Use of Reduced Graphs To Encode Bioisosterism for Similarity-Based Virtual Screening

Kristian Birchall, Valerie J. Gillet, and Peter Willett*

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, 211 Portobello Street, Sheffield S1 4DP, United Kingdom

Pierre Ducrot

Discngine, 102 Avenue Gaston Roussel, 93230 Romainville, France

Claude Luttmann

Chemical and Analytical Sciences, Sanofi-Aventis, 94400 Vitry-sur-Seine, France

Received February 27, 2009

This paper describes a project to include explicit information about bioisosteric equivalences between pairs of fragment substructures in a system for similarity-based virtual screening. Data from the BIOSTER database show that reduced graphs provide a simple way of encoding known bioisosteric equivalences in a manner that can be used during similarity searching. Scaffold-hopping experiments with the WOMBAT database show that including such information enables similarities to be identified between the reference structures and active structures from the database that contain different, but equivalent, fragment substructures. However, such equivalences also contribute to the similarities between the reference structures and inactives, and the latter equivalences can swamp those involving the actives. This presents serious problems for the routine use of information about bioisosteric fragments in similarity-based virtual screening.

INTRODUCTION

Research programs in the pharmaceutical and agrochemical industries seek to identify novel, synthetically feasible molecules that exhibit useful biological activity while having minimal side-effects. This is an extremely difficult task, and many chemoinformatics approaches have been developed over the years to increase the cost-effectiveness of lead discovery and optimization.[1−3] One such approach is the study of bioisosterism, the identification of molecular fragments that are interchangeable without significant perturbation of a molecule's biological activity.[4,5] The identification of such fragments is of potential benefit in both the lead-discovery and lead-optimization stages of a research program, providing facilities for both scaffold-hopping and ADME enhancement.[6−13]

Analyzing bioisosteric fragments requires consideration of the factors affecting the ligand-target interaction, including size and shape, electronic properties (such as charge, H-bonding ability, and lipophilicity), flexibility, and group positioning. The detailed role of these factors in ligand-protein interactions is often poorly understood, especially as there is often a strong dependence on the target and ligand combination, i.e., fragments found to be bioisosteric in one pair of ligands against one target may not necessarily be bioisosteric in a different pair of ligands against a different target. Complications such as these make the development of computational tools for the identification of bioisosteric pairs a challenging task.

* Corresponding author phone: +44-114-2222633; e-mail: p.willett@sheffield.ac.uk.

Much of the information about bioisosterism in the published literature has been compiled into a database known as BIOSTER,[14] which provides a valuable resource for use in data mining and modeling bioisosterism and which has been used in several studies. For example, Schuffenhauer et al. investigated the effectiveness of 2D fingerprints and electrostatic-field descriptors for the retrieval of pairs of bioisosteric molecules in the BIOSTER database.[15] The motivation behind this work was to establish whether a particular approach, e.g., a particular type of descriptor, might be able to recognize putative bioisosteres by computing a higher degree (on average) of structural similarity between bioisosteric pairs from the BIOSTER database as compared to the degree of similarity between pairs of structures (either from BIOSTER itself or from another database) that are not known to be bioisosteric. A similar methodology was used in two subsequent studies to identify similarity measures that could suggest substructural fragments similar to a known substituent and that might thus be bioisosteric.[16,17] Holliday et al. investigated techniques for identifying pairs of molecular substituents that had a high similarity in terms of their sums of physicochemical properties, e.g., atomic weight, molar refractivity, hydrophobicity and H-bonding capability, at progressively greater numbers of bonds from the point of substitution.[16] Wagener and Lommerse focused on fingerprint techniques, using a version of the atom-pair approach of Carhart et al.[18] in which the atoms are coded using pharmacophoric features such as H-bond donors, hydrophobic centers, and attachment points.[17]

Rather than investigating bioisosterism explicitly, Sheridan sought to find the most common chemical replacements in

10.1021/ci900078h CCC: $40.75
Published on Web 06/01/2009

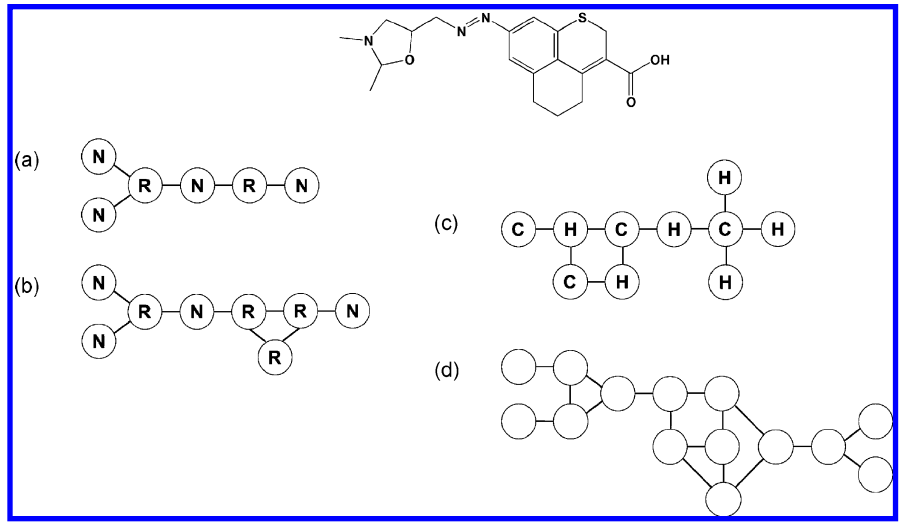


Figure 1. Examples of different reduction schemes applied to the same chemical structure. A ring/nonring reduction where fused rings are considered as a single node is shown in (a), while (b) shows the reduction where each smallest ring is treated as an individual node. A carbon/heteroatom reduction is shown in (c), while a homeomorphic reduction is shown in (d). The node types here are as follows: R = ring, N = nonring, C = carbon, and H = heteroatom.

druglike compounds[19] by comparing molecules within each activity class in the *MDL Drug Data Report* database. Pairs of molecules that were found to differ in only one region were identified using a maximum common substructure (MCS) procedure, with the different fragments then being regarded as potential bioisosteres. This provides a simple way of identifying possible fragment-pairs but does mean that only groups appearing with a one-to-one substitution in the database can be identified; furthermore, many of the frequent substitutions were found to be both small and relatively "obvious". Substituent replacements also underlie two recent studies, by Haubertin and Bruneau[20] and Leach et al.,[21] that involve detailed analyses of the changes in property resulting from one-to-one chemical replacements. The basic idea here is to suggest substituent replacements that are most likely to bring about some desired change in a property that is to be optimized, but the property data could also clearly be used to identify pairs of potentially bioisosteric substituents. Procedures for the identification of appropriate substituent replacements are described by Sheridan et al.[22]

In addition to approaches making use of 2D and property information, bioisostere identification can also make use of knowledge of 3D interactions. Watson et al.[23] applied geometric similarity measures to data from the IsoStar database[24] to identify substituents that made similar nonbonded interactions, and that might thus form bioisosteric replacements. Kennewell et al. focused on identifying activity class-specific bioisosteric equivalences, using classes of biological activity for which several ligands have been cocrystallized with the same target protein. The aim of this study was to find fragments from different ligands that possessed bound conformations exhibiting substantial steric overlap and that might hence be regarded as bioisosteric for that particular binding site.[25]

Although perhaps not specified in that way, most of the studies to date have sought to answer the following problem: given a specific molecular fragment, how can we find other fragments that are bioisosteric to it. The problem addressed in the study reported here is rather different and, arguably, more challenging: given information about potential bioisosteric replacements, how might we use that information to enhance similarity-based virtual screening. The approach we have adopted is based on the use of reduced graphs, a structure representation that has been the focus of considerable recent interest and that provides a generalization capability that may enable multiple structural types (such as bioisosteric fragments) to be encoded in the same way. The organization of the paper is as follows. Following a description of work to date on reduced graphs in the next section, we describe the graph representation used in our experiments and the matching of pairs of such graphs by a clique-detection algorithm. We then report extended experiments using the BIOSTER and WOMBAT databases, before presenting our conclusions.

REDUCED GRAPHS

Background. Graph reduction involves representing collections of atoms as single entities referred to as *nodes*. A reduced graph (RG) is a simplified version of a conventional chemical graph (CG), consisting of a set of nodes connected by edges and allowing CGs with different structures to be represented by the same RG. There are several types of RG that can be derived from a single chemical structure according to the *reduction scheme*, i.e. the definitions that govern which groups of atoms are reduced to which nodes.

The application of different reduction schemes to the same CG often results in RGs with different structures, as exemplified by the CG and RGs shown in Figure 1. A ring/nonring reduction tends to emphasize the structural topology of the molecule by grouping all adjacent ring atoms into a single ring node type, while all bonded nonring atoms are grouped into a nonring node type. It should be noted that there are two possible ways in which rings could be dealt with in this reduction scheme; fused rings could be represented as a single node (a), or each smallest ring could be represented by an individual node (b). The carbon/heteroatom reduction displayed in (c) groups adjacent heteroatoms into a single node, while grouping adjacent carbon atoms into a different node type. Finally, the homeomorphic reduction displayed in (d) assigns a node for each atom with more than or less than two connections, resulting in an RG that

emphasizes the shape of the CG by effectively encoding branch and terminal points. An important point to note is that in each case in Figure 1 the general topology of the chemical graph is retained, although to differing extents. The retention of topology is important since topology is one of the key factors that defines a molecule's shape, and steric properties are often closely associated with biological activity: an RG hence provides an obvious mechanism for relating different molecules or fragments with related shapes.[26]

For a given reduction scheme it is possible to add further detail to the description by implementing different labeling schemes. For example, the description of the ring/nonring reduction could be augmented by assigning different node types (labels) according to the number of atoms in the ring, e.g. 5-membered or 6-membered. Furthermore, multiple descriptions could be systematically combined in a hierarchical manner to generate RGs at increasing levels of discrimination, as discussed by Gillet et al.[26]

The ability of RGs to permit several different structures to reduce to the same graph is their primary attraction. Thus using RGs instead of CGs for similarity searching allows molecules with different (or even significantly different) structures to be considered similar, which provides an obvious mechanism for scaffold-hopping in the pursuit of novel leads. Further advantages of RGs are their functional description (which is readily interpretable by the chemist) and the availability of a range of node definitions (which permits tuning a search system toward high recall or high precision); their smaller sizes when compared to CGs may also be advantageous for searches based on computationally demanding graph-matching algorithms. There is, however, an obvious disadvantage, *viz.* the loss of structural information inherent in the reduction of multiple distinct CGs to the same RG, information that may be useful in some circumstances.

Applications of Reduced Graphs. The first application of RGs in chemoinformatics was to the representation and searching of Markush structures, which encompass a variety of possible chemical structures in a single representation.[27] The major application of Markush structures is in the filing of patents, although they are also used to give a concise description of any other set of related molecules, such as bioisosteres or molecules present in a combinatorial library. However, the main application of RGs in recent years has been in similarity searching and HTS analysis, for which a number of different RG implementations have been developed.

Perhaps best known is the feature tree, a molecular representation that seeks to generalize chemical structures by emphasizing functional features.[28] A ring/nonring reduction similar to the example in Figure 1(a) is used except that separate nodes are assigned to each nonterminal acyclic atom, with the reduction being carried out such that the resulting structures are trees rather than graphs. This allows significant improvements in speed when calculating their similarity since tree-matching algorithms are much more efficient than graph-matching algorithms. Each node in the tree is labeled with a range of features derived from its constituent atom(s) such as their volume and molecular interaction capabilities. Similarity between two trees is then based on the matching of subtrees with a numerical similarity based on a weighted combination of their feature label similarities. There is an extensive literature on the use of feature trees for ligand-based virtual screening (see, e.g., refs 29 and 30).

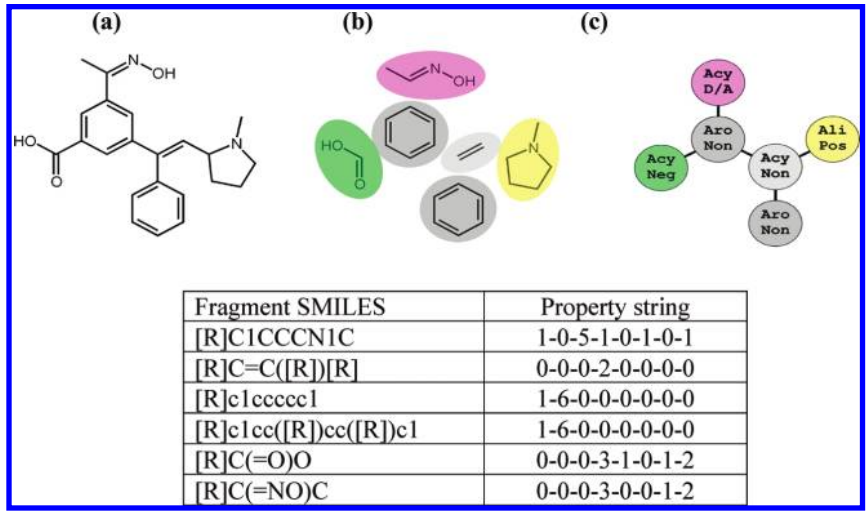
A comparison of the effectiveness of different types of graph reduction for similarity searching was reported by Gillet et al.[26] Daylight-like fingerprints generated from the resulting RGs were used in simulated virtual screening experiments across a range of activity classes. The results showed that it was possible to retrieve different actives from those retrieved by conventional Daylight fingerprints based on CGs, with the RGs often enabling the retrieval of actives that were noticeably different in structure from the query molecules. Subsequent work by Barker et al.[31] considered RG-based fingerprints analogous to atom-pairs[18] and showed that they were generally superior to the Daylight-like fingerprints studied previously, although it was not possible to identify one type that was consistently the best. Harper et al.[32] implemented a variation of the Ar/F(4) RG encoding scheme developed by Gillet et al.,[26] including a more detailed node-labeling scheme and an edit-distance algorithm for measuring the similarity of RGs. This algorithm quantified the degree of similarity of two RGs based on the number and type of operations required to convert one graph to the other; subsequent work by Birchall et al. described the use of a genetic algorithm (GA) for training the weights of the different edit-distance operations, this resulting in a substantial increase in screening performance.[33]

The extended reduced graph (ErG) approach was developed by Stiefl et al. to improve the likelihood of scaffold hopping in similarity searches.[34] The approach exploits the reduced reliance on structural similarity offered by RGs in several ways: by encoding molecular features that are commonly involved in mediating drug-target interactions (such as H-bonding, charged and hydrophobic features); by an enhanced coding method for rings that better conserves positional information; and by using holograms that code the frequencies of occurrence, rather than just incidence as in conventional binary fingerprints. Simulated virtual screening experiments showed that while the RGs were comparable in performance to Daylight fingerprints, they tended to retrieve more diverse sets of actives across a range of activity classes. Based on knowledge of the drug-target interaction gained from experimental data, Stiefl and Zaliani subsequently described a weighting scheme for those RG features that could form similar interactions with the target and found that this improved the screening performance of ErGs.[35]

Drawing on earlier work by Takahashi et al.,[36] Gardiner and co-workers reported the matching of RGs directly by graph-matching algorithms, rather than indirectly by matching fingerprints derived from those RGs. Their studies involved virtual screening experiments that demonstrated clearly the scaffold-hopping capabilities of graph-matching[31] and the use of RGs to summarize the centroids of clusters of CGs.[37] Finally, Birchall et al. have recently described the use of a multiple-objective GA to evolve multiple structure−activity relationships in a form that is readily interpretable by a chemist. The relationships are encoded as RG queries describing features that are preferentially present in active compounds compared to inactives and are applicable to heterogeneous data sets such as those obtained from HTS experiments.[38,39]

METHODS

The basic approach that we have adopted in the work to be described here involves the following main stages:



Fragment SMILES	Property string
[R]C1CCCN1C	1-0-5-1-0-1-0-1
[R]C=C([R])[R]	0-0-0-2-0-0-0-0
[R]c1ccccc1	1-6-0-0-0-0-0-0
[R]c1cc([R])cc([R])c1	1-6-0-0-0-0-0-0
[R]C(=O)O	0-0-0-3-1-0-1-2
[R]C(=NO)C	0-0-0-3-0-0-1-2

Figure 2. Generation of a reduced graph representation. Cleaving all permitted cuttable bonds in the chemical structure (a) yields the fragments indicated in (b). These are reconnected and assigned a node type as indicated by the abbreviated labeling in (c). Colors are also used here to aid in the recognition of commonality between the nodes in the different representations. The table lists the fragment SMILES and the property counts for each of the six RG nodes.

creating RG representations from a set of molecules using a fragmentation procedure; matching pairs of RGs using a clique-detection algorithm; and using the resulting matches for similarity-based virtual screening. These stages are described below.

Creation of Reduced Graphs. Generating RGs involves partitioning a chemical structure into discrete fragments such that each fragment is represented by a single node. The resulting set of nodes can then be labeled with one or more fragment-derived properties and finally reconnected to maintain the topology of the original chemical structure.

The fragmentation procedure used here simply involves recursively cutting all nonterminal, acyclic single bonds with the following three exceptions to ensure that chemically sensible and interesting fragments result: acyclic sp3 carbon to acyclic sp3 carbon bonds; acyclic heteroatom to acyclic heteroatom bonds; and acyclic heteroatom to acyclic sp2 carbon bonds. A more complex procedure has been implemented at Sanofi-Aventis since the work reported here: this new procedure contains a total of 21 rules coded as SMARTS, e.g., acyclic heteroatom to acyclic sp2 carbon bonds are now cut when the carbon is involved in an enamine (-C=CN) or a hydrazone (-C=NN).

The reduction procedure used in our experiment shares many similarities with the modified Ar/F(4) reduction scheme used by Harper et al.[32] For example, rings and functional groups remain intact, rings are divided into aromatic and aliphatic types, terminal feature atoms are not separated from rings, and positively and negatively ionizable groups are given priority over hydrogen-bonding features (which are classified into donor, acceptor, and joint donor/acceptor node types). The principal difference is that our encoding results in a single node for a fused ring system (although the multinode encoding of fused ring systems was also investigated as discussed later in the paper).

After the cuttable bonds have been cleaved, each of the resulting fragments corresponds to a node in the RG for which the node type is assigned based on the properties of the fragment (as described in the next paragraph). The nodes are then reconnected in the same configuration as the fragments in the original chemical structure, resulting in a RG molecular object that can be handled by conventional chemical software. The processing is illustrated in Figure 2, with the RG nodes each being associated with a string encoding its property count: this string is used in deriving the node type and also in restricting the graph matching between RGs (as discussed further below). The values reading from left to right in the second column of the table encode the number of rings, aromatic atoms, aliphatic atoms, acyclic atoms, negatively ionizable atoms, positively ionizable atoms, donor atoms, and acceptor atoms. This numeric information is used to define the types of the nodes.

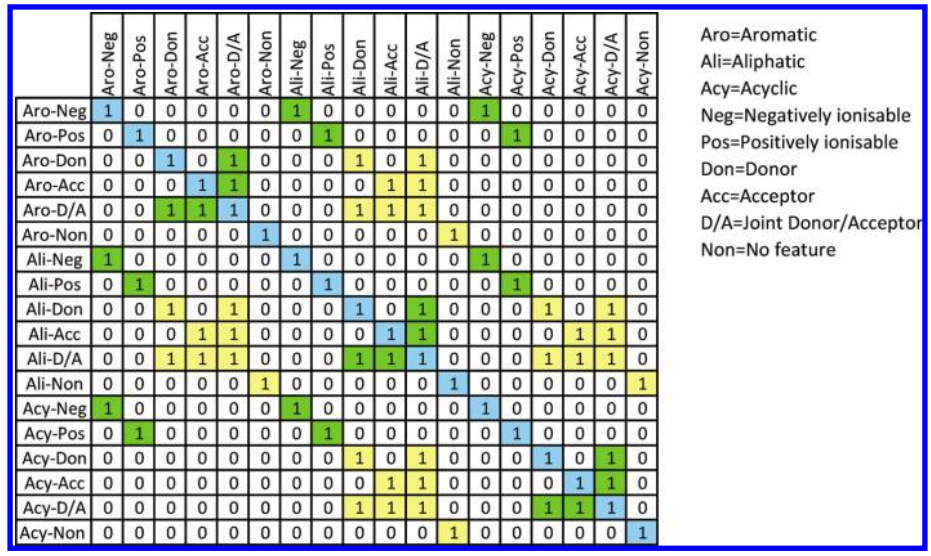
Table 1. Types of RG Node

types of RG node	
aromatic negatively ionizable	aliphatic donor
aromatic positively ionizable	aliphatic acceptor
aromatic joint donor−acceptor	aliphatic featureless
aromatic donor	acyclic negatively ionizable
aromatic acceptor	acyclic positively ionizable
aromatic featureless	acyclic joint donor−acceptor
aliphatic negatively ionizable	acyclic donor
aliphatic positively ionizable	acyclic acceptor
aliphatic joint donor−acceptor	acyclic featureless

When assigning a node type to a fragment, there may be atoms of different types present and there hence needs to be a system of prioritizing which node type should be assigned. Two types of object are considered: structure types and feature types. The following order of priority is applied for structure type: aromatic > aliphatic > acyclic.

The following order of priority is applied for feature type: negatively ionizable > positively ionizable > joint-HBD and HBA > HBD or HBA > featureless, with these features being based on the definitions used in Pipeline Pilot and with the featureless classification indicating the absence of atoms with charge or hydrogen-bonding capability. Combining each of the three structure types with each of the six feature types results in the 18 distinct node types listed in Table 1.

Matching of Reduced Graphs Using Clique-Detection. The similarity between a pair of reduced graphs is computed using an MCS procedure that identifies the number of matching nodes; this number is then used to compute the



	Aro-Neg	Aro-Pos	Aro-Don	Aro-Acc	Aro-D/A	Aro-Non	Ali-Neg	Ali-Pos	Ali-Don	Ali-Acc	Ali-D/A	Ali-Non	Acy-Neg	Acy-Pos	Acy-Don	Acy-Acc	Acy-D/A	Acy-Non
Aro-Neg	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
Aro-Pos	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
Aro-Don	0	0	1	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0
Aro-Acc	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0
Aro-D/A	0	0	1	1	1	0	0	0	1	1	1	0	0	0	0	0	0	0
Aro-Non	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
Ali-Neg	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
Ali-Pos	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
Ali-Don	0	0	1	0	1	0	0	0	1	0	1	0	0	0	1	0	1	0
Ali-Acc	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0	1	1	0
Ali-D/A	0	0	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0
Ali-Non	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
Acy-Neg	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
Acy-Pos	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
Acy-Don	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	1	0
Acy-Acc	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	0
Acy-D/A	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	1	0
Acy-Non	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1

Figure 3. Matching of node types during clique-detection. A "1" (or "0") in this compatibility table indicates that the indexed node types are permitted (or are not permitted) to match. Three sets of parameters were investigated in this study: the "identical" set of parameters only allowed identical node types to match (shown in blue); the "close" set of parameters additionally allowed closely related node types to match (shown in green); the "related" set of parameters additionally allowed a number of more distantly related node types to match (shown in yellow).

Dice coefficient. Given two RGs containing *A* and *B* nodes, respectively, and an MCS containing *C* nodes, then the Dice coefficient is given by

$$\frac{2C}{A + B}$$

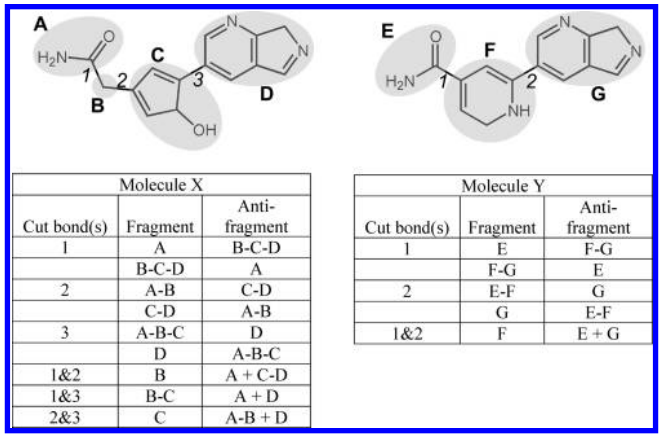
The common nodes in the MCS are identified using a version of the Bron-Kerbosch clique-detection algorithm that has been developed for matching RGs.[31] Each RG is described by a distance matrix in which the *XY*-th element contains the number of edges in the RG separating the *X*-th and the *Y*-th nodes. A correspondence matrix is then defined where the *IJ*-th element is set to unity if the *I*-th pair of nodes from the first graph (call these *P* and *Q*) can be mapped to the *J*-th pair of nodes from the second graph (call these *R* and *S*). A mapping is allowed if two conditions hold (these conditions being checked using the information in the distance matrices describing the two RGs that are being compared): the first condition is that *P* and *Q* are separated from each other by the same number of edges as *R* and *S* are separated from each other; the second condition is that the node types of *P* and *Q* are compatible with (and can hence be mapped to) the node types of *R* and *S*.

Compatible node types are detailed in Figure 3, where it will be seen that we have allowed three levels of matching. The obvious way is to allow nodes to match only if they are of the same type ("identical"), but it is also possible to allow related node types to match, for example, aliphatic positively ionizable with acyclic positively ionizable. Less restrictive matching may enable the retrieval of more diverse structures, and we have thus allowed two levels of partial matching - "close" and "related" - as detailed in Figure 3. Matching nodes must not only be of compatible types but also be of comparable size (in terms of the number of heavy atoms comprising the node), particularly as fused rings are normally encoded in a single node. Accordingly, two nodes are not permitted to match if they differ in size by more than five atoms or the size of the smaller of the two nodes (whichever is the lesser; the minimum node size is one atom): the necessary size information is obtained from the property strings shown in Figure 2. The second, alternative condition arises from the fact that there is a much greater relative mismatch if the nodes are, e.g., of sizes two and seven than if they are of sizes five and ten.

Processing of the BIOSTER Database. Our initial experiments involved applying our fragmentation and graph-reduction methods to structures from the BIOSTER database (available from Accelrys Software Inc. at http://www.accelrys.com).

BIOSTER encodes pairs of molecules that have been reported in the literature as having the same biological activity and that differ in the presence of a particular structural fragment. Each such pair of molecules was fragmented using the bond-cleavage procedure described in the previous section. Each of the pair of fragments produced on cutting a bond is considered both as a *fragment* and as an *antifragment*, where the antifragment is defined as the remainder of the molecule following excision of the fragment. Consequently, while the antifragment may consist of multiple disjoint fragments, the fragment is always a single continuous substructural moiety. Fragments with less than three atoms are typically too small to be of interest, and any such fragment pairs are not recorded for output.

For a pair of molecules from BIOSTER with the same activity, a pairwise comparison is made between the two sets of antifragments by comparison of the corresponding canonical SMILES strings. Once a pair of identical antifragments has been found, three conditions must be met before the remaining fragments are regarded as putative bioisosteres. First, the fragment (variable part) must be smaller than the antifragment (constant part) so as to ensure that the antifragment is the major factor accounting for the retention of activity between two molecules. Second, the fragments must differ in size by no more than five atoms. For example, even for molecules with the same binding mode it may be possible to replace a fragment with a much larger fragment where



Molecule X		
Cut bond(s)	Fragment	Anti-fragment
1	A	B-C-D
	B-C-D	A
2	A-B	C-D
	C-D	A-B
3	A-B-C	D
	D	A-B-C
1&2	B	A + C-D
1&3	B-C	A + D
2&3	C	A-B + D

Molecule Y		
Cut bond(s)	Fragment	Anti-fragment
1	E	F-G
	F-G	E
2	E-F	G
	G	E-F
1&2	F	E + G

Figure 4. Extraction of fragments from a pair of BIOSTER molecules. The accompanying tables give the combinations of fragments that result from cutting bonds in the above molecules. Comparing all the antifragments in molecule X to those in molecule Y identity is first found between A and E (although this would not be considered since the fragment is larger). The procedure continues to find the largest valid antifragment identity, that between A+D and E+G such that B−C and F are isolated as the variable fragments.

the fragment simply protrudes into the binding pocket rather than interacting with the receptor: such replacements would be highly context dependent and cannot be regarded as being truly bioisosteric. Third, the fragments must be 3−15 atoms in size, since chemical replacements outside this size range are likely to be of little interest to the medicinal chemist.

The procedure seeks to maximize the size of the antifragment identity between a pair of molecules, so as to ensure that the single smallest possible region of variability is identified per pair of molecules. The operation of the procedure is shown in Figure 4. The cuttable bonds in the two molecules are labeled with numbers and the fragments with letters. Fragments consisting of a number of connected smallest fragments are indicated by concatenation with "−" such that A-B-C-D is the whole molecule. A set of fragments that is not connected is indicated by a "+". 2425 pairs of putative bioisosteres were extracted from the BIOSTER database using the procedure shown in Figure 4. Graph reduction was then applied to the resulting pairs of fragments to ascertain whether the reduction had been able to encode bioisosteric fragments (as indicated by their presence in the BIOSTER database) with the same RG (as generated using our bond-cleavage and node-labeling routines).

Processing of the WOMBAT Database. The simulated virtual screening experiments were carried out using Version 2007.1 of the World Of Molecular Bioactivity database (WOMBAT) from Sunset Molecular Discovery LLC (at http://sunsetmolecular.com/products/?id=4).

WOMBAT contains activity information extracted from the medicinal chemistry literature on a large range of targets. Six activity classes were chosen for investigation based on the amount of available data in WOMBAT and on their molecular diversity and chemical interest. For consistency, in each activity class only records with IC_{50} data were examined, and only pure numerical values (e.g., excluding entries such as ">4") with a minimum pIC_{50} value greater than or equal to 5 were selected. Only the single species for which there was the largest amount of data was selected: "rat" for 5HT1A, "human immunodeficiency virus type 1" for HIV1-P, and "human" for the remaining four classes. These six classes are listed in Table 2, together with the number of active molecules and the number of distinct ring scaffolds (as defined by the Murcko frameworks) for each class.

Table 2. Activity Classes Used in the Experiments, with Their Numbers of Active Molecules and of Distinct Scaffolds (As Defined by the Murcko Frameworks)

activity class	actives	scaffolds
5HT1A	588	241
AChE	446	198
fXa	823	346
HIV1P	749	303
MMP1	683	287
PDE4	413	212

The WOMBAT database was filtered to remove the following: molecules containing atoms other than C, H, N, O, P, S, and halogen; molecules having more than 50 heavy atoms or more than 20 nonterminal rotatable bonds or more than one fragment after stripping salts in Pipeline Pilot; and molecules with missing data in any of the fields of interest (SMILES, species, target, and IC_{50}). The filtered database was then further refined by the removal of all molecules belonging to any of the six chosen classes and by the removal of molecules that could potentially be cross-reactive with any of these classes (e.g., many of the 5HT1A actives have also been found to be active at the 5HT2A receptor and so any 5HT2A actives were also removed). The remaining molecules were then input to a selection procedure that produced a set of 10,000 inactives for which the physicochemical properties (in terms of the distributions of molecular weight, ClogP, number of rotatable bonds, number of H-bond donors, and number of H-bond acceptors) matched as closely as possible the observed distributions for the set of actives pooled across the six activity classes. The selection procedure involved comparing the binned active and inactive property profiles, with adjustments in the number of compounds in each bin being made using random selection employed in a stepwise manner. Manual supervision and refinement of selection frequency and bin size was used to maintain balance between the different property profiles.

Ten reference structures were randomly selected from each activity class to search through a data set consisting of the 10K inactives together with the remaining actives, with the similarities being computed using the clique-detection process described above. The overall similarity of each molecule in the 10K WOMBAT data set was then taken as its maximum similarity over all of the ten reference structures, i.e., a group fusion approach was used.[40] The data set was ranked in descending similarity order with the top N molecules being taken as the hitlist (where N is the number of actives in the activity class, minus the 10 that are used as queries); the proportion of the data set actives found in the hitlist (recall) and number of different scaffolds retrieved in the hitlist was then calculated. For each activity class, ten different searches were carried out in this way.

Thus far, the search protocol is entirely conventional, mirroring many previous studies of similarity-based virtual screening. However, an additional search criterion was adopted to ensure that the search focused on scaffold-hopping, rather than on the retrieval of simple analogues of

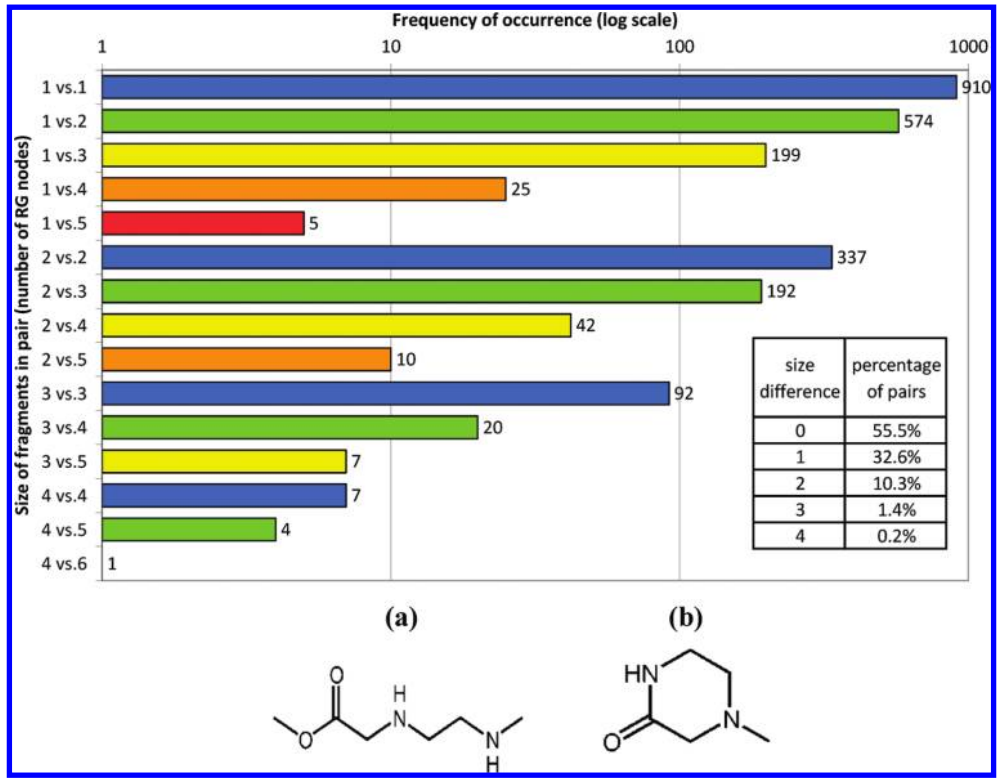


Figure 5. The plot depicts how often pairs of putative bioisosteres extracted from the BIOSTER database are encoded as RGs of the same size (blue bars), compared to different sizes. The inset table provides a summary of the percentage of the pairs that exist with a particular size difference. The structures shown at (a) and (b) are an example of one of the few cases where the RGs have a large size difference.

the reference structures, which is a known, common problem with 2D similarity measures. Specifically, similarity was only calculated between reference structures and data set structures that did not have the same scaffold. This was done in an effort to minimize the retrieval of molecules belonging to the same chemical series as the reference structure (since these are typically of less interest in the operational context of lead-hopping) and hence to maximize the opportunities for scaffold-hopping. For comparison with the RG results, some of the searches were repeated using Pipeline Pilot FCFP_4 fingerprints.

RESULTS AND DISCUSSION

Fragmentation of the BIOSTER Database. As described in Methods, our initial experiments used the BIOSTER database to ascertain whether the RG nodes that we identified could, in fact, be used to match fragments that the database suggests are bioisosteres.

Figure 5 shows the number of times that an RG of size N nodes is found to be bioisosteric with a fragment encoded by an RG of size M nodes. A summary of the percentage of fragment pairs encoded by RGs that differ in size by X nodes (i.e., $X = |M\text{-}N|$) is given in the accompanying table. It will be seen that a substantial proportion of pairs are encoded by RGs of the same (blue bars in the figure) or similar size, suggesting that the reduction scheme used here is quite reasonable in operation. That said, there are a very few instances where the difference in size between the RGs is surprisingly large given the restriction that the fragments must differ in size by no more than five atoms. One example of such a pair is shown in the lower part of the figure where the corresponding RGs are five (a) and one node (b) in size, and where fragment (b) is essentially a cyclized variant of (a).

As shown in Figure 5, the largest group (~38% of total pairs) comprised those encoded by a pair of single RG nodes. Graph matching provides the possibility of allowing different node types to match, and it is hence of interest to find how frequently particular node types are found to be equivalent and hence which node types are co-occurring more frequently than would be expected if the node occurrences were statistically independent of each other. Figure 6 displays the color-coded enrichment factors calculated in this manner. Unsurprisingly the highest values tend to occur along the diagonal, i.e. where the node types are the same, with ~46% of the pairs involving identical node types. The majority of other high values occur where the node types are related, e.g., where a fragment encoded as an aliphatic featureless node is found to be bioisosteric with a fragment encoded as an acyclic featureless node. This equivalence information can be used to determine which node types should be permitted to match during graph matching, which can be achieved simply by applying a cutoff to the enrichment factor values: the information in this matrix helped in the identification of the "close" and "related" matching categories noted in Figure 3. Alternatively, instead of using the matrix to define which node types are permitted to match during graph matching, the relative frequency of co-occurrence could be interpreted as the likelihood that a pair of nodes should be regarded as equivalent. This could then be used to determine a node similarity score rather than a binary match/nonmatch; a similar concept to the amino acid similarity matrices that are widely used in bioinformatics.[41] This was, however, not done in our experiments in view of the very limited amount of data available (just 910 pairs for the 324 different node pairs possible with the 18 different node types used here and listed in Table 1).

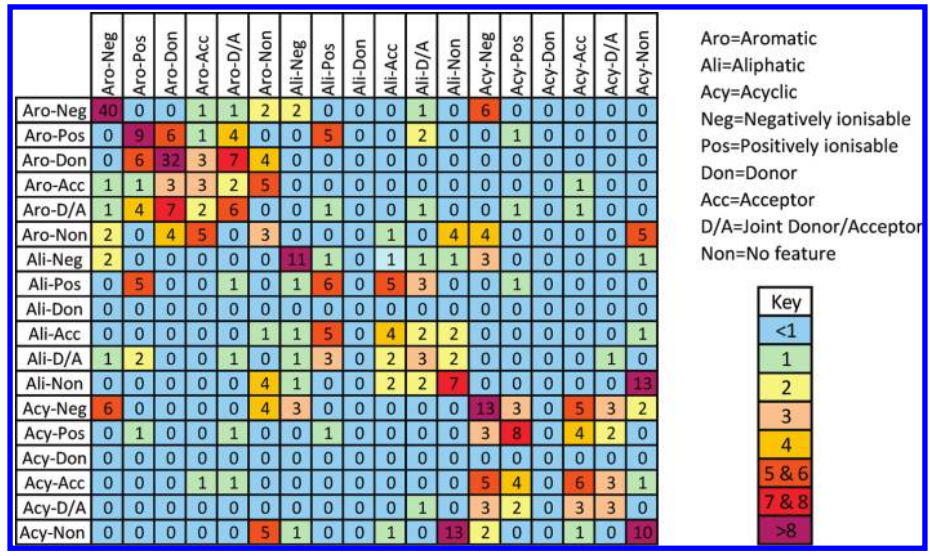


Figure 6. The matrix presents the number of fold greater than random that the indexed node types were found to occur in the putative bioisosteric pairs extracted from the BIOSTER database. The values are rounded down to the nearest integer and color coded at the cut-offs indicated in the key.

Table 3. Percentage Recall of (a) Actives and (b) Scaffolds Using FCFP_4 Fingerprints and the Basic Version of the RG Search

activity class	FCFP_4	identical	close	related
		(a) Actives		
5HT1A	74.9	59.6	59.1	59.7
AChE	62.4	36.7	37.2	36.4
fXa	67.3	58.6	58.5	56.9
HIV1P	63.0	56.2	56.5	55.1
MMP1	67.9	58.4	58.9	58.4
PDE4	64.9	49.9	49.8	50.9
mean	66.8	53.2	53.3	52.9
		(b) Scaffolds		
5HT1A	73.9	59.0	58.7	59.9
AChE	62.2	45.5	46.1	45.4
fXa	66.2	57.3	57.5	56.7
HIV1P	63.5	58.0	58.5	57.5
MMP1	69.7	63.4	63.6	63.1
PDE4	66.7	54.6	54.6	56.1
mean	67.0	56.3	56.5	56.4

These experiments have hence demonstrated that our fragmentation and label-assignment procedures can identify pairs of fragments that are regarded as bioisosteric in the BIOSTER database.

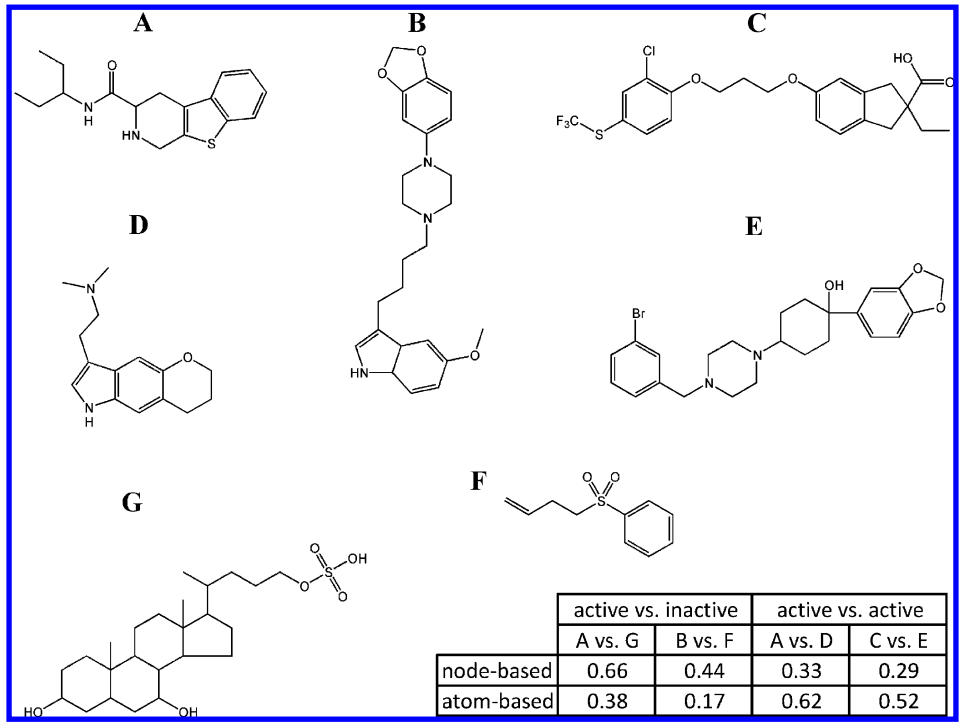
Initial WOMBAT Searches. Table 3 lists the initial results obtained using the protocols detailed in the previous section: it will be seen that there are only slight performance differences between the three RG matching criteria ("identical", "close" and "related"), and hence, for simplicity, only the basic "identical" criterion was used in the subsequent searches. The results also show that the RG results are consistently inferior to the fingerprint searches, although the overall differences are marginally less in the scaffold searches of Table 3b. A comparison of the actives retrieved in the RG and fingerprint searches showed that both methods retrieved unique structures, i.e., molecules that were not retrieved by the other method: combining the results of the two searches would on average improve recall by about 5% across the different activity classes.

A potential problem when calculating the similarity between RGs is that there is a loss of sense of structural size. For example, it is possible for one node to encode a small monocycle while another node encodes a complex fused ring system containing many atoms, and yet these two nodes may be matched at search time. We have sought to alleviate this problem in two ways, as detailed below (see also the Section "Splitting of fused rings" later in the paper).

First, the assignment of a RG node type is absolute and takes no account of the difference in the number of features that may be present within a particular fragment. Consequently, fragments with relatively large differences in polarity/hydrophobicity may still be assigned the same node type and hence permitted to match during the graph matching. This is particularly important for fused ring systems where there is a greater scope for feature variation due to their larger size, with an increased possibility of false positives being retrieved. Accordingly, restrictions were placed on the difference in feature counts such that a pair of H-bonding nodes with a difference of more than two donor or acceptor atoms was not allowed to match.

Second, the overlay of a pair of nodes each representing a single atom contributes the same amount to the overall similarity as a pair of nodes representing much more significant portions of the structure, such as large ring systems. Accordingly, an atom-based similarity measure was devised, in which each node makes a contribution to the similarity based on the number of heavy atoms contained in the fragment represented by that node. Each pair of matched nodes is compared in size; the size of the smaller of the two nodes is added to the numerator in the calculation, while the size of the larger of the two nodes is added to the denominator. The number of atoms in the unmatched nodes is then calculated for each structure, and the larger of the two values is added to the denominator to reflect the increasing dissimilarity for structures that have larger mismatched regions. Thus, if we denote two matched nodes by A and B, containing a and b atoms, respectively, and two unmatched nodes by C and D, containing c and d nodes, respectively, then the similarity is computed as

$$\frac{\sum \min\{a,b\}}{\sum \max\{a,b\} + \sum \max\{c,d\}}$$



	active vs. inactive		active vs. active	
	A vs. G	B vs. F	A vs. D	C vs. E
node-based	0.66	0.44	0.33	0.29
atom-based	0.38	0.17	0.62	0.52

Figure 7. Effect of using node-based and atom-based similarities for 5HT1A reference structures (structures A, B, and C) and other actives (structures D and E) and inactives (structures F and G).

In effect, larger matched nodes contribute more to the similarity than do smaller matched nodes, with similarly sized nodes resulting in a larger contribution than nodes with a larger difference in size.

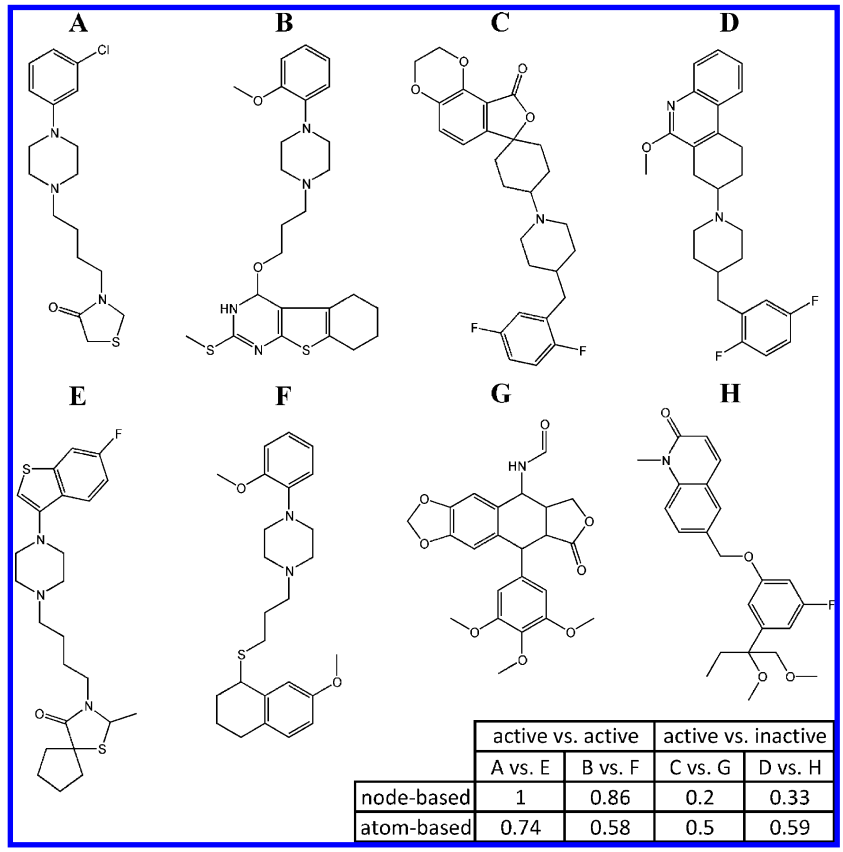
Figure 7 illustrates several real examples where the atom-based similarity gives an arguably better reflection of the intuitive similarity than the previous, node-based similarity, both for active−active and active−inactive pairs of structures. That said, there are also cases where the use of the atom-based similarity would appear to give a worse reflection of the intuitive similarity than the node-based similarity, both for active−active and active−inactive pairs of structures. Examples are shown in Figure 8 where, as can be seen in pairs A vs E and B vs F, the penalty for mismatches in size between nodes increases the stringency of a search, which can sometimes be detrimental. The fact that the atom-based similarities are higher than the node-based similarities in pair C vs G to some extent reflects the fact that a large similarity is given to the large fused ring systems. However, for both pairs C vs G and D vs H it is perhaps more the case that the node-based similarity is relatively low as a result of the high number of small nodes, e.g. the methoxy groups in structure G.

Neither of these procedures - restrictions on H-bond matching and atom-based similarity - resulted in an overall increase in performance, improving the mean recall for some activity classes but worsening it for others, but it was decided to include both of them in the subsequent experiments.

Incorporation of Bioisostere Information. While the initial BIOSTER runs showed that the RG nodes were at least partially successful in conflating bioisosteres, i.e., by assigning the same node type to bioisosteric fragments, bioisostere information has not been used explicitly thus far in the matching procedure. This can be effected by allowing nodes represented by known bioisosteric fragments, identified by a simple SMILES comparison, to match regardless of their node type or other restrictions. The fact that the graph reduction and bioisostere extraction procedures are based upon the same fragmentation scheme helps ensure that the RG node and bioisostere fragment SMILES have the necessary textual correspondence in terms of cut-points (e.g., rings intact) and R-atoms etc. However, many of the putative bioisosteres are encoded by multiple nodes that would not have such correspondences. Consequently, the simplest case considered initially is where a fragment represented by a single node is putatively bioisosteric with another fragment also represented as a single node. The latter part of this section explores the more complex question of how multi-node bioisostere information can be handled.

Single-Node Bioisosteres. 910 pairs of single-node bioisostere-pairs were extracted from the BIOSTER database. To gauge the likely benefit of incorporating such information, the set of fragments contained within each activity class was examined to see if they contained any known bioisosteres present in this set. Table 4 presents the total number of fragments contained within the molecules of each activity class (including the inactives) and the number of these that are unique. The percentage of these that are identical to one of the 436 unique fragments present in the set of 910 single-node bioisostere pairs is also given, along with the percentage of molecules in each class that contain one or more of the known bioisostere fragments.

The data in Table 4 show that on average 42% of the fragments in the activity classes have one or more single-node putative bioisosteres, indicating that the bioisosteric equivalences extracted from BIOSTER are relevant to the chemical space covered by the six WOMBAT activity classes. Furthermore, the large percentage of molecules that contain at least one node with a known equivalence is very high, which suggests that there is significant potential for making use of the bioisostere information. However, it is worth noting that the number of unique fragments in each



	active vs. active		active vs. inactive	
	A vs. E	B vs. F	C vs. G	D vs. H
node-based	1	0.86	0.2	0.33
atom-based	0.74	0.58	0.5	0.59

Figure 8. Effect of using node-based and atom-based similarities for 5HT1A reference structures (structures A-D) and other actives (structures E and F) and inactives (structures G and H).

Table 4. Number of Fragments Contained within the Molecules of Each Activity Class and the Percentage of These That Are Identical to One of the Known Bioisosteric Fragments (BI)[a]

activity class	molecules	fragments	unique fragments	% molecules with known BI	% fragments with known BI	% unique fragments with known BI
5HT1A	588	3070	322	92	45	17
AChE	446	2557	273	82	33	18
fXa	823	6332	322	100	50	21
HIV1P	749	6842	322	99	53	20
MMP1	683	5172	307	100	48	21
PDE4	413	2388	227	77	25	24
inactives	10000	71777	3876	91	38	5

[a] The percentage of molecules in each class that contain one or more of these fragments is also given.

class is very much smaller than the total number of fragments, i.e. there is a large amount of redundancy, with fragments such as amide or phenyl tending to occur very frequently: it is thus not surprising that most molecules contain one of these fragments for which bioisosteres are known. It is also notable that due to the generality of such fragments, a large percentage of the inactives contain fragments with known bioisosteric equivalences, an important fact to which we shall return later in the paper.

The results of including this single-node bioisostere information in RG-based similarity searches of the six WOMBAT activity classes is presented in Table 5 as the change in recall compared to the results previously obtained in the original RG-based similarity searches. It is clear from the data in the table that the application of single-node bioisostere information has had little or no impact upon the results.

There are several possible explanations for these rather disappointing results. First, although the prevalence of known bioisosteres in the set of nodes contained within each activity class is an indicator of how relevant the bioisosteric fragments are to an activity class, it does not answer the question of how often such equivalences can be made use

Table 5. Percentage Change in Recall When Carrying out Graph Matching Making Use of Known Single-Node Bioisosteric Equivalences Compared to Graph Matching without Making Use of Such Information[a]

activity class	mean	minimum	maximum
5HT1A	0.6	−0.5	2.4
AChE	−0.6	−1.8	0.0
fXa	−1.1	−3.4	0.0
HIV1P	0.5	−2.0	2.8
MMP1	0.6	0.1	2.4
PDE4	0.0	−0.5	0.5

[a] Results are presented as the mean, minimum, and maximum paired differences observed when using the 10 different sets of reference structures.

Table 6. Percentage of Reference-Active and Reference-Inactive Comparisons Where the Overlay Makes Use of the Single-Node Bioisostere Information and Percentage of These Pairs Where the Bioisosteric Fragments Are of the Same Node Type

activity class	% reference-active pairs where the best overlap contains known single-node BI fragment pair(s)	% reference-inactive pairs where the best overlap contains known single-node BI fragment pair(s)	% bioisosteric fragment pairs identified that have the same RG node type
5HT1A	17.4	9.4	43
AChE	1.0	4.9	27
fXa	8.9	14.9	45
HIV1P	24.7	13.8	27
MMP1	23.9	20.4	72
PDE4	4.2	6.1	30

of in a similarity search. In order to be useful, known bioisosteric fragments must co-occur, i.e. one fragment must be present in the reference structure while one of its known bioisosteres must be present in a database structure. Second, there is no guarantee that the equivalence may be of any extra benefit since the fragments may be encoded as the same RG node type and would have matched anyway (see Figure 5). Third, the presence of known bioisosteric fragments in a pair of molecules does not necessarily mean that these fragments would be overlaid when the molecules are compared using the graph-matching algorithm, since their overlay may not give the best alignment of the other features in the remainder of the molecules. Fourth (and most importantly as discussed below), consideration must be given to the extent to which matches are found between the reference structures and the inactives in the search file since, as indicated in Table 4, there is a very considerable chance that equivalences may be found in the inactives (of which there are many, many more than there are actives). Consequently, it is possible that just as the bioisostere information may help improve the reference structure's similarity to the actives, it may also improve its similarity to the inactives. The overall change in the hitlist is then dependent on how many actives have their similarity increased compared to how many inactives have their similarity increased.

In light of these considerations, a more detailed analysis was carried out to determine the number of cases where the best overlay between the query and reference RGs made use of one of the known single-node bioisosteric equivalences. Counts were made separately for when the reference structure was being compared to an active and for when it was being compared to an inactive. The results of this analysis are presented in Table 6 as percentages of the total number of reference-active and reference-inactive pairs. The table also presents the percentage of the bioisosteric equivalences found within the reference-active and reference-inactive pairs that are encoded as the same node type. The results in the table show clearly that although a high proportion of the actives in each class contain a fragment that is identical to a known single-node bioisostere, the requirement for co-occurrence of bioisosteric fragments means that the proportion of pairs of molecules that contain an equivalence is substantially lower. The two figures are not even directly related, for example, 100% of the actives in the fXa class contained a fragment that is identical to one of the single-node bioisosteres, yet only 8.9% of reference-active pairs contained a single-node equivalence, whereas 92% of the actives in the 5HT1A class contained a fragment that is identical to one of the single-node bioisosteres, with 17.4% of reference-active pairs containing a single-node equivalence. Performance is not just dependent on the number of equivalences found with the actives but also on the number of equivalences found with the inactives. It must be remembered that there are far more reference-inactive pairs than there are reference-active pairs: ten reference structures are used and there are 10,000 inactives, giving a total of 100,000 reference-inactive pairs, as against, e.g., 5780 reference-active pairs for the 5HT1A class. This difference of scale may help to explain why the fXa class with a moderate percentage of reference-active equivalences and a high percentage of reference-inactive equivalences ends up with a net decrease in performance. The final factor to consider is the extent to which the bioisosteric equivalences are the same node type and would thus have been matched without the inclusion of bioisostere information (see the right-hand column of Table 6). For example, the high value for the MMP1 class (72%) is why there is little change in the performance despite the large number of equivalences in the actives and inactives.

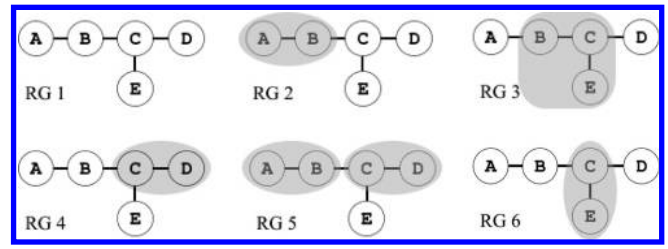


Figure 9. Processing of RGs to allow multinode bioisostere information to be easily incorporated into the graph matching. Shading is used to indicate which nodes are merged into a single node. Given that fragments represented by the nodes AB, CD, and BCE have known bioisosteres, the merged variants RG2 and RG4 would be output in addition to the original, RG1. RG3, RG5, and RG6 would not be generated (see text).

Multiple-Node Bioisosteres. Having demonstrated that inclusion of information regarding single-node bioisosteres had little effect on performance, we then investigated whether including information about multiple-node bioisosteres could do any better. This requires a more sophisticated approach to the coding and matching of the fragments; specifically, a procedure is required to partition the RGs such that fragments normally represented by multiple nodes can now be encoded within a single node.

This problem is addressed in feature trees using the procedures described by Rarey and co-workers;[28,29] the procedure adopted here is illustrated in Figure 9. Considering RG1 in the figure: if the fragment represented by AB has known bioisosteres in the BIOSTER data, then a "merged variant" (RG2) is generated whereby nodes A and B are represented by a single node as indicated by the shading. Furthermore, if the fragment represented by CD has known bioisosteres, then the merged variant RG4 would also be

Table 7. Percentage Recall of (a) Actives and (b) Scaffolds Using the Original RG Search and Including Both Single- and Dual-Node Bioisosteric Equivalences

activity class	original	single-node	multinode
	(a) Actives		
5HT1A	53.2	53.8	53.7
AChE	40.5	39.9	40.2
fXa	57.6	56.5	57.7
HIV1P	62.9	63.5	67.4
MMP1	53.0	53.6	56.5
PDE4	49.8	49.8	50.3
mean	52.8	52.9	54.3
	(b) Scaffolds		
5HT1A	55.1	55.9	56.2
AChE	43.3	42.7	44.1
fXa	53.6	52.8	55.8
HIV1P	63.8	64.9	77.2
MMP1	56.3	57.1	63.8
PDE4	50.8	50.8	52.2
mean	53.8	54.0	58.2

output. The merged variant RG5 would not be generated since for simplicity only one merged node per molecule is currently permitted. Furthermore, also for simplicity only pairs of nodes are considered for merging, and hence the merged variant RG3 would not be generated. Finally, merged variants are only generated for known bioisosteric fragments, i.e. if there are no known bioisosteres of the fragment CE, then the merged variant RG6 will not be generated. Despite these restrictions, the procedure enabled the exploitation of a substantial proportion of the known bioisostere information: 574 pairs where a fragment represented by a single node is found to be bioisosteric with a fragment represented by two nodes, and 337 pairs where a fragment represented by two nodes is found to be bioisosteric with another fragment represented by two nodes.

The recall results from similarity searching in the six WOMBAT activity classes making use of both the single and multinode equivalences are presented in Table 7. The results obtained when using the standard graph matching (only identical node types permitted to match) are included for comparison. The tables show that recall of actives is improved through the incorporation of multinode bioisostere information with an average value of 1.4% over the six activity classes compared to the use of single-node bioisostere information. However the main benefit is an increase in the diversity of the search outputs, with the number of scaffolds increasing by an average of 4.2% over the six activity classes.

The percentage of the reference-structure and database-structure pairs containing a multinode equivalence was counted for the actives and inactives and is displayed in Table 8. For all of the activity classes, the great majority of the cases where there was a multinode equivalence gave rise to an increased similarity, with about half of these increased similarities resulting from the overlaying of the multinode bioisosteres by the clique-detection procedure. A comparison of Tables 6 and 8 shows that the percentage of pairs that contain a multinode equivalence is much less than the percentage of pairs that contain a single-node equivalence (this is to be expected given that multinode fragments are by definition larger and therefore less likely to be found by chance). Even so, they can improve considerably the search performance. This is most clear in the case of the HIV1P

Table 8. Percentage of the Reference-Active and Reference-Inactive Pairs That Contain a Multinode Bioisosteric Equivalence

activity class	reference-active pairs	reference-inactive pairs
5HT1A	0.5	2.4
AChE	1.8	3.9
fXa	7.4	8.7
HIV1P	12.5	7.7
MMP1	5.9	6.4
PDE4	5.9	6.4

Table 9. Percentage Recall of Actives and Scaffolds Using Both Fused and Split Ring Nodes

	actives		scaffolds	
activity class	fused	split	fused	split
5HT1A	53.2	65.7	55.1	63.6
AChE	40.5	42.8	43.3	44.9
fXa	57.6	59.9	53.6	58.5
HIV1P	62.9	64.4	63.8	65.3
MMP1	53.0	53.9	56.3	58.7
PDE4	49.8	49.1	50.8	52.6
mean	52.8	56.0	53.8	57.3

class, where the percentage of reference-active pairs with a known equivalence is highest (in Table 8) and where the effect on search performance is greatest (Table 7). However, the percentage of reference-inactive pairs that have a known equivalence will also influence the search rankings, and there are, as we have noted previously, vastly more of these pairs.

Splitting of Fused Rings. Thus far, the reduction scheme has kept fused ring systems intact. An alternative approach would be to encode each ring in a fused system (specifically, the smallest set of smallest rings) with a different node. This gives a greater level of discrimination, which should help to reduce the number of false positives. When encoding fused ring systems as multiple nodes, atoms that are shared between adjacent rings are counted multiple times, e.g., the two fragments that would result from splitting naphthalene would each contain six atoms. The effect of splitting fused rings during graph reduction is shown in Table 9.

The results show that splitting rings improves performance in terms of recall for five out of the six activity classes. Although the improvement in the majority of the classes is quite modest, the performance in the 5HT1A class is greatly improved and is worth considering in more detail. The change in the similarity distribution was calculated for the 5HT1A data set by subtracting the similarity for each molecule obtained when splitting rings from the similarity obtained in the original search. This analysis showed that while about the same percentage of actives (18%) as inactives (21%) saw no change in their similarity to the closest reference structure, there was a marked difference in behavior for the remainder of the cases where the similarity changed after ring-splitting. Thus, 44% of the inactives had their similarity decreased and 35% increased, whereas for the actives the corresponding figures were 27% and 55%, respectively, i.e., much more discriminating searches resulted from the splitting. Moreover, this increased discrimination is achieved without detriment to the diversity of the actives retrieved, as demonstrated by the increased recall of scaffolds in Table 9.

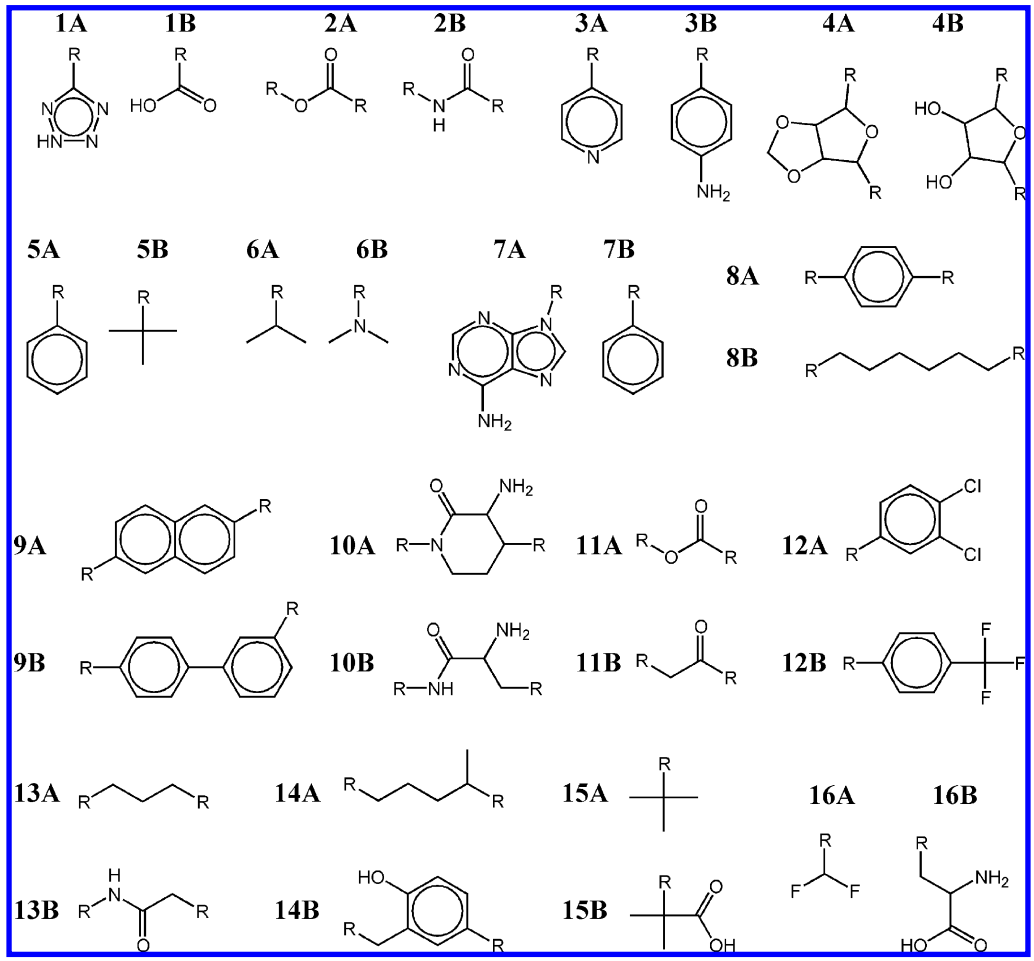
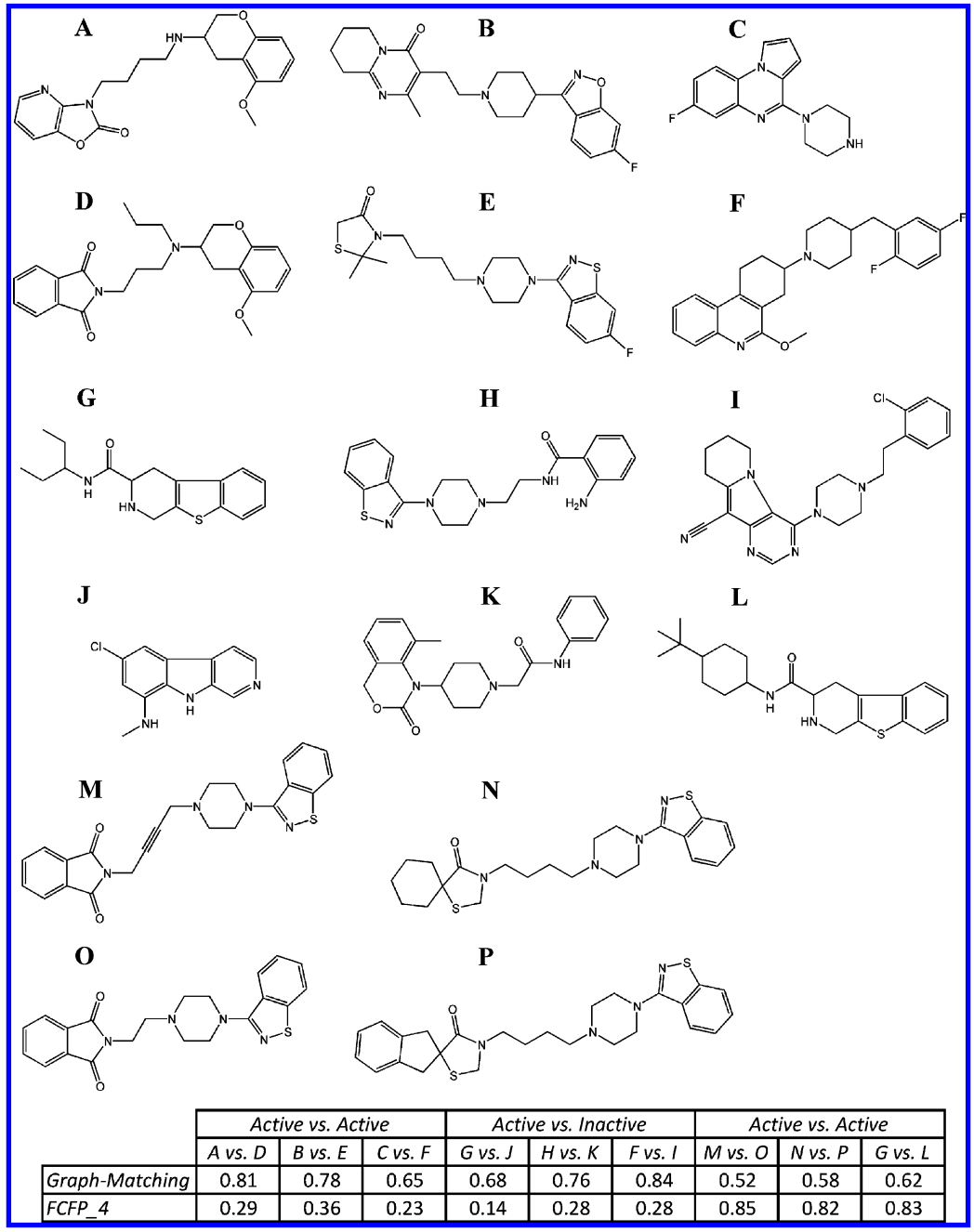


Figure 10. Example pairs of putatively bioisosteric fragments extracted from the BIOSTER database. Pairs 1 to 8 are single-node pairs where each fragment is represented by a single RG node, whereas 9 to 16 are pairs where the fragment A is represented by a single RG node and fragment B is represented by two nodes.

Examples of Bioisosteric Pairs. Some one-to-one node and one-to-two node putative bioisostere pairs are shown in Figure 10 to exemplify some of the interesting equivalences found and also to exemplify some of the more dubious equivalences that were extracted from the BIOSTER database. All the examples of single-node equivalences shown are of a different node type, but many of the more recognizable equivalences have similar node types. For example, both fragments in pair-1 have a negative charge, with the difference in node type resulting from the aromatic (A) versus acyclic (B) structure types; similarly, the fragments in pair-5 differ in their structural classification (aromatic and acyclic) but are related by their both being featureless (as well as being typically hydrophobic). The relatedness of the feature types in the fragments of pairs 2 to 4 is slightly more distant; H-bond acceptor (A) versus joint donor/acceptor (B) in each case, although the structure type is identical in each case. Whether or not replacement of these fragments would result in the maintenance of activity would depend on the specific interactions that the fragment makes with the target. Consequently, while pair-1 could be confidently labeled as truly bioisosteric, it seems likely that the next three pairs are less reliable, in the sense that they might be bioisosteric in some contexts but not in others. Another notable point is that these fragments tend to be similar in shape and size and where appropriate (pair-2 and pair-4) the relative positioning and distance between the R groups is maintained. This is something of great importance in allowing similar pharmacophoric configurations to be adopted in the parent molecules upon substitution of the bioisosteric fragments. For example, although there is similarity with regard to the size and functional node type in pair-8 (both featureless), fragment 8A has a much shorter distance between the R-groups and lacks the flexibility that may be required to adopt the pharmacophoric configurations achievable by fragment 8B: accordingly, these fragments should probably not be considered as generally bioisosteric. In the case of pair-6, although the fragments are of a similar size, shape, and structure type, the difference in feature type suggests that they would not be considered bioisosteric in many cases. It would however be much easier to conclude that the fragments in pair-7 are not bioisosteric since they differ quite significantly in terms of their size and features. The inclusion of dubious equivalences such as these in a similarity search would be expected to degrade performance due to an increased number of false positives. It would hence be beneficial to filter out these equivalences in the context of an operational implementation of our ideas.

Regarding the multinode equivalences shown in the lower half of Figure 10, pairs 9−12 seem potentially valuable, whereas pairs 13−16 are perhaps more dubious. The equivalence between the single aromatic featureless node and the two aromatic featureless nodes in pair-9 is quite reasonable given the similarity in shape, size, and R-group positioning. Pair-10 is also quite reasonable given the similarity in the positioning of the features. Fragment 11B



	Active vs. Active			Active vs. Inactive			Active vs. Active		
	A vs. D	B vs. E	C vs. F	G vs. J	H vs. K	F vs. I	M vs. O	N vs. P	G vs. L
Graph-Matching	0.81	0.78	0.65	0.68	0.76	0.84	0.52	0.58	0.62
FCFP_4	0.29	0.36	0.23	0.14	0.28	0.28	0.85	0.82	0.83

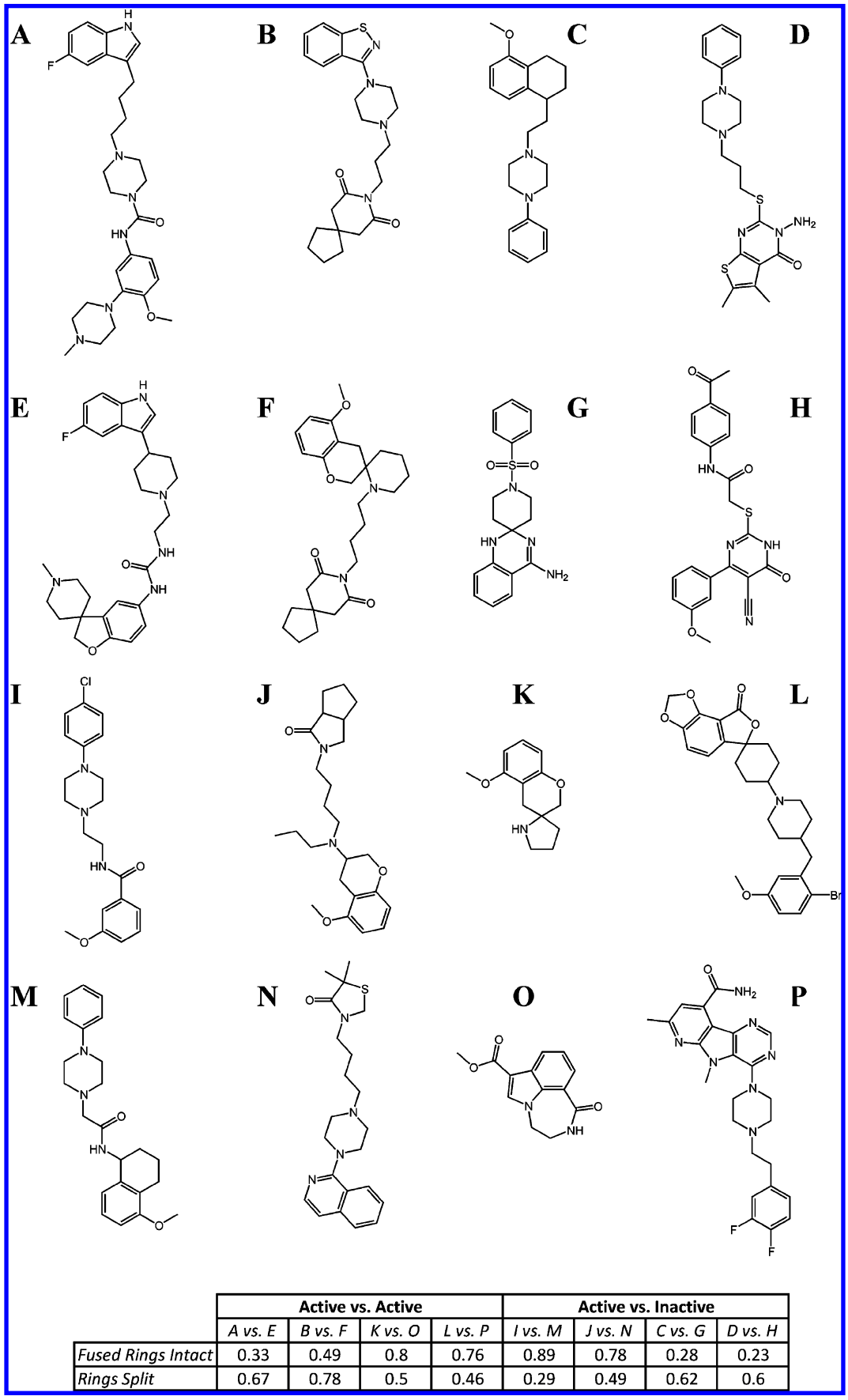
Figure 11. Examples of molecules retrieved uniquely by the FCFP4 and graph matching methods when searching through the 5HT1A data set.

differs from 11A in the loss of the ether-oxygen acceptor functionality, which seems more significant than the change in pair-2 where the acceptor functionality at that position is maintained and augmented with a donor functionality. However, pair-11 should probably still be considered fairly reliably bioisosteric since the stronger acceptor functionality of the carbonyl oxygen is maintained. In comparison, pair-13 is less reliably bioisosteric given the much greater loss of functionality. Pair-15 is an interesting case since fragment 15A is simply a substructure of 15B; however, given that the added functionality of the negatively ionizable carboxyl group is significantly different from the purely hydrophobic character of fragment 15A, the fragments are very unlikely to be bioisosteric. Pair-16 is also clearly unlikely to be bioisosteric given the difference in size and features.

From the examination of these bioisosteric pairs it seems relatively easy to identify by inspection pairs that are clearly likely or clearly unlikely to be truly bioisosteric, i.e., generally equivalent across a range of targets. It is, however, more difficult to make accurate absolute judgements on the reliability of pairs in between these two extremes. In reality such pairs are likely to be bioisosteric in some contexts but not in others, depending on the interactions that a particular fragment makes in a particular binding mode of a particular target.

Typical Search Outputs. Previous sections have discussed the overall performance of the RG searches; here we discuss the sorts of molecule that are retrieved, using outputs from the 5HT1A searches. Specifically, Figure 11 shows examples of molecules that were retrieved either by the RG search or by the FCFP_4 search but not by both to highlight the different characteristics of the two methods.

The actives retrieved uniquely by the RG search (pairs A and D, B and E, and C and F) are all examples of hits that



	Active vs. Active				Active vs. Inactive			
	A vs. E	*B vs. F*	*K vs. O*	*L vs. P*	*I vs. M*	*J vs. N*	*C vs. G*	*D vs. H*
Fused Rings Intact	0.33	0.49	0.8	0.76	0.89	0.78	0.28	0.23
Rings Split	0.67	0.78	0.5	0.46	0.29	0.49	0.62	0.6

Figure 12. Similarities of pairs of structures in 5HT1A searches when fused rings are split or kept intact during the generation of the RGs.

have only limited structural resemblance at the fingerprint level. The actives retrieved uniquely by the FCFP_4 method (pairs M and O, N and P, and G and L) are relatively close misses and exemplify the general difficulties encountered when applying graph reduction and graph matching methods, i.e., how best to partition a structure and how best to assign

a node type or permit matching. For example, in the pair M vs O, the acyl group forms a separate node that prevents the two halves of the structures from aligning correctly: a higher score would be attained for this match if gaps were tolerated. In the case of the pair N vs P the phenyl ring in the thiazolidinone-containing ring system changes the classification of the node type to aromatic such that it does not match with the corresponding ring system in structure N, when the ring systems remain intact. Similarly, the discrete classification of node types results in failure to recognize the commonality between the featureless terminal nodes of the pair G vs L. While it would be simple to alter the definitions to remedy these equivalences by altering the partitioning or graph matching definitions, the difficulty lies in the fact that there is no single set of rules that will be the best in every case. What may be reasonably regarded as equivalent in one situation may not be so in many others, as indicated by an increase in false positives. For example, in the example of pair G vs L in Figure 11, the equivalence between the para-t-butyl-cyclohexyl group of structure L and the 1-ethyl-propyl group of structure G is so specific as not to be documented in the BIOSTER database, and neither should it necessarily be, given that there would be many examples where this replacement would result in loss of activity due to the much larger volume occupied by the para-t-butyl-cyclohexyl group.

Finally, Figure 12 provides examples of some of the actives and inactives that were added and removed from the 5HT1A hit-lists when using the multinode encoding of fused ring systems. The examples in the figure show that both positive and negative outcomes are obtained, as listed in the table of similarity values. However, the recall figures presented previously in Table 9 show that ring splitting is generally beneficial.

CONCLUSIONS

In this paper we have reported the principal findings from a project to exploit the use of information about bioisosteric equivalences to enhance the performance of systems for similarity-based virtual screening. Our methods are based on the use of RGs, since these provide a simple and direct way of encoding multiple chemical fragments, such as bioisosteres, in a single node according to the functional properties of the fragments. The suitability of the reduced graph descriptor for encoding bioisosteric equivalences was demonstrated by the large proportion of BIOSTER-derived putatively bioisosteric fragments that were found to be encoded as the same node type. It was hence rather disappointing to find that using RGs for scaffold-hopping searches had little effect on search performance. The lack of improvement is indicative of the complexity of bioisostere information that cannot always be so easily generalized into the functional RG node type definitions. Consequently, attempts were made to incorporate bioisostere information more directly into the similarity calculations, by allowing matches between RG nodes encoding fragments found to be putatively bioisosteric. Our approach provides a simple and efficient means of including information when the bioisosteric fragments are encoded by one or (with some additional processing) two RG nodes. The improvements in performance obtained when incorporating bioisostere information in this way were quite modest although improvements in the diversity of the retrieved scaffolds were more pronounced. In comparison with more established similarity searching methods such as structural fingerprints, the RG descriptor is a viable alternative that although lower in recall offers interpretable results that would be of more interest in a scaffold hopping context; our findings here further support previous findings regarding the complementary nature of RG-based and fingerprint-based virtual screening.

There are many ways in which the work reported here could be extended. For example: the co-occurrence matrix (Figure 6) could be used to provide a probability-based approach to the calculation of internode similarities; consideration could be given to including matching, but non-overlaid, nodes in the similarity calculations; or more specific node descriptions could be used, e.g., the ErG approach of Stiefl et al.[34,35] However, enhancements such as these are unlikely to overcome the most significant problem that we have identified, *viz.* the fact that bioisosteric equivalences are at least as likely to occur between active reference structures and inactive structures in the database that is being searched as they are likely to occur in active structures. Given the much greater numbers of inactives than actives, it is very easy for equivalences involving the former type of molecule to swamp those involving the latter, as demonstrated by the figures in Tables 6 and 8. This problem is not inherent to the RG-based approaches studied here: it would affect any technique for including bioisosteric information unless there was already considerable information available as to the structural characteristics of the active and inactive molecules for the target of interest. Such information is unlikely to be available during the early stages of a lead-discovery project, which is where similarity searching is most likely to be used; accordingly, we believe it may prove quite difficult to use generalized bioisostere information in an effective manner for virtual screening.

ACKNOWLEDGMENT

We thank Sanofi-Aventis for funding, Accelrys Software Inc. for software and the BIOSTER database, and the Royal Society and the Wolfson Foundation for laboratory support.

REFERENCES AND NOTES

(1) Gasteiger, J. *Handbook of Chemoinformatics*; Wiley-VCH: Weinheim, Germany, 2003.
(2) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*, 2nd edition ed.;Kluwer: Dordrecht, The Netherlands, 2007.
(3) Willett, P. From Chemical Documentation to Chemoinformatics: Fifty Years of Chemical Information Science. *J. Inf. Sci.* **2008**, *34*, 477–499.
(4) Friedman, H. L. *Influence of Isosteric Replacements Upon Biological Activity 206*; National Academy of Sciences-USA: Washington, DC, 1951; pp 295−300.
(5) Thornber, C. W. Isosterism and Molecular Modification in Drug Design. *Quart. Rev. Chem.* **1979**, *8*, 563–579.
(6) Burger, A. Isosterism and Bioisosterism in Drug Design. *Prog. Drug Res.* **1991**, *37*, 287–367.
(7) Patani, G. A.; LaVoie, E. J. Bioisosterism: A Rational Approach in Drug Design. *Chem. Rev.* **1996**, *96*, 3147–3176.
(8) Kubinyi, H. Similarity and Dissimilarity: A Medicinal Chemist's View. *Perspect. Drug Discovery Des.* **1998**, *9−11*, 225–232.
(9) Olesen, P. H. The Use of Bioisosteric Groups in Lead Optimisation. *Curr. Opin. Drug Discovery Dev.* **2001**, *4*, 471–478.
(10) Böhm, H.-J.; Flohr, A.; Stahl, M. Scaffold Hopping. *Drug Discovery Today: Technol.* **2004**, *1*, 217–224.
(11) Lima, L. M.; Barreiro, E. J. Bioisosterism: A Useful Strategy for Molecular Modification and Drug Design. *Curr. Med. Chem.* **2005**, *12*, 23–49.

(12) Schneider, G.; Schneider, P.; Renner, S. Scaffold-Hopping: How Far Can You Jump. *QSAR Comb. Sci.* **2006**, *25*, 1162–171.
(13) Brown, N.; Jacoby, E. On Scaffolds and Hopping in Medicinal Chemistry. *Mini-Rev. Med. Chem.* **2006**, *6*, 1217–1229.
(14) Ujváry, I. Bioster - a Database of Structurally Analogous Compounds. *Pest. Sci.* **1997**, *51*, 92–95.
(15) Schuffenhauer, A.; Gillet, V. J.; Willett, P. Similarity Searching in Files of Three-Dimensional Chemical Structures: Analysis of the Bioster Database Using Two-Dimensional Fingerprints and Molecular Field Descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 295–307.
(16) Holliday, J. D.; Jelfs, S. P.; Willett, P. Calculation of Intersubstutuent Similarity Using R-Group Descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 406–411.
(17) Wagener, M.; Lommerse, J. P. M. The Quest for Bioisosteric Replacements. *J. Chem. Inf. Model.* **2006**, *46*, 677–685.
(18) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular-Features in Structure Activity Studies - Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
(19) Sheridan, R. P. The Most Common Chemical Replacements in Drug-Like Compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 103–108.
(20) Haubertin, D. Y.; Bruneau, P. A Database of Historically-Observed Chemical Replacements. *J. Chem. Inf. Model.* **2007**, *47*, 1294–1302.
(21) Leach, A. G.; Jones, H. D.; Cosgrove, D. A.; Kenny, P. W.; Ruston, L.; MacFaul, P.; Wood, J. M.; Colclough, N.; Law, B. Matched Molecular Pairs as a Guide Inthe Optimization of Pharmaceutical Properties: A Study of Aqueous Solubility, Plasma Protein Binding and Oral Exposure. *J. Med. Chem.* **2006**, *49*, 6672–6682.
(22) Sheridan, R. P.; Hunt, P.; Culbertson, J. C. Molecular Transformations as a Way of Finding and Exploiting Consistent Local QSAR. *J. Chem. Inf. Model.* **2006**, *46*, 180–192.
(23) Watson, P.; Willett, P.; Gillet, V. J.; Verdonk, M. L. Calculating the Knowledge-Based Similarity of Functional Groups Using Crystallographic Data. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 835–857.
(24) Bruno, I. J.; Cole, J. C.; Lommerse, J. P. M.; Rowland, R. S.; Taylor, R.; Verdonk, M. L. Isostar: A Library of Information About Nonbonded Interactions. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 525–537.
(25) Kennewell, E. A.; Willett, P.; Ducrot, P.; Luttman, C. Identification of Target-Specific Bioisosteric Fragments from Ligand-Protein Crystallographic Data. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 385–394.
(26) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338–345.
(27) Gillet, V. J.; Downs, G. M.; Ling, A.; Lynch, M. F.; Venkataram, P.; Wood, J. V.; Dethlefsen, W. Computer-Storage and Retrieval of Generic Chemical Structures in Patents. 8. Reduced Chemical Graphs and Their Applications in Generic Chemical-Structure Retrieval. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 126–137.
(28) Rarey, M.; Dixon, J. S. Feature Trees: A New Molecular Similarity Measure Based on Tree Matching. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471–490.
(29) Rarey, M.; Stahl, M. Similarity Searching in Large Combinatorial Chemistry Spaces. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 497–520.
(30) Gerlach, C.; Broughton, H.; Zaliani, A. F. Tree Query Construction for Virtual Screening: A Statistical Analysis. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 111–118.
(31) Barker, E. J.; Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Gillet, V. J.; Kitts, P.; Willett, P. Scaffold-Hopping Using Clique Detection Applied to Reduced Graphs. *J. Chem. Inf. Model.* **2006**, *46*, 503–511.
(32) 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.
(33) 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.
(34) Stiefl, N.; Watson, I. A.; Baumann, K.; Zaliani, A. ERG: 2D Pharmacophore Descriptions for Scaffold Hopping. *J. Chem. Inf. Model.* **2006**, *46*, 208–220.
(35) Stiefl, N.; Zaliani, A. A Knowledge-Based Weighting Approach to Ligand-Based Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 587–596.
(36) 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.
(37) Gardiner, E. J.; Gillet, V. J.; Willett, P.; Cosgrove, D. A. Representing Clusters Using a Maximum Common Edge Substructure Algorithm Applied to Reduced Graphs and Molecular Graphs. *J. Chem. Inf. Model.* **2007**, *47*, 354–366.
(38) Birchall, K.; Gillet, V. J.; Harper, G.; Pickett, S. D. Evolving Interpretable Structure-Activity Relationships. 1. Reduced Graph Queries. *J. Chem. Inf. Model.* **2008**, *48*, 1543–1557.
(39) Birchall, K.; Gillet, V. J.; Harper, G.; Pickett, S. D. Evolving Interpretable Structure-Activity Relationship Models. 2. Using Multiobjective Optimization to Derive Multiple Models. *J. Chem. Inf. Model.* **2008**, *48*, 1558–1570.
(40) Whittle, M.; Gillet, V. J.; Willett, P.; Alex, A.; Loesel, J. Enhancing the Effectiveness of Virtual Screening by Fusing Nearest Neighbor Lists: A Comparison of Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1840–1848.
(41) Henikoff, S.; Henikoff, J. G. Amino Acid Substitution Matrices from Protein Blocks. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 10915–10919.

CI900078H