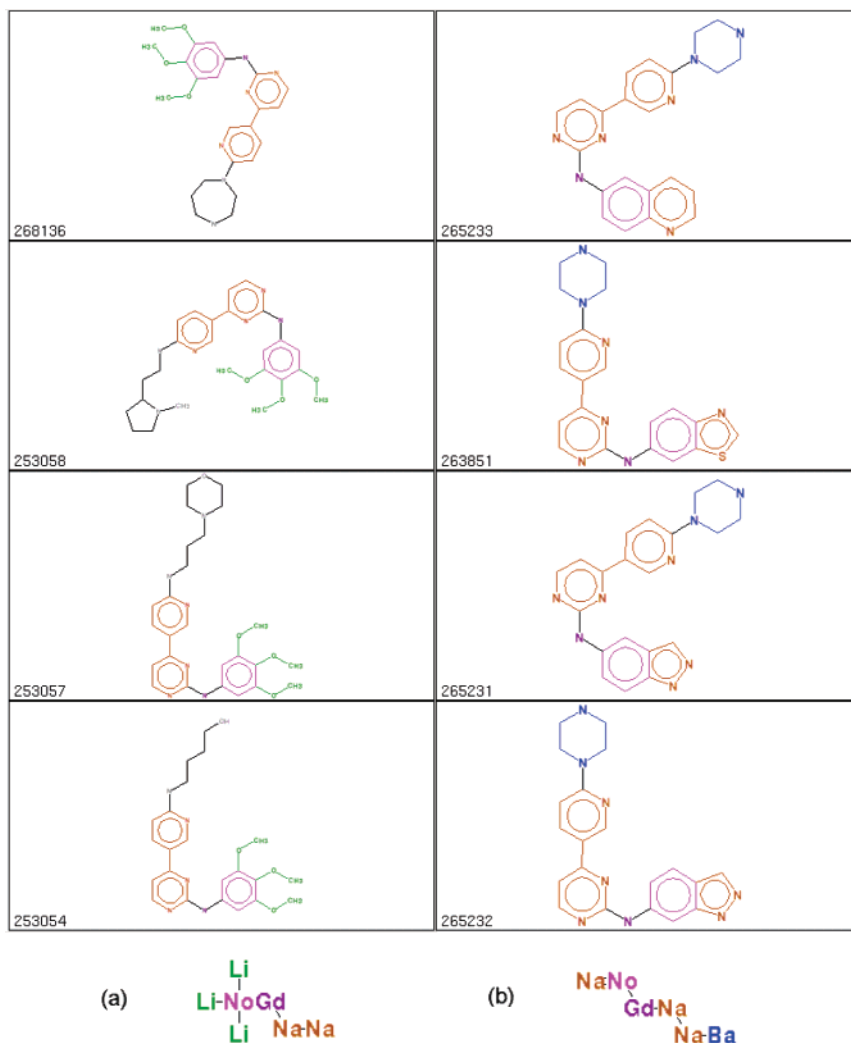


Figure 11. Cluster representation of cluster 1047. Ba = Base feature, Gd = acyclic donor feature, Li = linker, Na = aromatic ring—hydrogen-bond acceptor, No = aromatic ring—no hydrogen bonding. Left column: four of the 12 molecules represented by RG (a). Right column: all four molecules represented by RG (b).

1. See if there is an identical substituent in the same position in the two molecules (in this case, relative to the methyl carbon C1).

1.a. If there is a mapping that makes these equivalent, accept that mapping. In the molecules depicted in Figure 13, this would map atom 3 of molecule A onto atom 3 of molecule B.

1.b. If no exact mapping is possible, choose the mapping that has the two identical substituents as close to each other as possible, in terms of number of bonds between them. In Figure 13, this would map atom 4 of molecule C onto atom 5 of molecule A.

2. If there is no pair of identical substituents, choose the two largest, and apply the same rules as above; that is, first try for an exact mapping, then the nearest mapping. If the search reaches this last test, then the results are tending toward arbitrariness, but the results are at least consistent. In Figure 13, these tests result in mapping atom 4 of molecule D onto atom 7 of molecule A.

When these rules are applied to the molecules in Figure 13, the final mappings of the cores thus become as follows: C1—C2—C3—C5—C6—C7—C10 in molecule A corresponds to C1—C2—C3—C5—C6—C7—C8 in molecule B, C1—C2—

C3—C4—C6—C7—C9 in molecule C, and C1—C2—C10—C9—C8—C4—C3 in molecule D.

Results of applying this core-perception and R-group stripping to the clusters of Figures 9 and 10 are shown in Figures 14 and 15, respectively. Figure 14 shows the full R-group table for the structures; Figure 15 shows the summary of the R groups at each position, giving the number of times that each group appears.

An advantage, in this instance at least, of the RASCAL algorithm for MCES detection is that it allows for the production of fragmented cores with variable linker groups being identified, as shown in Figure 16. Care must be taken with the fragmented cores, however, to ensure that the orientations of the parts of the core are consistent. In molecule d of Figure 16, for example, the pyridyl group is attached to the rest of the molecule para to the nitrogen, rather than meta. This molecule does not, therefore, share the same core, so it is dropped from the subsequent R-group analysis, and the molecule is drawn in red. When allowing fragmented cores, another problem concerns the relative separation of the parts. For example, should a biphenyl be considered as sharing a core with two phenyl groups separated by a 10-

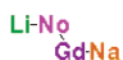
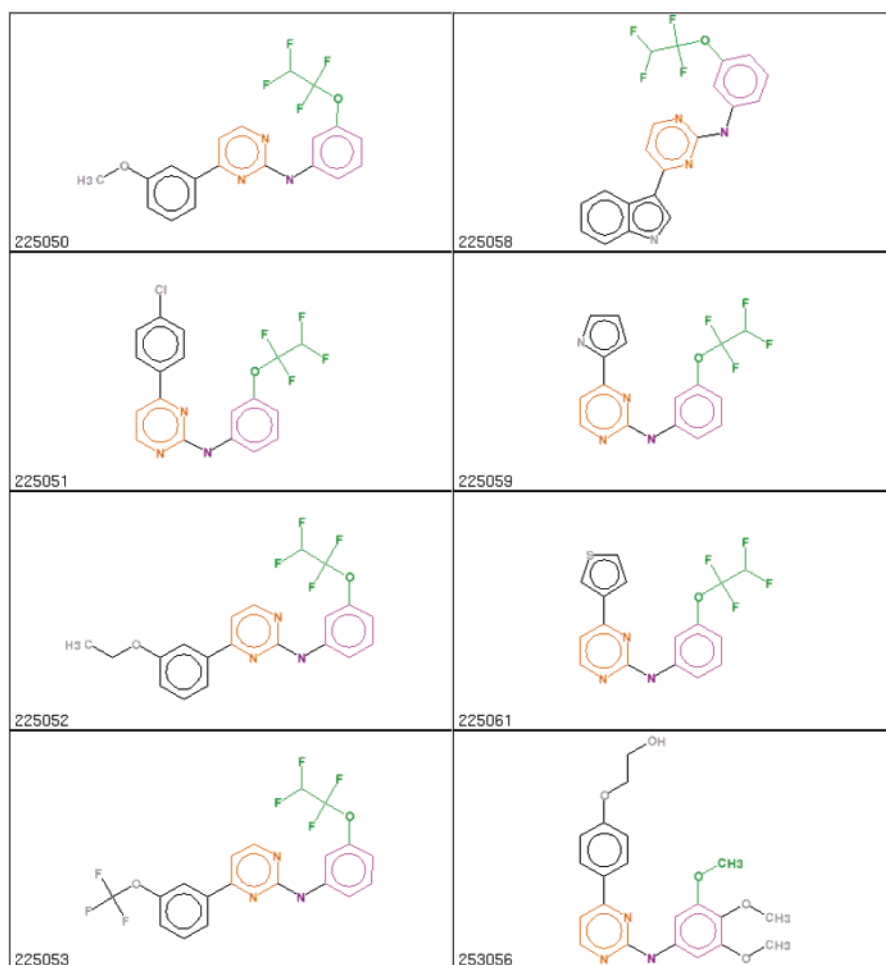


Figure 12. Molecules retrieved using the cluster representative of cluster 1047. Ba = base feature, Gd = acyclic donor feature, Li = linker, Na = aromatic ring—hydrogen-bond acceptor. Eight of the molecules of cluster 1146 which were retrieved by an RG similarity search using the RG of Figure 11a together with their RG representative.

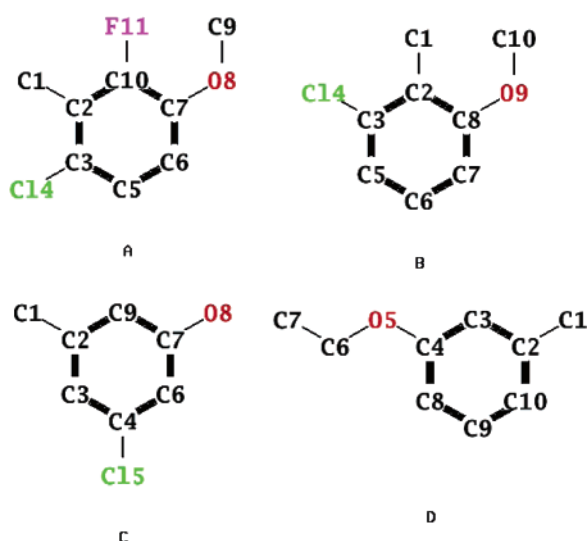


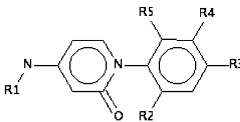
Figure 13. Resolving ambiguities in R-group mappings—example molecules. See text for details.

carbon chain, such that the linker in the former is one bond and in the latter is 11 bonds? We have left this as a decision for the user as an adjustable input parameter.

CONCLUSIONS



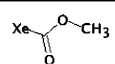



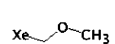
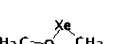

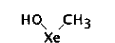



A RG represents a molecule as a topological graph where each node corresponds to a generalized functional group. Previously, we have used RGs as topological molecular descriptors for virtual screening. This paper has focused on the use of RGs as cluster representatives. We clustered the MDDR using Daylight fingerprints and a sphere exclusion clustering algorithm. We then represented the molecules in a cluster as RGs. By repeated application of a MCES algorithm, we obtained one or more reduced graph cluster representatives. The sparsity of the RG representation, in conjunction with the rapid RASCAL MCES algorithm, allowed the cluster representatives to be found interactively. The atoms of the cluster molecules which belong to the common cluster representative RG were colored according to their RG nodes type, giving the medicinal chemist a rapid means of assessing potential activities contained within the cluster.

The RG cluster representation showed the presence of multiple series within a cluster and identified outlying molecules within a cluster. Similarity searching with an RG cluster representative retrieved molecules from different



Molecule	R1	R2	R3	R4	R5
H	Xe	Xe	Xe	Xe	Xe
I	Xe	Xe	Xe-F	Xe	Xe
J	Xe	Xe	Xe-CH ₃	Xe	Xe
K	Xe	Xe	Xe	Xe	Xe-Cl
L	Xe	Xe	Xe-Cl	Xe	Xe
M	H ₃ C-Xe-CH ₃	Xe	Xe	Xe	Xe-Cl
N	Xe	Xe	Xe	Xe-CH ₃	Xe

Figure 14. R-group analysis of some of the molecules in cluster 1215. Each row in the table represents the substituent at the given position in the Markush diagram shown at the top. Full structures are shown in Figure 9. The Xe atom marks the point of attachment to the core, and a “bare” Xe indicates a simple hydrogen substituent.

R1	R2	R3	R4	R5
14 Xe-CH ₃	15 Xe	15 Xe	9 Xe	10 
5 	5 Xe-F	4 	6 	4 
1 		1 	3 	3 
		2 	1 	1 
				1 

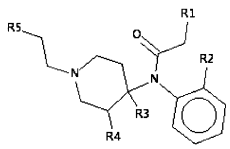


Figure 15. R-group summary for molecules in cluster 688. Full structures of some of the molecules are shown in Figure 10. Each column in the table represents the different substituents at the corresponding position in the Markush diagram. The number in the top left of each cell denotes the number of times that substituent appears in the molecule set. For example, at the R2 position (ortho to the aniline–nitrogen), there are 15 molecules with a hydrogen and five with a fluorine atom.

clusters with similar RGs, allowing the identification of related clusters.

Clusters of interest were then subject to a detailed R-group analysis using the same iterated MCES algorithm but applied to the molecular graphs. This process is too time-consuming

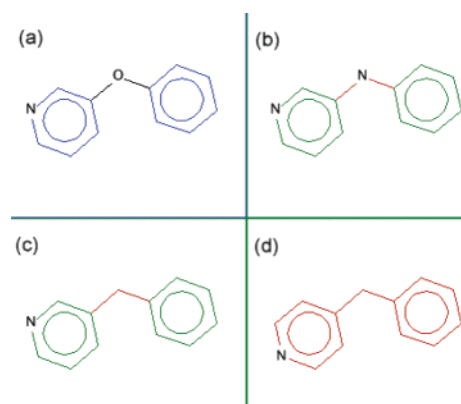


Figure 16. Molecules with a fragmented core and different linking groups. In molecule a, the core atoms are colored blue to show that this is the reference molecule; molecules b and c have the core colored in green. Molecule d is colored red because it does not exhibit the core—the pyridyl nitrogen is para to the methylene rather than meta.

to be applied indiscriminately, but RG cluster representation aided in the identification of those clusters appropriate for R-group analysis. After a further step, an unambiguous common core for the molecules in the cluster was generated and the variations across the molecules presented as a substituent or R-group table which can then be used, for example, to examine any SAR that may be present within the cluster.

ACKNOWLEDGMENT

This work was funded by AstraZeneca. We thank David Buttar and Paula Daunt for helpful discussions. We also thank Openeye Scientific Software Inc. and Daylight Chemical Information Systems Inc. for software support and MDL for provision of the MDDR database.

REFERENCES AND NOTES

- (1) Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. Assessment of the Consistency of Medicinal Chemists in Reviewing Sets of Compounds. *J. Med. Chem.* **2004**, *47*, 4891–4896.
- (2) Raymond, J. W.; Gardiner, E. J.; Willett, P. Heuristics for Similarity Searching of Chemical Graphs Using a Maximum Common Edge Subgraph Algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 305–316.
- (3) Stahl, M.; Mauser, H.; Tsui, M.; Taylor, N. R. A Robust Clustering Method for Chemical Structures. *J. Med. Chem.* **2005**, *48*, 4358–4366.
- (4) Wilkens, S. J.; Janes, J.; Su, A. I. HierS: Hierarchical Scaffold Clustering Using Topological Chemical Graphs. *J. Med. Chem.* **2005**, *48*, 3182–3193.
- (5) Wolohan, P. R. N.; Akella, L. B.; Dorfman, R. J.; Nell, P. G.; Mundt, S. M.; Clark, R. D. Structural Unit Analysis Identifies Lead Series and Facilitates Scaffold Hopping in Combinatorial Chemistry. *J. Chem. Inf. Model.* **2006**, *46*, 1188–1193.
- (6) Stahl, M.; Mauser, H. Database Clustering with a Combination of Fingerprint and Maximum Common Substructure Methods. *J. Chem. Inf. Model.* **2005**, *45*, 542–548.
- (7) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way to Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747–750.
- (8) Raymond, J. W.; Kibbey, C. E. An Automated Method for Exploring Targeted Substructural Diversity within Sets of Chemical Structures. *J. Chem. Inf. Model.* **2005**, *45*, 1195–1204.
- (9) Barker, E. J. Chemical Similarity Searching Using Reduced Graphs. Ph.D. Thesis, University of Sheffield, Sheffield, United Kingdom, 2004.
- (10) 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.

- (11) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338–345.
- (12) Birchall, K.; Gillet, V. J.; Harper, G.; Pickett, S. D. Training Similarity Measures for Specific Activities: Application to Reduced Graphs. *J. Chem. Inf. Model.* **2006**, *46*, 577–586.
- (13) Barker, E. J.; Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Kitts, P.; Willett, P.; Gillet, V. J. Scaffold Hopping Using Clique Detection Applied to Reduced Graphs. *J. Chem. Inf. Model.* **2006**, *46*, 503–511.
- (14) Rarey, M.; Dixon, J. S. Feature Trees: A New Molecular Similarity Measure Based on Tree Matching. *J. Comput.-Aided Mol. Des.* **1998**, *12* (5), 471–490.
- (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) MDL; MDL Information Systems, Inc.: San Leandro, CA. <http://www.mdli.com> (accessed Nov 30 2006).
- (18) Figueras, J. Ring Perception Using Breadth-First Search. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 986–991.
- (19) OEchem; OpenEye Scientific Software, Santa Fe, NM. <http://www.eyesopen.com> (accessed Nov 30 2006).
- (20) 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.
- (21) Daylight Chemical Information Systems, Inc., Los Altos, CA. <http://www.daylight.com> (accessed Nov 30 2006).
- (22) Matlab, version 7.1; The Mathworks, Inc.: Novi, MI. <http://www.mathworks.com> (accessed Nov 30 2006).
- (23) Raymond, J. W.; Gardiner, E. J.; Willett, P. RASCAL: Calculation of Graph Similarity Using Maximum Common Edge Subgraphs. *Comput. J.* **2002**, *45*, 631–644.
- (24) 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.
- (25) I. D. Alert Database; Current Drugs Ltd.: London, U. K.
- (26) Qt, version 4.1; Trolltech: Oslo, Norway. <http://www.trolltech.com> (accessed Nov 30 2006).
- (27) Kibbey, C.; Calvet, A. Molecular Property eXplorer: A Novel Approach to Visualizing SAR Using Tree-Maps and Heatmaps. *J. Chem. Inf. Model.* **2005**, *45*, 523–532.

CI600444G