

The Reduced Graph Descriptor in Virtual Screening and Data-Driven Clustering of High-Throughput Screening Data

G. Harper,* G. S. Bravi, S. D. Pickett, J. Hussain, and D. V. S. Green

GlaxoSmithKline, Gunnels Wood Road, Stevenage SG1 2NY, United Kingdom

Received April 27, 2004

Virtual screening and high-throughput screening are two major components of lead discovery within the pharmaceutical industry. In this paper we describe improvements to previously published methods for similarity searching with reduced graphs, with a particular focus on ligand-based virtual screening, and describe a novel use of reduced graphs in the clustering of high-throughput screening data. Literature methods for reduced graph similarity searching encode the reduced graphs as binary fingerprints, which has a number of issues. In this paper we extend the definition of the reduced graph to include positively and negatively ionizable groups and introduce a new method for measuring the similarity of reduced graphs based on a weighted edit distance. Moving beyond simple similarity searching, we show how more flexible queries can be built using reduced graphs and describe a database system that allows iterative querying with multiple representations. Reduced graphs capture many important features of ligand–receptor interactions and, in conjunction with other whole molecule descriptors, provide an informative way to review HTS data. We describe a novel use of reduced graphs in this context, introducing a method we have termed data-driven clustering, that identifies clusters of molecules represented by a particular whole molecule descriptor and enriched in active compounds.

1. INTRODUCTION

A number of 2D and 3D descriptors have been developed for the clustering and similarity searching of compounds. Development of new methods in this area is still being actively pursued, not least because no one method adequately describes the complexity of chemical space and its relationship to biological activity. High-throughput screening is a key component of lead discovery within the pharmaceutical industry. The challenge posed to the medicinal and computational chemist is how to sift through the potentially large number of hits that can be generated by a high-throughput screen in order to select interesting compounds for further follow-up.

In this paper, we explore the use of reduced graphs as a chemical descriptor both in the similarity searching problem and in the analysis of HTS data. Similarity searching using reduced graphs has been described previously by others.^{1,2} Here we introduce a new method for calculating the similarity between reduced graphs that overcomes some of the issues with standard fingerprint based methods. We also introduce a new tool for the analysis of HTS data, where reduced graphs are combined with other molecular representations to provide an ordered structural clustering of compounds to the scientist such that the early clusters are enriched in active (or hit) compounds.

Reduced graphs were originally developed for structure and substructure searching of generic chemical structures³ and since then have been further developed for similarity searching. Takahashi et al.⁴ and Fisanick et al.⁵ describe systems for similarity searching based on reduced graphs;

however, the work that is presented here builds on recent work from the University of Sheffield.^{1,2} In the Sheffield work, a specific kind of reduced graph based on ring systems and hydrogen-bonding groups was developed, the aim being to emphasize potential binding groups and their relative positions within a structure. It is this recent work that we develop further here.

2. REDUCED GRAPH DEFINITION

Our reduced graphs are based on the Ar/F(4) definition of Gillet et al.¹ However, we extend the range of pharmacophore types by additionally encoding positively and negatively ionizable features. The full list of basic feature types we include are hydrogen bond donor, hydrogen bond acceptor, positively ionizable group, negatively ionizable group, aliphatic ring, and aromatic ring. Following Barker et al.,² we merge adjacent donor and acceptor features into a distinct joint donor–acceptor feature. We also define joint ring/feature node-types such as aromatic ring/donor. However we stop short of encoding all possible combinations of features. A complete list of the groups that are encoded and their corresponding superatom codes is given in Table 1. The codes used in this table may initially seem unintuitive; however, encoding the superatoms in this way allows us to use standard pieces of software for handling molecules for reduced graph generation and handling. It also allows us to use standard structure visualization tools to visualize the reduced graphs. When an atom could be interpreted as having two or more pharmacophoric properties that have not been encoded as a combined feature in Table 1, some order of precedence is required to decide which feature-type should be set in the reduced graph. The order of precedence that we use to resolve clashes is (“positively ionizable” before

* Corresponding author phone: +44(0) 1438 768416; fax: +44(0) 1438 768232; e-mail: gavin.x.harper@gsk.com.

Table 1. Superatom Definitions Used in the Definition of the Reduced Graph

description	superatom code
aromatic ring nonfeature	Sc
aromatic ring donor	Ti
aromatic ring acceptor	V
aromatic ring donor & acceptor	Cr
aromatic ring positively ionizable	Mn
aromatic ring negatively ionizable	Fe
aliphatic ring nonfeature	Hf
aliphatic ring donor	Ta
aliphatic ring acceptor	W
aliphatic ring donor & acceptor	Re
aliphatic ring positively ionizable	Y
aliphatic ring negatively ionizable	Zr
feature node donor	Co
feature node acceptor	Ni
feature node donor & acceptor	Cu
feature node positively ionizable	Nb
feature node negatively ionizable	Mo
link node	Zn

“negatively ionizable” before “donor” or “acceptor”). If an atom matches one of the earlier definitions in this list, then it is excluded from matching subsequent ones. The existence of the joint donor–acceptor feature means that the order of precedence between donor and acceptor need not be defined.

3. SIMILARITY SEARCHING

A similarity score between molecules is usually defined based on the comparison of their chemical structure. The structure is often coded in a binary vector (or “fingerprint”) with each bit denoting the presence or absence of a particular substructure. The similarity between molecules is conventionally measured as a number between 0 and 1, calculated by comparing their fingerprints, 1 representing the highest degree of similarity.

Similarity scores can be used in various scenarios. One common use is to cluster molecules. A quite different use is for similarity searching. In similarity searching, a database of molecules is ranked in order of similarity to the “query molecule” provided by the user, and the molecules at the top of the ranked list are returned. There are several ways in which similarity searching can be applied. If a molecule with some desired biological activity (the “query” molecule) has been identified, similarity searching is commonly used to identify all close analogues to the molecule, helping to establish SAR quickly by testing these analogues at the biological target of interest. An alternative application of similarity searching, more akin to 2D pharmacophore searching, is where the aim is to search for molecules that are not obvious analogues of the query molecule yet retain the biological activity of interest. We will call the problem of identifying such compounds the “Scaffold-Hopping Virtual Screening Problem”.

The Scaffold-Hopping Virtual Screening Problem. *Given a molecule (the “query molecule”) that is known to be biologically active at a given target, select a list of a few hundred molecules from over a million available compounds (representing for example the space of compounds that could be acquired easily from external sources) such that at least one compound in the list is itself active at the biological target of interest, and ideally that it represents a lead series that is structurally distinct from the query molecule.*

Note the emphasis on identifying new chemotypes in this definition, rather than just finding active molecules. This definition of success in virtual screening is similar in spirit to the “chemotype hit rate” as discussed previously by Good et al.⁶ and has relevance in real-world virtual screening problems. Similarity searching may be used in this scenario to identify a novel lead series, either as a back-up or as an alternative lead series to the one represented by the query molecule. If a target were promiscuous, and hence was hit by a multitude of suitable ligands, it may be that reverting to a chemotype hit rate to judge success would be more appropriate than placing the emphasis on the success or otherwise of finding a single lead series. However, in our experience, the discovery of a single new lead series usually represents a successful outcome.

Several studies have shown 2D descriptors to be more effective than 3D descriptors for analogue-directed similarity searching and clustering.^{7–9} Such studies typically judge the success of methods on the basis of the extent to which the clustering succeeds in partitioning the data into clusters that contain predominantly active or predominantly inactive compounds. If the clusters are relatively pure in this sense, then the clustering is judged to have been successful, since the structural clustering follows the biological activity closely. This type of assessment is realistic when the application that is required is analogue hunting or structural clustering. It should come as no great surprise that 2D methods are superior in such assessments. The fingerprints produced for many 2D methods were developed for substructure searching and so faithfully return members of the same chemical series as the query molecule. Medicinal chemistry experience makes it clear that close analogues stand a good chance of exhibiting similar biological activity, and 2D methods often return these analogues. This is not a direct indicator of success in the Scaffold-Hopping Virtual Screening Problem.

3D pharmacophore fingerprints may be generated from a single conformer; however if this is done, there is a good chance that the conformer used is not the conformation adopted by the molecule when bound to the receptor of interest. Hence multiple conformers are usually generated, with each conformer setting multiple bits in the fingerprint.^{10,11} In this case, only a few bits are actually relevant to the binding of the molecule at a particular receptor, and the fingerprint itself is highly dependent on the precise method of conformer generation. Nonetheless, pharmacophore fingerprints have been shown to be quite successful in the Scaffold-Hopping Virtual Screening Problem.¹²

Given the complications introduced by conformational flexibility in 3D similarity searching methods, there does seem to be further room for alternative molecular descriptors that encode pharmacophore-type features (and hence have the ability to bridge across chemical series) but in a 2D rather than a 3D descriptor. Reduced graphs are one such alternative descriptor. Other descriptors that have this potential have been reported in the literature and include topological pharmacophores,^{13,14} topological binding property pairs,¹⁵ and feature trees.¹⁶ In this paper we develop similarity searching, aimed specifically at the Scaffold-Hopping Virtual Screening Problem and based on reduced graphs. The success of the method should be judged in terms of whether a molecule from a new lead series can be identified from a large database

(of order 10^6 or 10^7), when we assume that there is the capacity to establish the biological activity of a tiny fraction of that database (of order 10^3 or 10^4). The list of biologically active compounds identified need not be exhaustive in any sense. Our basic approach will be to define a molecular descriptor (and a method of establishing similarity from that descriptor) that is capable of spanning multiple chemical series, while trying to minimize false positive matches.

4. DEFINING THE SIMILARITY BETWEEN REDUCED GRAPHS

Fingerprints. Traditional fragment-based approaches to fingerprinting are unlikely to work particularly successfully for reduced graphs. As noted by Barker et al.,² because a small number of features are represented in a typical reduced graph, conventional fingerprinting using Daylight hashed fingerprints does not work particularly well, since such fingerprints are very sparse. Additionally, paths often occur multiple times in the reduced graph, and a fingerprint based on presence or absence of a feature loses this frequency information. Barker et al.² suggest setting bits in the fingerprint based on node-edge pairs. As they note, a small change in path length (in the reduced graph – the path length may be identical in the original molecules) means that two molecules will be considered dissimilar. An alternative suggestion that they make is to use node-bond pairs, where the distance used is the number of bonds separating the features in the original chemical graph rather than the reduced graph. This approach has some merit. Note however that these suggestions do not pay full attention to the topology of the reduced graph. For instance, there is no way of knowing when any of the nodes has degree three or greater. Neither is there any allowance for small changes in distance in either representation, so a donor separated by 4 bonds/edges from an acceptor sets completely different bits in the fingerprint from a donor separated by 5 bonds/edges from an acceptor. Nonetheless, using fingerprint similarity as a means of filtering down the molecules of potential interest in a database has clear merit, particularly since searching using fingerprints can be achieved very efficiently using standard packages. We implement our own version of the node-edge pairs approach mentioned above. However, our implementation is slightly different, in that we only encode paths of between zero and six bonds (inclusive) in the reduced graph (longer paths are treated as if they were of length 6), and we do not set any bits at all when either end-node is a linker (since the linker node is itself unlikely to be involved in binding). We set a separate bit for each unique occurrence of the node-edge pair, up to a maximum of 5 bits. This means that if a given node-edge pair occurs three times in the molecule, then it sets three bits. Additional bits are set for paths of length zero where the atom is adjacent to a double bond and paths of length one where a double bond connects the two atoms in the path, again up to a frequency of occurrence of five. Additional bits (up to a maximum of four occurrences) are also set for features of each type that are vertices of degree 3 or higher (i.e. are branch points in the reduced graph). These additional bits are set because fused rings and branch points were perceived to be relatively important features in the reduced graph. A bit is set for every 2,4,6 etc. heteroatoms in the original molecule that are not members of a ring (up to a maximum

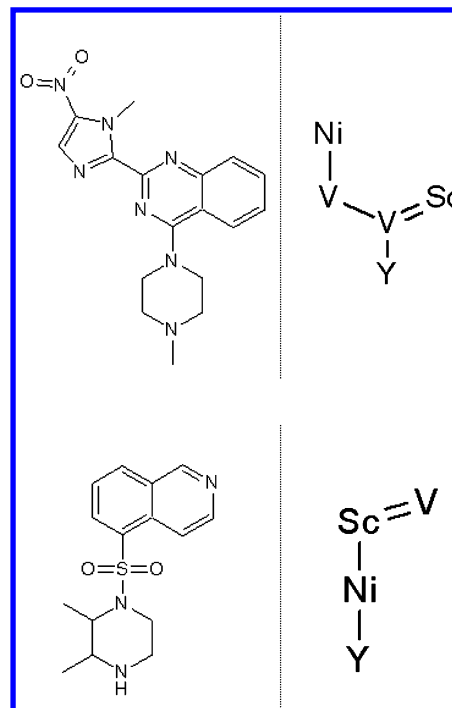


Figure 1. Query molecule (top) and returned molecule (below) with their corresponding reduced graphs (see Table 1 for an explanation of what the atoms represent). On the basis of the fingerprint alone, the molecule returned by fingerprint similarity searching shown here was 44th in the database. On the basis of the combined fingerprint and edit distance similarity, it was 1407th. Many of the features and the distances between them are similar in the two molecules, but the way in which the features are connected is substantially different. The edit distance takes better account of this than the fingerprint.

of 10 bits for 20+ atoms). Similarly, a bit is set for every 2,4,6 etc. heteroatoms in the original molecule that are members of a ring (up to a maximum of 10 bits for 20+ atoms). These two sets of bits are the only ones in the fingerprint that encode the original molecule directly rather than its reduced graph representation and can help to break ties when many molecules have the same reduced graph.

A danger with these fingerprints is that the same features at a very similar distance set no common bits, so that a single insertion or deletion of a superatom in the chain can stop two molecules with otherwise identical reduced graphs from being correctly identified as similar to each other. Hence, for node-edge pairs with edge length of 3 or greater, we also set bits in the fingerprint for the equivalent pair with edge length one shorter (so if the initial path length were 4, we also set additional bits for the equivalent path with length 3). This allows an association to be noted between paths varying only slightly in length in the reduced graph.

The bits generated by these rules can be fitted, without folding or hashing, into a standard 1024-bit fingerprint. These fingerprints still generate a lot of false positives when searching chemically diverse databases of the order of 10^6 compounds. This is largely due to reduced graphs with similar features but different connectivities between them generating a lot of the same node-edge pairs. In Figure 1, an example of a molecule returned at an inappropriately high point in the ranking demonstrates the problem. This shows the results of a search of the World Drug Index (WDI) database,¹⁷ containing 58945 compounds. Using the finger-

Table 2. Mutation and Insertion/Deletion Costs Used in the Edit-Distance Calculation

Matrix of Mutation Costs										
	aromatic ring ^a {Sc,Ti,V,Cr,Mn,Fe}	aliphatic ring ^a {Hf,Ta,W,Re,Y,Zr}	Nb	Mo	Co	Ni	Cu	Zn	—	=
aromatic ring ^a {Sc,Ti,V,Cr,Mn,Fe}	1	2	2	2	2	2	2	2	2	3
aliphatic ring ^a {Hf,Ta,W,Re,Y,Zr}	2	2	2	2	2	2	2	2	2	3
Nb	2	2	0	2	2	2	2	2	2	3
Mo	2	2	2	0	2	2	2	2	2	3
Co	2	2	2	2	0	2	1	2	2	3
Ni	2	2	2	2	2	0	1	2	2	3
Cu	2	2	2	2	1	1	0	2	2	3
Zn	2	2	2	2	2	2	2	0	2	3
— (single bond)	2	2	2	2	2	2	2	2	0	3
= (double bond)	3	3	3	3	3	3	3	3	3	0

Insertion/Deletion Costs					
feature	cost	feature	cost	feature	cost
aromatic ring {Sc,Ti,V,Cr,Mn,Fe}	2	Co	2	Zn	1
aliphatic ring {Hf,Ta,W,Re,Y,Zr}	2	Ni	2	— (single bond)	0
{Nb,Mo}	2	Cu	2	= (double bond)	3

^a For the sake of brevity, we include a single entry for all aromatic ring superatoms and a single entry for all aliphatic ring superatoms. Costs apply only when superatoms mismatch so a mutation of Sc to Ti incurs a cost of 1, but the trivial mutation of Sc to Sc incurs zero cost.

print for comparison, the second molecule is ranked 44th in the database in terms of the reduced graph similarity to the first molecule. This is exceptionally high in the ranking (within the top 0.1% of the database), but the reduced graphs are not highly similar by eye.

Path Comparison. The fingerprints generated by the above procedure provide a very useful and rapid first-pass filter in database searching. Therefore we have developed an alternative similarity measure based on the concept of edit distance¹⁸ as a second pass. This involves comparing all the pairwise paths in two molecules to calculate the weighted edit distance, as described below.

The *simple edit distance* between two strings is the minimum number of operations (insertions, deletions and mutations) to change string one into string two. A common extension to the simple edit distance is to associate different weights (also often termed costs) to each operation depending on the pair of characters (or in this case features) involved and whether the operation is an insertion, deletion, or mutation. When the minimum edit distance is computed using these weights, it is commonly called the *weighted edit distance*. For reduced graphs, we will want to penalize the insertion of a double bond into a path more heavily than the insertion of a single bond. Similarly, a mutation from a donor feature to a joint donor–acceptor feature should be penalized less heavily than a mutation from a donor feature to an aromatic ring. It is difficult to know exactly how to assign weights to different operations. Based largely on intuition, we have chosen to use the weights shown in Table 2. Obviously a more careful assessment of the effects of the weights on performance could be made and the weights adjusted accordingly. This could be done by training the weights using an optimization technique such as a genetic algorithm with an objective function that rewarded the retrieval of molecules that had the same known biological activity as the query molecule. If there is prior knowledge about the relevant importance of features in the query molecule and possible replacements for existing groups, the weights could also be adjusted for a specific application on the basis of this knowledge. However we do not explore the

choice of weights any further here. The weighted edit distance between two paths may be calculated using dynamic programming techniques.¹⁸

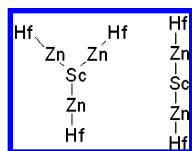
Thus far, we have only described how to compare paths, but not how we compare the molecules themselves. To calculate the distance between two molecules, we define the cost of each maximal path (path between two vertices of degree 1) in molecule 1 by the minimum weighted edit distance when compared with each maximal path in molecule 2 (including the paths compared both forwards and backward). Similarly, we define the cost of each path in molecule 2 by the minimum weighted edit distance when compared with each path in molecule 1. We define the weighted edit distance between two molecules by the maximum cost path, taken over all paths in both molecules. This gives us an idea of how dissimilar the two molecules are from each other. Traditionally however, comparisons between molecules are expressed as similarities. We transform the weighted edit distance between two molecules into a similarity by taking one minus the ratio of the weighted edit distance between molecules to twice the number of superatoms in the smaller of the two molecules. If this value is less than zero, then the similarity is taken to be exactly 0. This ensures that all similarities lie between 0 and 1 and is appropriate provided that the distribution of weighted edit distances is not so heavily skewed toward negative numbers that the vast majority of compounds would otherwise have similarities below zero. Note that for the application to virtual screening envisaged, the relative ranking of compounds further down the hitlist is relatively unimportant.

An example should help to clarify the procedure stated above. To compare the molecules in Figure 1 using weighted edit distance, we first compare all pairs of paths in the two molecules. The first molecule contains three maximal paths, while the second one contains only one. Hence there are six pairs of paths to be compared (There are twelve pairs of paths in total including each path written forward and backward; however, comparing paths 1 and 2 is equivalent to comparing the reverse path 1 with the reverse path 2, so there are six possible distinct comparisons between paths.).

Table 3. Weighted Edit Distances of Pairs of Paths in the Two Molecules Shown in Figure 1^a

path	weighted edit distance	alignment	
1 [Ni][V][V]=[Sc]	8	Ni — V — V = Sc	V = Sc — Ni — Y
1 reversed [Sc]=[V][V][Ni]	6	Sc = V — V — Ni	V = Sc — Ni — Y
2 [Ni][V][V][Y]	8	Ni — V — V — Y	V = Sc — Ni — Y
2 reversed [Y][V][V][Ni]	8	Y — V — V — Ni	V = Sc — Ni — Y
3 [Sc]=[V][Y]	4	Sc = V —	V = Sc — Ni — Y
3 reversed [Y][V]=[Sc]	6	Y — V = Sc	V = Sc — Ni — Y

^a The second path in the pair is the same in each case, being the single path in the reduced graph of the second molecule shown in Figure 1. Each of the three paths in the first molecule in Figure 1 are shown in the table aligned to the path from molecule 2 in both forward and reverse directions. Single bonds are denoted by a '—', double bonds by a '='. Weights, as given for each insertion, deletion, or mutation operation by Table 2, are noted under each alignment of a pair of paths. The weighted edit distance for the alignment is calculated as the sum of the individual weights incurred by the alignment.

**Figure 2.** On the basis of edit distance alone, these two reduced graphs have a similarity of 1. This is because all three maximal paths in the reduced graph on the left are matched identically by the single maximal path in the reduced graph on the right.

Through insertions and mutations, a cost for each pair of paths is calculated. Table 3 shows each of the three paths in molecule 1 in both forward and reverse directions lined up with the single path in molecule 2. The smallest edit distance for any alignment pair including the single path in molecule 2 is 4. The smallest edit distances for each of the three paths in molecule 1 are 6, 8, and 4, respectively. Hence the largest cost for any path is 8. Since the smaller of the two reduced graphs contains four superatoms, the similarity in this case is $1 - 8/(2 \times 4)$, which is zero.

The weighted edit distance similarity overcomes many of the difficulties of the fingerprint. In particular, it deals with insertions and deletions of superatoms gracefully, maintaining a high level of similarity when individual operations of this type are the only differences between reduced graphs. However, because it is based on the comparison of paths rather than whole molecules, it can also generate false positives, particularly in the case of molecules with symmetry (see Figure 2 for an example). The problem with symmetry is not present in the fingerprint-based comparison. Hence, where the weighted edit distance similarity exists (i.e. when there are no cycles in the reduced graph), we take the overall similarity to be the average of the fingerprint and edit distance similarities. Where the weighted edit distance

similarity does not exist, we define the overall similarity by the value of the fingerprint-based similarity alone.

In the example discussed earlier, and depicted in Figure 1, the second molecule in the figure is only ranked 1407th out of the 58945 compounds in the database using this averaged fingerprint and edit distance definition of overall similarity to the first (query) molecule, whereas it was ranked 44th on the basis of the fingerprint similarity alone. The movement down the ranked list of this molecule seems appropriate and reflects our earlier concern that the reduced graphs of the two molecules concerned are not all that similar in terms of how the features are connected together, despite containing broadly similar features.

5. SMARTS SEARCHING

Because reduced graphs as defined here can be written as valid SMILES strings, SMARTS¹⁹ and substructure querying is possible using the reduced graph representation. SMARTS queries allow the user to be quite specific about the kind of molecule that they wish to be returned. Consider for example a reduced graph SMARTS query of



This reduced graph SMARTS query looks for an acceptor or joint donor–acceptor feature, connected to a linker, connected to any aromatic ring superatom. By restricting the degree of the first two superatoms, but not the third, we allow additional superatoms to be present in matching molecules if they are connected to the aromatic ring but allow for no additional connections to the other two superatoms. Just as we encode in this example the feature “any aromatic ring”, various concepts such as “any acidic group” or “any nonring feature” are easy to encode.

Whereas reduced graph similarity searching can be run without any expert knowledge other than an initial active molecule with which to query, SMARTS querying provides great potential for encoding expert knowledge in a straightforward and flexible manner.

6. IMPLEMENTATION

We have built a database of compounds for searching using reduced graph querying in Oracle using DayCart.²⁰ The structure and fingerprint of both the molecule and its reduced graph are stored for each molecule, enabling substructure, SMARTS, and fingerprint-based similarity on both representations. Note that the fingerprint for the original molecule is the native Daylight one, while we set the bits in the reduced graph fingerprint as described earlier. This fingerprint can be searched in exactly the same way as for a normal Daylight fingerprint. When similarity searching using reduced graphs, we set a threshold on the level of fingerprint similarity to cut down the original database to a much shorter list of similar molecules, then pass these molecules to a C program using the Daylight toolkit to calculate edit distances, and hence calculate the overall similarity of these molecules. A simple Web interface allows the user to enter query molecules or SMARTS. It is straightforward for the user to proceed iteratively, trying out various queries and modifying them depending on the molecules returned by each. This iterative selection procedure enables an effective use of

expert knowledge. When a user finds a group of molecules that are of interest, they can download the list of molecules with their SMILES. These molecules will then typically be postprocessed using standard clustering and filtering tools to arrive at a final list of compounds for purchase and/or testing.

7. EXAMPLE: CB1 ANTAGONISTS

To investigate the power of reduced graph similarity searching for virtual screening, we used a database of compounds that were believed to be available for purchase from external suppliers. This database was built up on the basis of compounds offered to GlaxoSmithKline for purchase and contained 2056212 compounds. The database is of a suitable order of size and composition to judge success in the Scaffold-Hopping Virtual Screening Problem. To this database, we added known active compounds at given targets. If a query molecule can find a structurally distinct molecule which is one of the known active compounds in the first few hundred molecules from a database of more than two million, this provides a reasonable demonstration of the effectiveness of reduced graph similarity searching when applied to the Scaffold-Hopping Virtual Screening Problem. (Of course, it is also possible that other active compounds are in the list, but since this is a retrospective analysis we are unable to ascertain whether these other compounds are active or not.)

In this example, a known active compound from Sanofi²¹ (the earliest of the molecules to be patented as a CB1 antagonist) was used as the query compound, and 10 subsequently reported CB1 antagonists^{22–31} from eight different companies were added to the database to be searched. These are shown in Figure 3. Two of the nine turned out to have identical reduced graphs to the query compound. These were molecule 8 (Daylight similarity 0.468) and molecule 10 (Daylight similarity 0.420). There were only a total of nine compounds with identical reduced graphs to the query in the entire database of compounds. Molecule 4 had a reduced graph similarity of 0.815 (Daylight similarity 0.327). Only 66 compounds had a similarity equal to or greater than 0.815. Hence there were three active compounds, which would not have been retrieved by the Daylight method, found in the first 66 compounds that would have been screened out of over two million available compounds. Granted, the other six known active molecules would not have been found immediately from the reduced graph search – the next highest in the rankings was molecule 6 with a similarity of 0.646 tied for 6340th in the rankings and hence well outside of the rankings that we suggest necessary for discovery in our formulation of the Scaffold-Hopping Virtual Screening Problem. Nonetheless, the discovery of a single novel lead series often provides a satisfactory starting point for a medicinal chemistry project, so finding three when less than a hundred molecules are screened represents a very encouraging result. If instead of reduced graphs, we use Daylight similarity to search for molecules, then molecule 5 (third most similar, Daylight similarity 0.72), molecule 1 (twenty-fifth most similar, Daylight similarity 0.67), and molecule 7 (ninety-third most similar, Daylight similarity 0.60) are found. Note that there is no overlap with the molecules retrieved by the reduced graph search. Furthermore, the next

most similar molecule is molecule 8, 6717th in the list with a Daylight similarity of 0.47.

We suggested above that conventional similarity and substructure searching techniques could be used to expand out a single hit compound to molecules in the same lead series. As a second round following on from the reduced graph search, we took the original query molecule plus the three hits found from reduced graph searching and ranked the entire database based on maximum Daylight similarity to any of these four CB1 actives. The top 100 molecules in the database when ranked by this criterion now include molecules 1, 5, 6, and 7. Although three of these come from their similarity to the original query molecule, molecule 6 is identified as a result of searching with molecule 8, identified through our reduced graph search. Hence, after two rounds of screening, and only 200 molecules selected in total, seven of the ten known CB1 antagonists that we added to our database of over two million externally available compounds have been found in this exercise.

This single example is retrospective. Clearly to gauge accurately the effectiveness of reduced graph similarity searching, several prospective analyses should ideally be run. Nonetheless, this example demonstrates the ability of the method to identify molecules which would fail to be identified by the conventional similarity searching technique used here and shows how reduced graph methods may be used to complement other search techniques.

8. SUMMARIZING HIGH-THROUGHPUT SCREENING DATA – DATA-DRIVEN CLUSTERING

The reduced graph is a whole molecule descriptor. Using both the reduced graph and versions of another whole molecule descriptor (the “framework” defined below), to describe molecular structure, we have developed an algorithm that we call data-driven clustering to sift through high-throughput screening data in an informative manner. Data-driven clustering uses the biological activity data from the screen to identify sizable structural clusters of predominantly active compounds, all sharing a large common framework or reduced graph component, from screening data of hundreds of thousands of compounds.

The frameworks we use are based on those introduced by Bemis and Murcko.³² We use three basic types of framework. In all three cases we begin generation of the framework by recursively deleting atoms corresponding to vertices with degree one in the chemical graph of the molecule, until no further atoms can be deleted (no vertices of degree one remain in the chemical graph). This defines the framework of the molecule for the first type of framework. For the second type of framework, we delete the atomtype (C,N,O, etc.) labels but leave the bond-order information. In the third type of framework, we also delete the bond-order information, leaving a conventional graph made up of vertices and edges.

Data-driven clustering can be run exceptionally quickly (usually in a matter of a few minutes) once the reduced graphs and frameworks have been generated for the molecules and so gives an instant view of some of the main features in the results of the screen. It can easily be adapted to work with pooled data, where it identifies in clusters the component of each active pool that is likely to be responsible for the apparent activity.

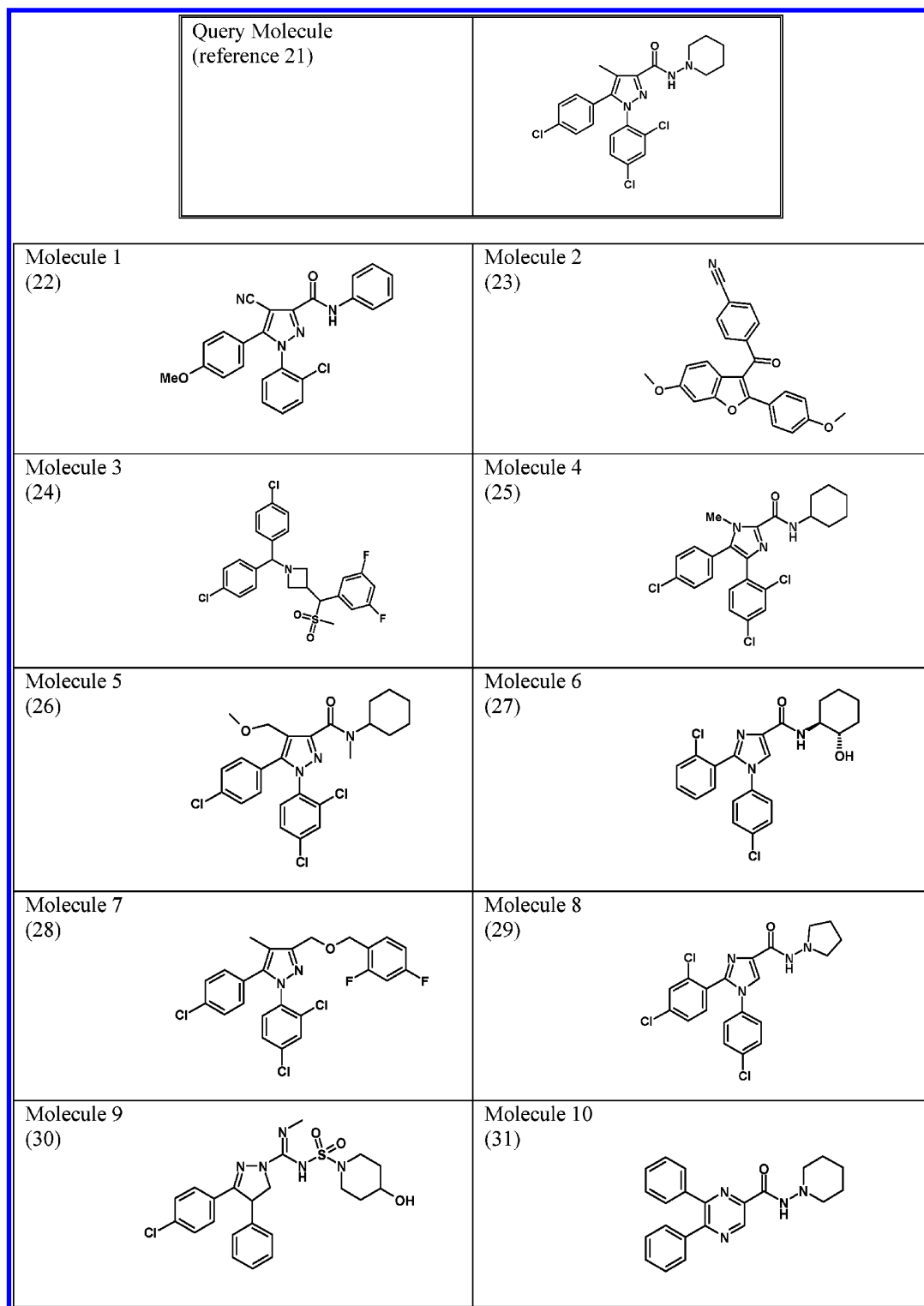


Figure 3. CB1 antagonists — reference numbers are given in brackets and correspond to the reference number given at the end of the article.

Clustering on frameworks and reduced graphs has the advantage that the scientist can immediately identify the precise criterion defining the cluster of compounds. This is in contrast to fingerprint-based methods, which rely on the more abstract criterion of sharing common bits in the fingerprint representation. As we use the activity of all the compounds screened, we return to the user a profile of all the compounds that match the given framework or reduced graph (provided that the compound has not already appeared

in an earlier cluster). This can provide some early ideas regarding SAR.

We would normally recommend a full visual analysis of the putative hits from a screen using the SIV (Selection through Interactive Visualization) process.³³ However the data-driven analysis often gives hints about the kinds of filters that may be appropriate to apply for compounds to progress or pools to deconvolute and in addition is a very quick and simple first view of the data. It is also the only view that we

give of the data where the inactive data is clustered along with the putative hits. This can draw attention to relatively small clusters that contain a particularly high proportion of apparently active compounds. It can also be used when reviewing the compounds selected for progression to double check that sufficient compounds have been selected from the potentially interesting clusters identified by the data-driven method.

Descriptor Generation. We generate the primary reduced graph from a molecule as described earlier. In addition to the simple reduced graph, we also generate a number of “near neighbor” reduced graphs that involve changing or deleting a single superatom in the primary reduced graph. Ring superatoms are never deleted; however, any nonring superatom may be deleted (or mutated to a linker node if degree 2 or greater) to produce a near neighbor reduced graph. Also joint donor–acceptor features may be changed to the equivalent donor feature or acceptor feature, and any other type of feature (donor, acceptor, positively ionizable, negatively ionizable) may be removed, resulting in a changed superatom type. Hence an aromatic ring donor–acceptor superatom may become an aromatic ring donor, an aromatic ring acceptor, or an aromatic ring, while the rest of the molecule is conserved, hence generating three near neighbors from the permutation of a single superatom. Where a superatom does not contain a ring feature, the superatom may be deleted altogether or changed into a linker superatom. Note that this may result in the near neighbor reduced graph needing to be further altered by two adjacent linker nodes being contracted to a single linker node.

For each of the three framework types, we generate near neighbors through the deletion of single nonfused rings from the primary framework. Following the deletion of the ring, any resulting side chains are also removed.

We will refer to the collection of frameworks, reduced graphs, and near neighbors of each as the “motifs” of the molecule. As an example, Figure 4 shows all the motifs that are generated from the query molecule used earlier in the paper and shown in Figure 1.

In data-driven analysis, molecules may only be clustered together when they all share a common motif. Since the motifs are whole-molecule descriptors, this generally guarantees the soundness of the structural clustering. No such guarantee exists with fragment-based recursive partitioning based methods.

Algorithm. Suppose that a collection of molecules C has been screened and that each individual molecule m has an activity $a(m)$ in the assay. Given a subset $B \subseteq C$, we define an associated score $f(a(B))$ which is a function of the activities of the molecules in B . We also define a threshold score K . If, at any point during the algorithm described below, all available clusters have scores below this threshold, we will not form any further clusters, judging that there is little evidence of consistent activity in any of the structural clusters available to us. At any particular step in the data-driven clustering algorithm, there is a set of molecules that are yet to be clustered U . At the beginning of the algorithm $U \equiv C$. The algorithm is defined as follows:

$$U \equiv C, n = 0.$$

1. Rank all motifs in descending order of $f(a(M_i))$ where M_i is the subset of the unclustered molecules that match the i th motif.

2. $n \rightarrow n+1$

3. Take the motif M_j giving the highest score f :

3a. If $f(a(M_j)) < K$: STOP. Clustering is complete.

Otherwise:

3b. Form cluster n : $C_n \equiv M_j$

3c. Remove cluster n from the set of unclustered molecules: $U \equiv U \setminus C_n$

4. Go to 1.

This defines the data-driven clustering algorithm completely given a scoring function f and threshold score K .

The scoring function for a cluster C of molecules matching a given motif that we use for high-throughput screening data is

$$\sum_{i \in C} [a(m_i) - k] \quad (1)$$

where k is a “break-even” point such that a molecule with activity higher than k makes a positive contribution to the score, while a molecule with a value less than k makes a negative contribution. We normally set k at twice the standard deviation of the middle (in terms of activity) 75% of the data. This provides a robust estimate of the general spread of the data. Lower multiples of this value could also be used if desired.

The scoring function is motivated by assuming that molecules in an active cluster have their activity distributed uniformly on the positive real line, while molecules in a cluster of inactive molecules follow an exponential distribution with density of form $\lambda e^{-\lambda x}$ — see the Appendix for further details.

For this scoring function there is a very natural cutoff value given by setting $K = 0$ in step 3a of the algorithm above. This is easy to interpret, since the average value of the activity of the cluster is exactly k , the break-even activity value for an individual molecule. Finally, we find it practically useful to add a restriction that clusters have a minimum size, typically set to 4 or 5, so that at each step, only motifs that would result in a cluster of size at least 4 or 5 are selected.

What results from the use of the algorithm described is a series of ordered clusters with large clusters of compounds with consistently high activity coming first. It must be stressed that we do not regard data-driven clustering as a prediction algorithm as such, since genuinely active small clusters and singletons will never be included in the clustering. Rather, it provides an exceptionally rapid method for highlighting the gross features apparent in the screening data. Our experience has been that the highest ranked clusters often contain compounds that interfere with the assay in some way since such compounds often produce the most consistent activity signal irrespective of minor structural variations. Libraries of compounds made for historical medicinal chemistry projects on targets related to the one being screened are also often easily identified and can be an indication early in the analysis that the screen has succeeded in picking up a true activity signal.

Note that although we use structural “whole molecule” descriptors, this approach could be extended to any kind of

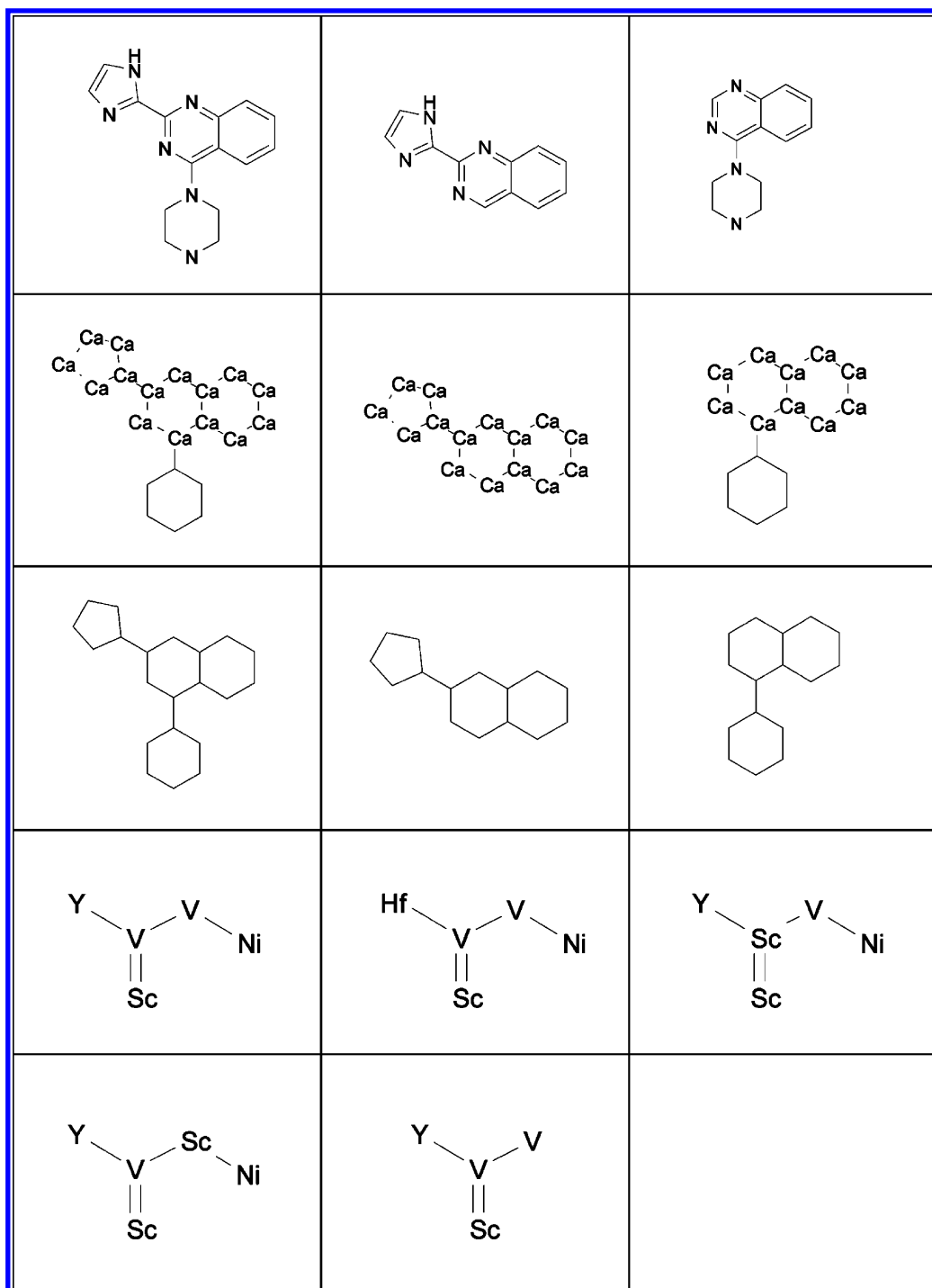


Figure 4. The motifs for the query molecule shown in Figure 1. Note that [Ca] represents an atom in an aromatic ring in the bond-framework representation.

descriptor that describes a meaningful class of molecules. For instance, one could imagine defining nonstructural classes of molecules such as “known to interfere with assay format X”, “synthesized for project Y”, or “binds to receptor Z” as motifs.

9. EXAMPLE: NCI AIDS DATA SET

Availability of large sets of high-throughput screening data in the public domain is poor. To demonstrate the use of data-driven clustering, we use here the much-studied NCI AIDS³⁴ data set. These data contain activity for each of over 40000 compounds on the basis of a dose–response curve. The

activity of each compound is summarized as belonging to one of three categories (High Activity/Medium Activity/Inactive). To simulate single shot screening using these data, we assigned a (uniformly) random activity to the inactive compounds in the range (−40,40), for the medium activity compounds in the range (10,90), and for the active compounds in the range (30,110). This effectively added a substantial degree of noise to the existing data, as might be expected from high-throughput single shot screening. The top seven motifs can be seen in Figure 5, along with one of the molecules in the cluster. It can be seen that a variety of frameworks and reduced graphs are identified. The exact

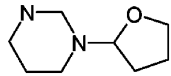
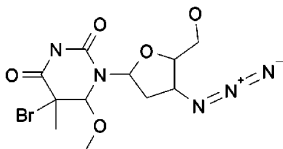
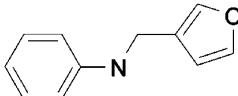
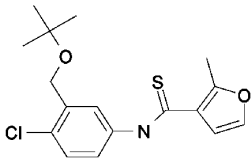
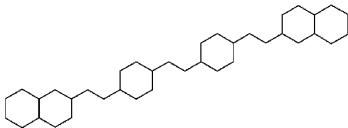
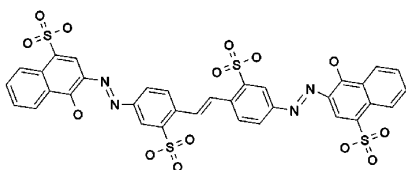
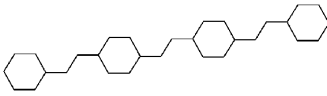
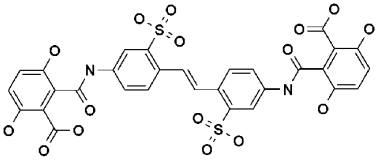
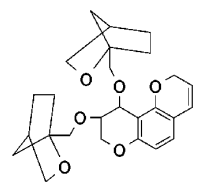
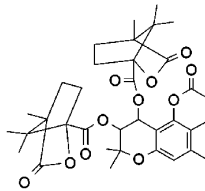
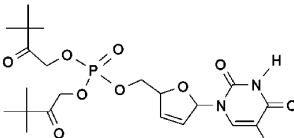
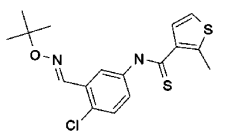
Motif	Molecule from Cluster	Cluster Constituents
		20 Active 1 Medium 9 Inactive
		21 Active
		22 Active 10 Medium 6 Inactive
		13 Active 21 Medium 19 Inactive
		12 Active 2 Medium 1 Inactive
Ni—Zn—Ni—Zn—W—Re		13 Active 2 Medium 3 Inactive
Ni—Zn—Sc—Cu—Sc		9 Active 5 Medium 7 Inactive

Figure 5. The first seven motifs from data-driven clustering of the AIDS data set and a molecule from the resulting clusters. The makeup of the cluster in terms of the original activity classes is also reported. Real-valued data was generated as described in the text, with overlapping ranges for different activity classes.

ordering of the clusters looks slightly unintuitive in terms of the original classification of the molecules. Note however that we added noise to the original data to generate artificial single shot data, with different activity classes mapping to activities chosen from overlapping ranges. It is clear that the clusters returned early on by the data-driven clustering are large clusters of predominantly active compounds. The motifs of these clusters are a mixture of framework and reduced graph motifs, showing how the different types of

molecular descriptor complement each other in describing the screening data.

10. CONCLUSIONS

We have demonstrated various uses and applications of the reduced graph descriptor. It would appear that it has good potential as an interpretable two-dimensional pharmacophoric descriptor. As such, it is a descriptor that complements more

traditional two-dimensional and three-dimensional molecular descriptors.

We have introduced edit distance as a new method for handling comparisons of reduced graph descriptors that can complement fingerprint-based techniques and explored the flexibility of querying that is possible with reduced graphs using substructure and SMARTS queries. In particular, the reduced graph appears to be a useful tool for virtual screening applications.

We have also shown how the reduced graph can be a useful whole-molecule descriptor that can be used to make sense of the considerable output from high-throughput screening in a straightforward and completely automated manner. The algorithm that achieves this, data-driven clustering, is a method for the automated identification of clusters of predominantly potent compounds. These clusters prove to typically be of high quality in terms of the chemical structure of the compounds clustered together because they are based on whole molecule descriptors rather than multiple molecular fragments.

ACKNOWLEDGMENT

The authors would like to thank Dr. Jason D. Speake for drawing their attention to the CB1 antagonist example used in this paper. We would also like to thank Dr. Andrew Leach and Dr. Francis Atkinson, who provided the SMARTS definitions used in the generation of the reduced graphs, and Dr. Val Gillet who provided the original reduced graph code.

APPENDIX

Suppose that the event A means that a given cluster C contains actives, and I means that the cluster does not contain actives. Suppose that \mathbf{x} is the vector of activities of the compounds in that cluster. For ranking clusters, we are directly interested in the probability that a cluster contains actives given the activity values of the compounds in that cluster, which we write as $\text{Pr}(A|\mathbf{x})$. Similarly, we write the unconditional probability that a cluster contain actives as $\text{Pr}(A)$. By Bayes' rule, and if we assume that the compounds within an active or inactive cluster have activity distributed identically and independently, then

$$\text{Pr}(A|\mathbf{x}) = \frac{\prod_{i \in C} f_A(a(m_i)) \text{Pr}(A)}{\prod_{i \in C} f_A(a(m_i)) \text{Pr}(A) + \prod_{i \in C} f_I(a(m_i)) [1 - \text{Pr}(A)]} \quad (\text{A1})$$

where $f_A(\cdot)$ is the density function for the activity of a compound from a cluster containing actives, and $f_I(\cdot)$ is the corresponding density function for a compound in a cluster of inactives. Ranking clusters by eq A1 is equivalent to ranking them by

$$\frac{\prod_{i \in C} f_A(a(m_i))}{\prod_{i \in C} f_I(a(m_i))} \quad (\text{A2})$$

Active clusters typically have a mix of active and inactive compounds contained in them. A simple proposed form for

the density function of the activity of a compound in an active cluster is that of a uniform distribution. Clusters of inactive compounds will on the other hand typically give activity values near to zero (although various errors can occur in screening, producing activity values that are not necessarily near to zero). If we restrict our attention to positive values of activity, then a simple form for the density function would be exponentially decreasing. If we assume these two basic forms of density function, then for some constants K_0 and K_1 , expression A2 above is proportional to

$$\prod_{i \in C} \frac{K_0}{\exp(-K_1 a(m_i))}$$

and, taking the natural logarithm (which is monotonic in its argument), ranking by this function is equivalent to ranking by

$$\sum_{i \in C} [a(m_i) - k]$$

for some constant k .

REFERENCES AND NOTES

- (1) Gillet, V. J.; Willett, P.; Bradshaw, J.; Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 338–345.
- (2) Barker, E. J.; Gardiner, E. J.; Gillet, V. J.; Kitts, P.; Morris, J.; Further Developments of Reduced Graphs for Identifying Bioactive Compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 346–356.
- (3) Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Lynch, M. F.; Dethlefsen, W.; Computer Storage and Retrieval of Generic Chemical Structures in Patents. 13. Reduced Graph Generation. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 260–270.
- (4) 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.
- (5) Fisanick, W.; Lipkus, A. H.; Rusinko, A., III.; Similarity Searching on CAS Registry Schemes. 2. 2D Structural Similarity. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 130–140.
- (6) Good, A. C.; Cheney, D. L.; Sitkoff, D. F.; Tokarski, J. S.; Stouch, T. R.; Bassolino, D. A.; Krystek, S. R.; Li, Y.; Mason, J. S.; Perkins, T. D. J. Analysis and optimization of structure-based virtual screening protocols: 2. Examination of docked ligand orientation sampling methodology: mapping a pharmacophore for success. *J. Mol. Graphics Modell.* **2003**, *22*, 31–40.
- (7) Brown, R. D.; Martin, Y. C.; Use of Structure–Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.
- (8) Brown, R. D.; Martin, Y. C.; The Information Content of 2D and 3D Structural Descriptors Relevant to Receptor–Ligand Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1–9.
- (9) Matter, H. Selecting Optimally Diverse Compounds from Structural Databases. A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219–1229.
- (10) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries (PDQ). *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214–1223.
- (11) Brown, R. D.; Martin, Y. C. Use of Structure–Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.
- (12) Pickett, S. D.; McLay, I. M.; Clark, D. E. Enhancing the Hit-to-Lead Properties of Lead Optimization Libraries. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 263–272.
- (13) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physicochemical Properties. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118–127.
- (14) Scheider, G.; Neidhart, W.; Giller, T.; Schmid, G.; “Scaffold Hopping” by topological pharmacophore search: A contribution to virtual screening. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 2894–2896.

- (15) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128–136.
- (16) 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.
- (17) The World Drug Index is available from Derwent Information, 14 Great Queen St., London W2 5DF, UK.
- (18) Gusfield, D. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*; Cambridge University Press: 1997; Chapter 11.
- (19) James, C. A.; Weininger, D.; Delaney, J. Daylight Theory Manual Daylight 4.82; Daylight Chemical Information Systems, Inc.: Los Altos, 2003.
- (20) Daylight Chemical Information Systems, Inc.: Los Altos. <http://www.daylight.com>.
- (21) Sanofi: EP658546 1994.
- (22) Meschler, J. P.; Kraichley, D. M.; Wilken, G. H.; Howlett, A. C. Inverse Agonist Properties of N-(Piperidin-1-yl)-5-(4-chlorophenyl)-1-(2,4-dichlorophenyl)-4-methyl-1H-pyrazole-3-carbazoxamide HCl (SR141716A) and 1-(2-Chlorophenyl)-4-cyano-5-(4-methoxyphenyl)-1H-pyrazole-3-carboxylic Acid Phenylamide (CP-272871) for the CB1 cannabinoid project. *Biochem. Pharmacol.* **2000**, *60*, 1315–1323.
- (23) Eli Lilly and Company: WO9602248 1996.
- (24) Aventis Pharma S. A.: WO001609 2000.
- (25) Merck & Co., Inc.: WO03007887 2003.
- (26) Sanofi: WO9719063 1997.
- (27) Bayer Pharmaceuticals Corporation: WO0340107 2003.
- (28) Virginia Commonwealth University; Organix, Inc.: US6509367 2003.
- (29) Solvay Pharmaceuticals B.V.: WO03027076 2003.
- (30) Solvay Pharmaceuticals B.V.: WO0170700 2001.
- (31) AstraZeneca AB: WO03051850 2003.
- (32) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
- (33) Leach, A. R.; Green, D. V. S.; Hann, M. M.; Harper, G.; Whittington, A. R. SIV: A synergistic approach to the analysis of high-throughput screening data. *Abstract of Papers*, 221st National Meeting of the American Chemical Society; American Chemical Society: San Diego, CA, 2001; 080-CINF.
- (34) The NCI AIDS data may be downloaded from http://dtp.nci.nih.gov/docs/aids/aids_data.html.
CI049860F