Detection and Assignment of Common Scaffolds in Project Databases of Lead Molecules

Alex M. Clark* and Paul Labute

Chemical Computing Group, Inc., 1010 Sherbrooke Street West, Suite 910, Montréal, Québec, Canada H3A 2R7

Received September 3, 2008

A method is presented for the detection and analysis of multiple common scaffolds for small collections of pharmaceutically relevant molecules that share a set of common structural motifs. The input consists of the molecules themselves, possibly some of the scaffolds, and possibly information about the relation between the substitution points of these scaffolds. Three new algorithms are presented: multiple scaffold detection, common scaffold alignment, and scaffold substructure assignment. Each of these steps is relevant for cases when either none, some, or all information about the common scaffolds and their substitution patterns is available. Each of these problems must be solved in an optimal way in order to produce useful structure−activity correlations. The output consists of a collection of scaffolds, a common numbering system, and a unique mapping of each molecule to a single scaffold substructure. This information can then be used to produce data for structure−activity analysis of medicinal chemistry project databases.

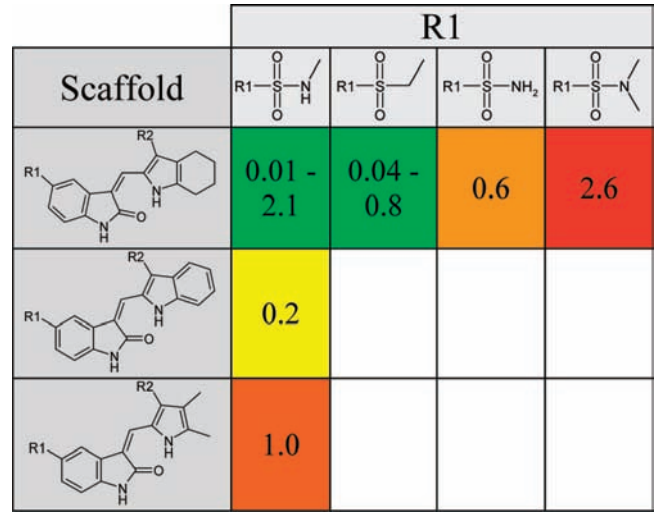
Introduction

A common theme in drug discovery projects is the accumulation of groups of small molecules based on several scaffold templates, which share some degree of structural similarity. By varying the composition of the scaffold, and the substituents with which the scaffolds are adorned, it is possible to extract trends that correlate ligand structure with activity against a chosen target. A popular method for elucidating structure−activity relationships is to represent the scaffold as a fragment, adorned with numbered R-group labels, and plot the pairwise combinations of possible substituents at each position in a series of tables. This approach has been described recently by Agrafiotis et al.,[1] and references therein.

When the data set is small, and the nature of the scaffolds is well understood, the structure−activity analysis can often be done manually or with partial automation using simple algorithms. Such circumstances often arise when a database has been constructed from a template synthesis, and the chemical structure changes have been recorded as part of a reaction description. For larger or less well annotated data sets, particularly those that are derived from studies not performed in-house, the prerequisite information necessary to examine the structure−activity relationships may not be readily available.

Furthermore, project databases often consist not only of molecules based on a single scaffold substructure but an analogous series of scaffold substructures, which can be labeled using a common R-group naming scheme, because they are presumed to bind in an analogous way. Describing such a database with multiple, related scaffolds introduces the scaffold itself as another degree of freedom, which can be tabulated in a manner analogous to the R-groups themselves. For example, when a group of Src inhibitors are each assigned to one of three scaffolds, which are clearly related, it is possible to plot scaffold vs R1, which is both visually revealing and suitable for Free-Wilson analysis:[2]

The cells for which the scaffold and the R1 substituent have available activity data are shown with activity against Src, measured as IC_{50} (μM),[3] and are also color-coded using a "heat-map" scheme, where the best binding affinities are green, the worst are red, and those in between are yellow. Those for which no compounds have been measured are left blank. Tables such as these can greatly simplify the process of gaining insight as to which scaffolds and substituents should be combined to improve the activity of lead compounds because it is easy to see which scaffolds and substituents are typically associated with high binding affinity and which of the most promising regions of the chemical space remain unexplored. By taking particular care to ensure that the scaffolds and substituents are depicted in such a way that common orientations are shown, the consequences of structural modifications become readily discernible to the viewer.



Whether the project database contains a single scaffold or an analogous series of scaffolds, it is necessary to ensure that the following information is available before such structure−activity analysis can be carried out:

(1) The *scaffold*, or *group of scaffolds*, upon which the majority of the molecules are based;

(2) In the case of multiple scaffolds, an *alignment* of the series which identifies analogous substitution points;

(3) For each of the database molecules, the scaffold which applies to it, and the atom-to-atom *substructure assignment*.

There are circumstances where none of these problems are particularly difficult to solve, e.g., when there is a single scaffold

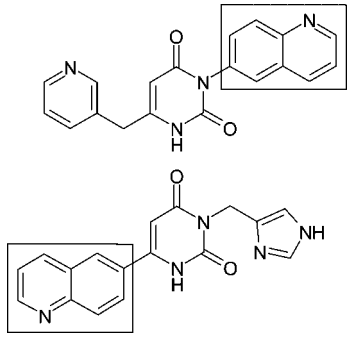
* To whom correspondence should be addressed. Phone: +1 514 393 1055. Fax: +1 514 874 9538. E-mail: aclark@chemcomp.com.

10.1021/jm801098a CCC: $40.75
Published on Web 12/18/2008

that is already known, and has no symmetry, and each molecule matches the scaffold in exactly one way. This being the case, structure−activity analysis of a minimally annotated database requires no algorithms more complex than substructure searching. However, there are a large number of practical situations when not all of these conditions are met. Scaffolds are not always known ahead of time, particularly for structures obtained from external sources, and automated methods for reconciling a series of multiple analogous scaffolds are difficult to come by. Furthermore, scaffold substructures are often symmetrical or degenerate, and the existence of multiple scaffolds introduces the possibility that a molecule may match more than one.

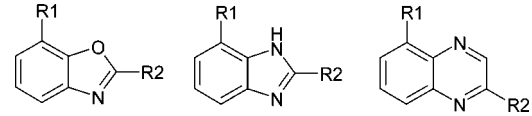
In general, an effective structure−activity analysis of a project data set requires a minimal number of common scaffolds, with an optimal alignment that conveys meaning throughout the series, and given multiple possibilities, an optimal assignment of scaffold substructures to the database constituents. The scaffolds should be selected for *scaffold-like* properties, and the variation of substitutents at analogous substitution points should be as low as possible in order to elucidate structure−activity trends. To effectively examine a comprehensive variety of databases of structure−activity data, it is necessary to solve the problems of scaffold detection, alignment, and assignment, in a stepwise fashion. Also, for pragmatic purposes, it is highly beneficial to design a method for solving these three problems such that any preexisting data is taken as definitive, and where possible, used to influence other decisions made during the process.

Topological scaffold detection is an issue that has been given significant attention in the literature,[4] though because it is surprisingly difficult, efforts to date tend to be based on the assumption that all molecules in the database are based on one single scaffold. A popular method for eliciting such information is the use of an implementation of the maximum common subgraph (MCS[a]) algorithm, which, though formally not solvable in polynomial time, can be effectively approximated for labeled chemical graphs.[5] The problem with this approach, other than being effectively limited to pairwise comparisons, is that it is notoriously easy to fool. While the largest contiguous common substructure between two chemical structures is frequently a good candidate for a scaffold, it is simple to propose an example where this is not the case, e.g.:



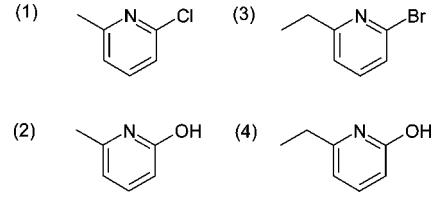
To the trained eye, the scaffold is obvious: the central six-membered ring with its two exocyclic carbonyl groups (8 heavy atoms) is almost certainly the best choice. However, most variations of MCS, as applied to chemical graphs, would find that the largest contiguous common fragment is the quinolyl moiety (10 heavy atoms). The frequency of such poor choices resulting from pairwise MCS analysis is high, and so additional strategies must be formulated, e.g., adding a large and complex collection of rules to bias the search in favor of medicinal chemistry applications or performing a pairwise comparison of all molecules in order to analyze the results statistically (requiring $^1/_2N(N-1)$ calculations, all of them costly). Extending such approaches to include the possibility that there are multiple common scaffolds within the data set, as well as decoy molecules that are based on scaffolds that are rare or unique, imposes escalating demands which mandate a simpler and more efficient approach.

One of the primary benefits of elucidating the relevant common scaffolds is that they are likely related in a way such that some of the substitution points are analogous throughout the series. Consider for example the following three templates:

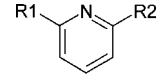


It might be imagined that for molecules based on any of these three scaffolds, the binding site orientation of active molecules might be such that the substituents labeled as R_1 and R_2 are oriented in a similar direction, providing a valid motivation for common annotation. While the substitution points are clearly indicated in the diagram shown above, if the scaffolds were calculated or provided without the Markush-style substituent notation, it is not at all unreasonable to expect that an unsupervised algorithm could align these structures to each other, such that the common relation is between them is evident. The structure composition often contains many clues as to what an optimal alignment might be but also the possible ways in which these substructures can be matched to the input molecules can provide a wealth of further clues because the variety of possible substituents and substitution points is usually restrictive.

Once the scaffolds are all known, and the alignment between them has been resolved, the final challenge is assignment of the scaffolds to the input molecules in a unique way, which can be deceptively troublesome. Because a substructure search does in many cases yield a unique result for each molecule, this is frequently all that is required to solve the problem. However, if any of the scaffold substructures has permutational symmetry, or a molecule has multiple occurrences of a scaffold substructure, or matches more than one scaffold, then the circumstances are less forgiving. A trivial example is the following data set, consisting of 4 molecules:



An obvious candidate for a scaffold substructure is the 2,6-disubstituted pyridine template:



Unfortunately, the pyridine substructure can match each of the constituent molecules in two ways due to internal symmetry, which means that the total number of ways to assign the scaffold to the input molecules is $2^4 = 16$. A quick examination of the

[a] Abbreviations: MCS, maximum common subgraph; UQO, unary quadratic optimization; SVL, scientific vector language.

input molecules reveals that, although these 16 states contain much degeneracy, they are not all equal if the objective is to have the minimum number of possible R-groups at each position. If molecules (1) and (2) are assigned such that $R_1 = CH_3$, and molecules (3) and (4) are assigned such that $R_1 = CH_2CH_3$, then the total diversity is such that $R_1 \in [CH_3, CH_2CH_3]$ and $R_2 \in$ [Cl, Br, OH], which is optimal by this criterion. An arbitrary assignment would most likely produce a less homogeneous set, which would obscure attempts to understand the structure−activity relationships. While this data set is sufficiently small to examine all possibilities, most data sets are large enough that enumerating an exponential number of potential solutions is not viable. Scaffolds with internal symmetry, and other causes for assignment degeneracy, are sufficiently common that this problem needs to be addressed in order to make best use of structure−activity data.

If a scaffold analysis method fails to find the most relevant common scaffolds, or does not perceive the way in which the scaffolds are related, or does not propose an effective way of mapping scaffolds to molecules when the possibilities are degenerate, then attempts to understand correlations between structure and activity will at best be hindered by high signal-to-noise ratio if not swamped or entirely meaningless.

The remainder of this work describes the implementation of algorithms to solve the scaffold detection, alignment, and assignment problems, for moderately sized project databases, followed by a number of case studies that identify the various difficulties involved in scaffold analysis and the manner in which the algorithms overcome these difficulties.

Methods

The three important algorithms described in this section are outlined as follows:

(1) Scaffold Detection. Inputs: *molecule database, optional prespecified scaffolds*. Outputs: *list of scaffold substructures*.

(2) Scaffold Alignment. Inputs: *list of scaffold substructures, optional atom numbering hints.* Outputs: *common atom numbering scheme for all scaffolds.*

(3) Scaffold Assignment. Inputs: *molecule database, list of scaffold substructures, common atom numbering scheme*. Outputs: *mapping to one scaffold substructure for each molecule.*

The scaffold detection algorithm is used to find the set of scaffolds that best suits the data set or to look for additional scaffolds in the event that scaffolds were supplied by the user. The scaffold alignment algorithm is used to find correlation mappings between the scaffolds, in the event that there is more than one, and any of the scaffolds that were prespecified did not already contain a common numbering scheme.

The scaffold assignment algorithm makes use of all the available information (the input molecules, the list of scaffolds, and the relation between the scaffolds) to assign each of the input molecules to a single scaffold with a single mapping. The first two steps can be controlled by the user if scaffold information is known a priori, but the final assignment step is necessarily fully automated. Once the series is complete, the database can be studied using standard cheminformatics techniques in order to break down the list of R-groups at each position and chart the effects that interchanging substituents and scaffolds has on activity.

1. Scaffold Detection. For our analysis of molecules, we begin by applying a method recently described by Schuffenhauer et al., which produces a linear fragmentation sequence by iteratively cleaving substituents from a molecule, starting with the least *scaffold-like*.[6] Figure 1 illustrates the process. The first step is to remove terminal substituents, while the following steps involve removing pendant or fused ring blocks, in reverse order of the likelihood that the removed component is a part of the core scaffold.

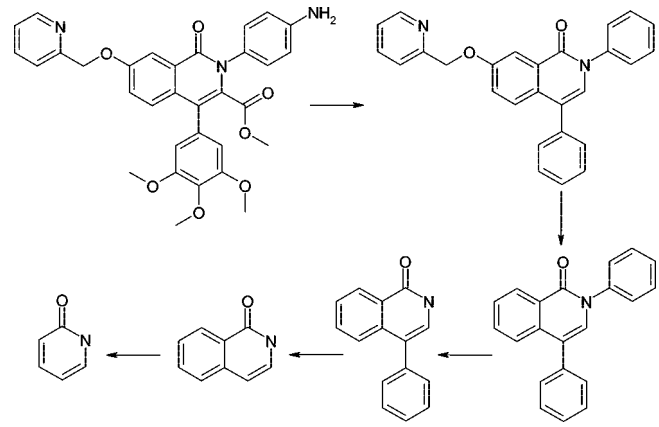


Figure 1. The fragmentation pattern of a single molecule, according to the Schuffenhauer scheme.

If a collection of molecules are decomposed in this way, and represented as a string of molecular fragments, smallest first, and sorted after converting into a canonical string identifier (such as SMILES),[7] then the rows and columns can be decomposed into a tree hierarchy, such as is illustrated in Figure 2. In this example, seven molecules, for which the terminal substituents have been removed for clarity, are decomposed into fragments and sorted, with common ancestors merged into parent nodes.

As can be seen, all of these molecules fragment into a single root fragment, (a), which consists of a single six-membered ring. The root fragment has two descendents, (b) and (c). Because fragment (c) has just one descendent molecule, and (b) has six, it is worth considering the possibility that (b) is a better scaffold than (a), even though it has one less descendent, because it is larger and less generic. The principle is to look for large fragments with low symmetry and high membership, whose immediate descendents have much smaller and more fragmented descendents.

The method of scaffold selection used in this work is based on assigning a score for each distinct possible fragment is given as:

$$O \cdot N \cdot \frac{U}{Y} \cdot \left(\frac{\sum_a S_a}{A} + 1 \right)$$

where O is the fraction of the remaining molecule set which contains the fragment; N is the number of heavy atoms in the fragment; U is calculated by counting the maximum number of bonds in all possible fragment substructures (i.e., theoretical maximum number of R-groups): if the largest number of severed bonds is 2 or 3, then $U = 3$, if the largest number is 1 or 4, then $U = 2$, otherwise $U = 1$; Y is the number of ways the structure can be topologically permuted onto itself, plus one; A is the number of fragments selected in previous iterations; and S_a is the similarity of the proposed fragment to each previously selected fragment, calculated as the Tanimoto coefficient using the MACCS fingerprints.[8] The final term is set to 1 for the selection of the first fragment.

The first scaffold is selected by picking the fragment with the highest score. Molecules that feature this substructure are removed from the set, and the remaining fragments are rescored. When the best score is less than 1, or no molecules remain, the scaffold selection stops.

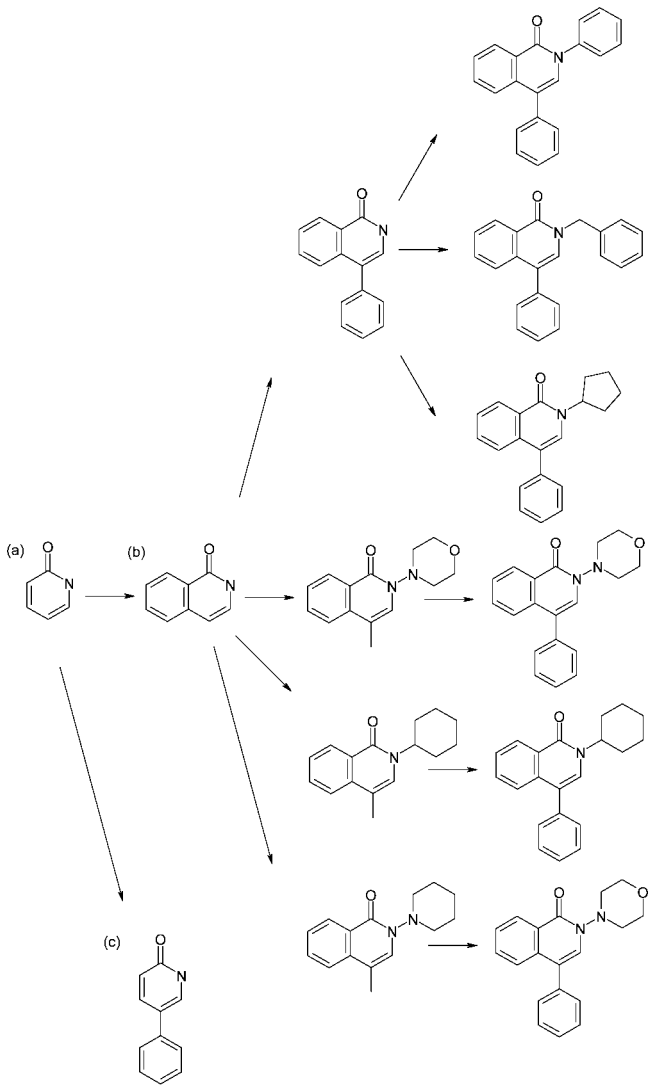


Figure 2. Linear fragmentation pattern of a set of molecules, presented as a tree, with a single root.

If some of the scaffolds are already known, then all molecules that are found to contain any of the provided scaffolds as a substructure are excluded prior to the determination of additional scaffolds.

The use of linear fragment decomposition is not only conceptually simple, but it is also very fast and runs with effectively linear time dependence on database size, given that the rate limiting step is the generation of fragments and conversion into canonical identifiers for rapid comparison.

The simple scoring function, which is primarily driven by collective population, makes the method somewhat resistant to the vagaries of graph comparisons, even for input data sets that contain a significant number of stray molecules not based on any reasonable notion of a common scaffold. Even for homogeneous inputs, molecules with a particularly *scaffold-like* substituent may be fragmented differently to the majority of molecules based on the same scaffold whose substituents are not *scaffold-like*. As long as such cases are in the minority, however, they will appear in the tree hierarchy as small clumps with a unique or rarely occurring root fragment, which typically score well below the threshold, so that their constituent fragments will not be selected as scaffolds. As will be seen in the following sections, the final assignment of scaffolds to molecules is not dependent on the origin of the scaffold, so molecules that were fragmented in a way that did not reveal its common scaffold can still be correctly assigned as long as the appropriate common scaffold was discovered elsewhere.

2. Scaffold Alignment. When a database of drug activity data contains more than one scaffold, it becomes necessary to develop a notion of how these scaffolds relate to each other so that the analogous substitution points can be identified.

In cases when the scaffolds are provided and fully annotated by the user, typically using the Markush-style R-group notation, it is taken as given that the numbering system is consistent across the series of scaffolds, i.e., the substitution point labeled as R_1 on one scaffold is conceptually equivalent to the substitution points denoted as R_1 on the other scaffolds. When the scaffolds have been detected automatically, or the use of R-groups is insufficient to correlate all substitution points, an analysis of the scaffolds must be performed in order to find common points of reference across the series.

To do this, it is necessary to make use of all available hints, which includes: (1) any partially prespecified mapping information, (2) molecular graphs, (3) chemical properties (atom and bond labels), and (4) information derived from all possible ways in which the scaffolds can be mapped as substructures of the input molecules.

For aligning a set of scaffolds, the working property is taken to be the mapping number for each of the non-hydrogen atoms of the scaffold, where the default value for each of these atoms is *unassigned*. In cases where prespecified scaffolds with Markush-style annotations have been provided, for each R-group label, the atom adjacent to the label is assigned a mapping number equal to the R-group number, and the label placeholder atom is removed. In most cases, not all of the scaffold atoms are adorned with a label-bearing substituent, and so user-annotated scaffolds will begin in a partially assigned state, which may or may not be sufficient to describe all of the possible substitution positions subsequently encountered.

Because the information contained in the chemical graph of the scaffolds is often insufficient to make a reasonable guess as to how two scaffolds relate to each other, it is necessary to acquire additional information by preanalyzing the ways in which the scaffolds can occur as substructures for all of the input molecules, which is done by the following algorithm:

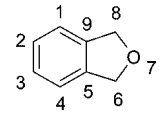
(1) For every scaffold, iterate over all of the molecules;

(2) Find all of the ways in which the scaffold substructure can be matched to the molecule;

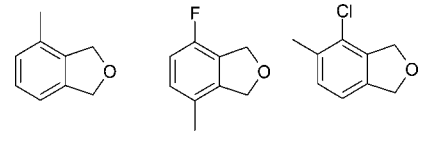
(3) Collect all of the implied substituents that would be created by removing the scaffold substructure and note the atom indices of the attachment points;

(4) For every scaffold, compile a list of the unique substituents connected at each atom position.

For example, consider the following scaffold, with atom ordering as shown:



in a database containing the molecules:



The union of substituent possibilities for all degenerate substructure matches are: $A_1 \in [CH_3, F, Cl]$; $A_2 \in [CH_3]$; $A_3 \in [CH_3]$; $A_4 \in [CH_3, F, Cl]$; the substituents of A_5, A_6, A_7, A_8, and A_9 all form empty sets. For this purpose, that being coarse scoring, a hydrogen atom is equivalent to the absence of substituents. Because the scaffold in this example is symmetrical, each of the molecules maps to it twice, so A_1 and A_4 are equivalent as are A_2 and A_3. As the scaffold alignment proceeds, a significant distinction is made between a scaffold atom, which does not bear substituents in any of its possible mappings, and one which does. The actual list of possible substituents is used where possible to further distinguish between putative alignments.

Once preprocessing is complete, the scaffold alignment algorithm operates as follows:

(1) For each pair of scaffolds (*i*, *j*), where $i < j$, find and record the best way to partially align the scaffolds *i* and *j*;

(2) Exclude any of the pairwise partial alignments that did not produce a valid result, and stop if none remain;

(3) For each qualifying partial alignment between scaffolds (*i*, *j*), assign a score;

(4) Pick the scaffold pair (*i*, *j*) with the best score and use the partial alignment to update the atom mapping numbers for both scaffolds;

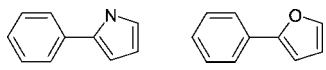
(5) Goto 1.

For each pass of the main loop, the pairwise combinations of scaffolds are enumerated and each is examined in order to find a good way to map scaffold(*i*) onto scaffold(*j*). If no reasonable mapping is found, or the only reasonable mappings require that previously determined mapping numbers clash, then the result is invalid. Similarly, if the best mapping includes only atoms that share the previously assigned mapping number or none of the freshly assigned atoms are suitably adorned by potential substitution points, the result is not stored.

If none of the pairwise comparisons are valid, the loop is terminated. Otherwise, the mapping that yielded the best score is selected. The atoms that were mapped from scaffold(*i*) to scaffold(*j*) are compared with regard to mapping number: if the atom in scaffold(*i*) has a defined mapping number but the atom in scaffold(*j*) does not, the number is copied from *i* to *j*, or vice versa. If neither atoms have a mapping number, then a previously unused number is assigned to both.

Should the loop be terminated before all atoms are given a mapping number, then unique mapping numbers are selected for each unassigned case.

Determination of the best-possible mapping between two scaffolds, either or both of which may be fully or partially populated with preexisting mapping numbers, begins by considering the possibility that the two heavy-atom graphs are identical, such as is the case for:



In the above example, there are four ways to map the two graphs onto each other if the chemical labels are ignored. Each of these isomorphisms is checked to see if previously assigned mapping numbers clash, and when this occurs, these mappings are rejected. The remaining mappings are subjected to a score, which is computed as according to the following scheme:

+1 for each mapped atom;

+1 for each mapped atom with the same element;

+0.1 for each mapped atom with the same number of heavy neighbors on both sides;

+0.1 for each mapped atom for which both sides would have a substituent in at least one possible substructure match;

+0.1 for each identical substituent shared at a substitution point;

+1 for each mapped bond;

+1 for each mapped bond of same type;

+5 for each ring fully and contiguously mapped on both sides.

If any graph isomorphism is valid, then the highest scoring case is returned. Otherwise, the next phase is activated, which is a greedy search to find the mapping with the highest score, as described above, with sufficient constraints to prefentially identify the most promising candidates first.

In principle, it may be considered necessary to examine all possible mappings of atoms between the two scaffolds, which has an exponential time dependence. However, because the value of assigning atoms such that they are contiguous on both sides is so high, then if several atoms from the optimal solution can be correctly locked into place, the remainder can usually be found quickly by sequentially adding atoms to this set. Additionally, it is usually the case that analogous scaffolds have graphs that share at least one ring. For example, if the unlabeled graphs of two scaffolds both have a single six-membered ring, then there are 12 ways of superimposing these rings. The mapped sets of size 6 can be scored and ranked according to the score described for graph isomorphisms, i.e., common elements, bond types, and substitution positions will favor large, contiguous mappings that have more common functionality.

The algorithm proceeds as:

(0) Seed the list of partial alignments with all possible ways to map a single small ring from one scaffold to the other;

(1) Score and rank all of the postulated partial alignments, and keep the best 50;

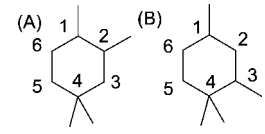
(2) For each partial alignment that was added in the previous step, grow the partial alignment by the following methods: (a) chain growth, (b) ring growth;

(3) Add nondegenerate new partial alignments that do not clash with preassigned atom mapping numbers;

(4) If new partial alignments were found, Goto 1.

At each stage of the iteration, the existing mappings are explored for valid ways to make them larger. The enlarged mappings are added to the collection, which is scored and ranked. Because the score improves when atoms and bonds are added, larger mappings have a tendency to displace smaller mappings, encouraging the search to find a more complete alignment between the two scaffolds.

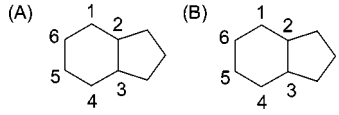
Two methods are used to grow an existing mapping and postulate some number of larger overlaps, which will subsequently be checked for validity and then assigned a score. Chain expansion is straightforward, and is performed one atom at a time. For every pair of mapped atoms, the nonmapped atoms that they are bonded to in either structure are enumerated and the pairwise sets of expansions are submitted. For example, for the following two structures, which are annotated by their partially assigned mapping numbers:



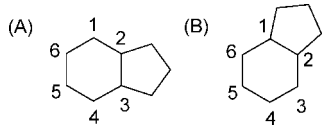
the atoms mapped as 1 both have a single substituent, and so a single additional mapping possibility is generated. The atoms mapped as 2 and 3 each have a substituent on one of the structures but not the other, so no additional mapping possibilities are generated. The atoms mapped as 4 each have two

substituents, so all four of the possible pairwise combinations are added to the list.

New possibilities generated by ring expansion are added whenever possible. A pairwise list of bonds mapped in both structures is generated, and this list is restricted to those that are part of a fully mapped ring, on both sides. For each such pair of qualifying bonds, for both structures, a list is made of the rings that share an edge with the bond but are not fully mapped. If for both structures the set is nonzero, then these rings are compared pairwise. In the following simple example, both structures share a six-membered ring that has been mapped between the two, coincidently in the optimal way. The bond between the atoms mapped as 2 and 3, in both cases, happen to have five-membered rings fused to them, which are presently unassigned:

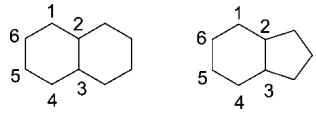


Because the rings are of the same size, the expanded mapping can be unambiguously added. In the next example:

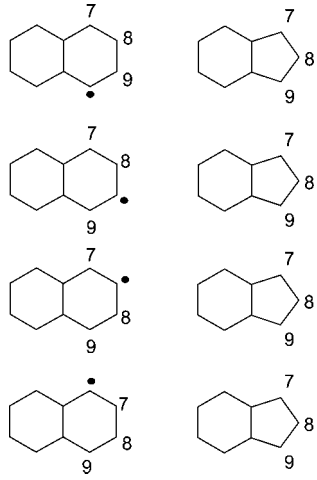


no further mappings are generated because the unassigned fused ring shares an edge with atoms 2 and 3 in structure A and with 1 and 2 in structure B.

In the following example:



both partial mappings possess as yet unmapped rings that share an edge with atoms 1 and 2. Because they differ in size by one, it is worth postulating a new mapping for each of the ways in which the two rings can be mapped together, with the six-membered ring leaving one atom unmapped. For the 6:5 combination, there are four possibilities, each of which is registered as a possible expansion:

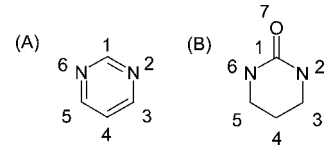


In most cases, the mapping of additional rings is achievable by exploring a sufficiently large number of single unit growth steps, but in practice, matching ring blocks to each other is the most effective way to quickly generated the best-scoring alignments as early as possible, which improves the likelihood of obtaining a high quality match before the cutoff buffers fill up.

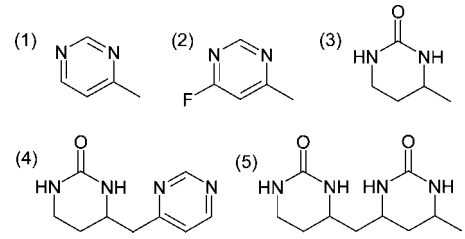
3. Scaffold Assignment. Once the scaffolds are known and their alignment relative to one another has been determined, it is necessary to find out for each input molecule which scaffold, if any, it should be assigned to and, in the case of multiple mappings, which of those should be selected.

The trivial case is commonly encountered, that being when each of the input molecules contains only one of the scaffolds as a substructure, and there is only one way to map this substructure or the molecule matches no scaffolds at all. If this is true for all of the input molecules, then no further effort need be expended. If this is not the case, however, deciding upon the best combination of scaffold assignments has a formally exponential time complexity, and cannot realistically be explored exhaustively nor can the problem reasonably be ignored.

Consider the following scaffolds, where the alignment between the atoms in both cases is indicated with numeric labels:



When applied to a database containing the following hypothetical molecules:



it is clear that no aspect of scaffold assignment is unambiguous. Molecule (1) matches the first scaffold but does so in two ways because the scaffold has symmetry: one match is such that atom 3 has a methyl substituent, and the other places the methyl substituent on atom 5. Molecule (2) matches the same scaffold, and substituents 3 and 5 can be assigned to methyl and fluoro, or fluoro and methyl, respectively. In isolation, these choices are arbitrary, but when considered together, it is far more useful to assign the mappings that place the common methyl group in the same position in both cases, e.g., atom 3 is always methyl substituted, while atom 5 can be attached to either a hydrogen or a fluorine atom.

Molecule (3) matches the second scaffold, but because the two scaffolds were aligned so that they have a high degree of commonality in their numbering system, the decision as to whether to assign the methyl substituent to atom 3 or atom 5 is not independent of the decisions made regarding the previous two molecules, i.e., if the first two molecules were assigned such that atom 3 featured a methyl substituent, then it would be desirable to do likewise for molecule (3) because the scaffolds have previously established positional equivalence.

Molecule (4) matches both of the incoming scaffolds, while molecule (5) has two occurrences of the second scaffold and both have internal symmetry. The choice of which scaffold to assign a molecule to could be assisted by the possibility of other

preferences, e.g., favoring larger scaffolds, more-buried substructure matches, the production of not too few and not too many implicit substituents, etc. However, this information is seldom sufficient to make the best choice overall, and the selections still need to be made in the context of a global optimum. In this example, if molecule (4) were assigned to the first scaffold such that the implied substituent of atom 5 is the cyclourea ring, and molecule (5) were assigned to the second scaffold such that atom 5 bears the same incarnation of the cyclourea ring and atom 3 is methyl substituted, then this can be combined with nominal assignments for the first three molecules, which produces an assignment result with optimally low substituent diversity.

To address the problem of scaffold assignment, we consider all of the possible combinations of assignments and pick the combination for which the diversity of substitution patterns is minimal. A naïve way to solve this problem is simply to perform a brute-force comparison; while this is appropriate for very small data sets, or data sets in which the degrees of freedom is very low, the number of combinations of assignments is exponentially proportional to the number of molecules that have nonunique solutions, and so a more sophisticated approach is required. In general, each of the molecules offers one or more possibilities, or *states*, where a state is a specific atom-by-atom assignment to a particular scaffold substructure, and for each of the molecules exactly one state must be selected.

More precisely, for every possible scaffold substructure match, a candidate state is produced, which consists of the following information: (a) a molecule index, (b) a scaffold index; (c) an atom-to-atom substructure mapping, and (d) the composition and location on the scaffold of each of the implied substituents. Figure 3 shows the possible assignments for the example used at the beginning of this section, the substructure matches for which divide the five molecules into a total of 14 states, each of which has a specific atom-to-atom mapping to one of the scaffolds, and a collection of implied substituents at each of the mapped positions.

For each pair of states, *i* and *j*, a score is calculated reflecting the compatibility of the assignment; this score is 0 if the two states have the same molecule index, otherwise $-S_{ij}$, the negative of a measurement of similarity between the implied substituents.[9] The pairwise scores are arranged into a symmetric matrix **S**. Let **x** denote a binary vector with each element corresponding to a particular state; a value of "1" means that a particular state is chosen and a value of "0" means that a particular state is not chosen. An *admissible* vector is a binary vector for which collection of states with the same molecule index has exactly one state chosen, i.e., exactly one "1" value in the vector elements corresponding to that molecule index. The scaffold assignment problem is then solved by computing the admissible vector **x** that minimizes the multidimensional quadratic function $\mathbf{x^TSx}$. This combinatorial optimization problem is an instance of unary quadratic optimization (UQO) and in the present work we use the large-scale UQO solver used in Protonate3D;[10] alternatively, dead end elimination[11] techniques can be used. In practice, those molecules with only a single produced state (singletons) are separated from the others because these singletons affect only scoring. The optimization is then cast into the mathematically equivalent form $s_0 + \mathbf{s^Tx} + {}^1/_2\mathbf{x^TSx}$ were **x** now contains only nonsingleton molecules, **S** contains the pairwise scores between states corresponding to nonsingleton molecules, the **s** vector contains the scores with respect to the singleton molecules (the "self" scores), and s_0 is a constant.

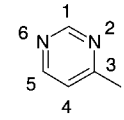
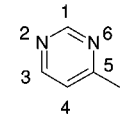
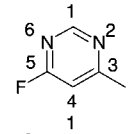
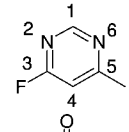
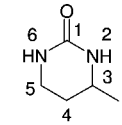
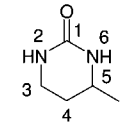
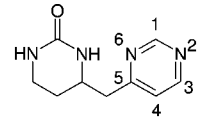
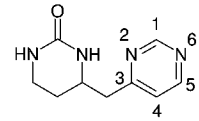
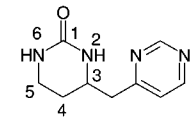
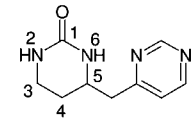
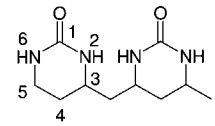
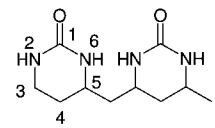
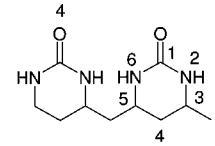
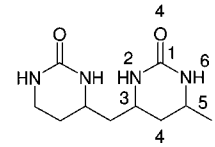
state	molecule	scaffold	mapping	substituents
1	1	1		A_3=CH_3
2	1	1		A_5=CH_3
3	2	1		A_3=CH_3 A_5=F
4	2	1		A_3=F A_5=CH_3
5	3	2		A_3=CH_3
6	3	2		A_5=CH_3
7	4	1		A_5=$C_5H_9N_2O$
8	4	1		A_3=$C_5H_9N_2O$
9	4	2		A_3=$C_5H_5N_2$
10	4	2		A_5=$C_5H_5N_2$
11	5	1		A_3=$C_6H_{11}N_2O$
12	5	1		A_5=$C_6H_{11}N_2O$
13	5	2		A_3=CH_3 A_5=$C_5H_9N_2O$
14	5	2		A_3=$C_5H_9N_2O$ A_5=CH_3

Figure 3. Assignment of states to specific scaffold mappings.

Figure 4 shows a matrix of the 14 states produced by the example used in this section. Each element in the matrix is marked with the pairwise score of the two states, except those that belong to pairs of states that are both from the same

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	n/a		-1.00	-0.00	-1.00	-0.00	-0.00	-0.03	-0.07	-0.00	-0.03	-0.00	-1.00	-0.03
2			-0.00	-1.00	-0.00	-1.00	-0.03	-0.00	-0.00	-0.07	-0.00	-0.03	-0.03	-1.00
3	-1.00	-0.00	n/a		-1.00	-0.00	-0.00	-0.03	-0.07	-0.00	-0.03	-0.00	-1.00	-0.03
4	-0.00	-1.00			-0.00	-1.00	-0.03	-0.00	-0.00	-0.07	-0.00	-0.03	-0.03	-1.00
5	-1.00	-0.00	-1.00	-0.00	n/a		-0.00	-0.03	-0.07	-0.00	-0.03	-0.00	-1.00	-0.03
6	-0.00	-1.00	-0.00	-1.00			-0.03	-0.00	-0.00	-0.07	-0.00	-0.03	-0.03	-1.00
7	-0.00	-0.03	-0.00	-0.03	-0.00	-0.03	n/a				-0.00	-0.80	-1.00	-0.03
8	-0.03	-0.00	-0.03	-0.00	-0.03	-0.00					-0.80	-0.00	-0.03	-1.00
9	-0.07	-0.00	-0.07	-0.00	-0.07	-0.00					-0.39	-0.00	-0.07	-0.33
10	-0.00	-0.07	-0.00	-0.07	-0.00	-0.07					-0.00	-0.39	-0.33	-0.07
11	-0.03	-0.00	-0.03	-0.00	-0.03	-0.00	-0.00	-0.80	-0.39	-0.00	n/a			
12	-0.00	-0.03	-0.00	-0.03	-0.00	-0.03	-0.80	-0.00	-0.00	-0.39				
13	-1.00	-0.03	-1.00	-0.03	-1.00	-0.03	-1.00	-0.03	-0.07	-0.33				
14	-0.03	-1.00	-0.03	-1.00	-0.03	-1.00	-0.03	-1.00	-0.33	-0.07				

Figure 4. Construction of the unary quadratic pairwise terms. Each matrix element which is not forbidden by the *unary* constraint is assigned a number equal to the negative sum of the Tanimoto coefficients, using the MACCS fingerprint keys, between each of the implied substituents at analogous positions on the two scaffolds.

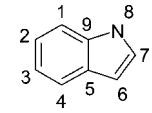
molecule. In this example, the self-score terms are all zero because all of the scaffolds have internal symmetry, which means that the solution to the scaffold assignment is based entirely on the pairwise score terms. The optimal solution shown in this case is the selection of states 1, 3, 5, 7, and 13, which are indicated in the matrix by light-gray shading. The pairwise score in this case is −14.0, which is obtained by adding up the values of the cells from each of the selected rows vs each of the selected columns. For this solution, the distribution of implied substituents is $A_3 \in [H, CH_3]$; $A_5 \in [H, F, C_5H_9N_2O]$. A second optimal solution exists in which the substituents connected to A_3 and A_5 are reversed.

In cases where some of the molecules match just one scaffold uniquely and the scaffold alignment mappings show common substitution points, the self-score terms will play a significant role in the selection of scaffold assignments for those molecules that have degenerate possible assignments. However, the method used to solve the unary quadratic optimization problem must take into account the possibility that pairwise score terms dominate the solution, with potentially degenerate multiple solutions, and a number of possible assignment combinations that grows exponentially with increasing data set size. Also, for data sets that have little or no overlap of implied substituents at their assigned scaffold positions, the solutions may be effectively partitioned into multiple groups of subsolutions whose outcomes do not significantly affect the others.

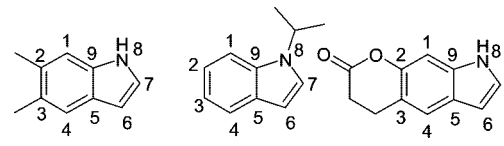
The final result from the optimization is that each of the molecules that contained any of the scaffolds as a substructure is now assigned to a single scaffold, and the mapping between the scaffold and the molecule is recorded.

4. Post-Assignment. Once the scaffolds are known, aligned, and optimally assigned to their corresponding molecules, the scaffold:molecule pairs can be analyzed using routine molecular graph methods. For each input molecule, removing the atoms assigned to the scaffold produces a list of substituents making up the remaining connected components. Recording the scaffold-invariant assignment identifiers to which each of these substituents was attached allows them to be analyzed collectively, and a list of R-group positions for the database can be assigned. The molecular compositions of R-groups, whether terminal or cyclic, are available for use in structure−activity analysis.

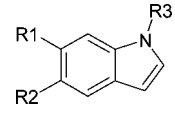
For example, if the scaffold:



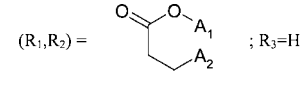
were mapped to the following three input molecules as-numbered:



then a collective analysis would reveal that scaffold atoms mapped as atoms 2, 3, and 8 are all adorned with non-hydrogen substituents in at least one case, and so the scaffold can be subsequently decorated with consecutively numbered R-groups at these positions, e.g.:



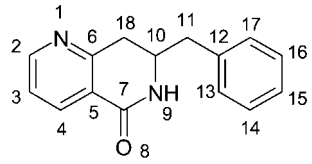
For the first molecule, R_1 = Me, R_2 = Me, R_3 = H, and for the second molecule, R_1 = H, R_2 = H, $R_3 = {}^{i}$Pr. For the third molecule, the chain disconnect would be expressed as:



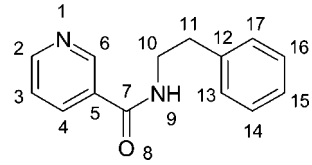
For purposes of subsequent analysis, it is advisable to record the composition of the substituents such that the attachment points are noted, allowing them to be unambiguously reconnected, as is shown for the cleaved-ring substituent above, where the numbers for A_n match the corresponding scaffold substituent point R_n.

In addition to collating and enumerating R-group substituents, it is also possible to make use of the fact that the scaffold alignment algorithm assigns matching identifiers to all of the scaffold atoms when it comes to providing 2D depiction layout coordinates. Consider the following two scaffolds, with their respective mapping indices:

(a)



(b)



If the scaffold with fewer degrees of freedom, (a), were to be depicted as shown above, a result that might be obtained from an unconstrained depiction layout algorithm, the depiction of (b) would do well to take into account the distances between analogous atoms found in the resulting layout for (a). For example, the pyridine nitrogen labeled as "1" could be placed on either the same or the opposite side of the amide functional group. This could be encouraged be affecting a bias toward distances between analogous atoms, such as "1" and "8". Similarly, the pendant phenyl ring should be arranged in a manner that is not optimal in terms of maximal dispersion but most closely reproduces distances found in the example. We have previously described 2D molecule depiction with additional distance constraints.[12]

5. Implementation. The scaffold detection, alignment, and assignment algorithms, as well as the Schuffenhauer fragmentation patterns, were implemented using scientific vector language (SVL), running under molecular operating environment (MOE).[13] The variant of substructure searching was implemented using distance matching constraints to postulate possible matches between chemical graphs, which had been preprocessed to convert Kekulé-style aromaticity into resonance form. Substructures that split double bonds or aromatic ring blocks were rejected. The unary quadratic optimizer used for scaffold assignment is a part of the SVL function library.

Performance metrics vary depending on the degeneracy of the scaffolds. Using contemporary hardware (e.g., single Intel x86 CPU, 3.0 GHz), data sets consisting of thousands of molecules can be analyzed in a matter of seconds when a small number of rigid scaffolds with no symmetry are detected. For larger sets of scaffolds with rotatable bonds, or nonsymmetrical mappings, the processing time becomes longer, and for cases when all the scaffolds are symmetrical or otherwise degenerate, the full UQO equation can take as long as several minutes to solve for data sets of up to a thousand molecules.

Results

Case 1: Homogeneous Single Scaffold. A collection of 136 molecules studied for inhibition of HIV-1 reverse transcriptase was assembled, selected examples of which are shown in Figure 5a. In this data set, there is a single scaffold framework, which can be readily detected by the observer:

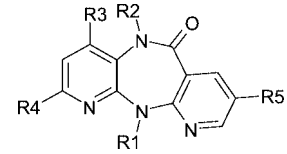
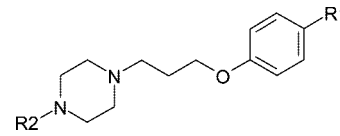


Figure 5b shows the root and stem of the tree of linear fragmentation patterns. All of the input molecules are ultimately peeled down to a single seven-membered ring, originally obtained from the 6:7:6 fused ring block that forms the core of each molecule. Note that from this base, there are two fused pyridine rings that may be peeled off stepwise, which introduces an issue of ordering. For 134 of the molecules, the ring adjacent to the carbonyl group was peeled first, and for the remaining two, the ring furthest from the carbonyl group was peeled first. The alternate fragmentation pattern was induced by the difference in the pendant substituents of both sides. The reason why these minority cases were fragmented in a different way is because of the presence of *scaffold-like* substituents, which influenced the peeling order.

Analysis of the fragmentation tree picks the best scaffold as (iii) because the scoring method is proportional to fragment size and to the number of descendents. While fragment (i) is the root of the whole database (136 counts), fragment (ii) has an additional four atoms, and only two less descendents. Fragment (iii) is better still, as it has another four atoms and the same number of descendents as its parent. Progressing any further along this branch, however, the score reduces drastically as the fragments split into a multitude of much smaller groups. Once fragment (iii) has been selected as the first candidate for a scaffold, it is found that the two molecules that did not descend from this fragment do in fact contain the selected fragment as a substructure, and so they are removed from the list of fragmentation sequences that could potentially lead to new scaffolds being selected, rendering the alternate fragmentation pathway a harmless artifact.

The scaffold assignment is, in this case, straightforward, because only one scaffold was detected, which contains no symmetry. Also, none of the input molecules contain multiple instances of this substructure, and so the matches are completely unambiguous.

Case 2: Misleading Nonscaffolds. A collection of 146 molecules tested for activity against the histamine H1 receptor (HRH1) features an optimal scaffold based on two six-membered rings joined by an aliphatic linker:



The chemical features of the scaffold do not distinguish themselves from other potential scaffold hypotheses, and as can be seen from parts a, b, and c of Figure 6, which show selected portions of three roots of the fragment decomposition tree, which indicates to some extent the apparent disarray of the fragmentation patterns. However, most of the molecules are peeled down to one of two roots: benzene (67) or piperazine (65), with nine other root fragments covering the remaining 14 molecules. The molecules that were reduced to a single benzene were fragmented in a way that does not lead to the most promising

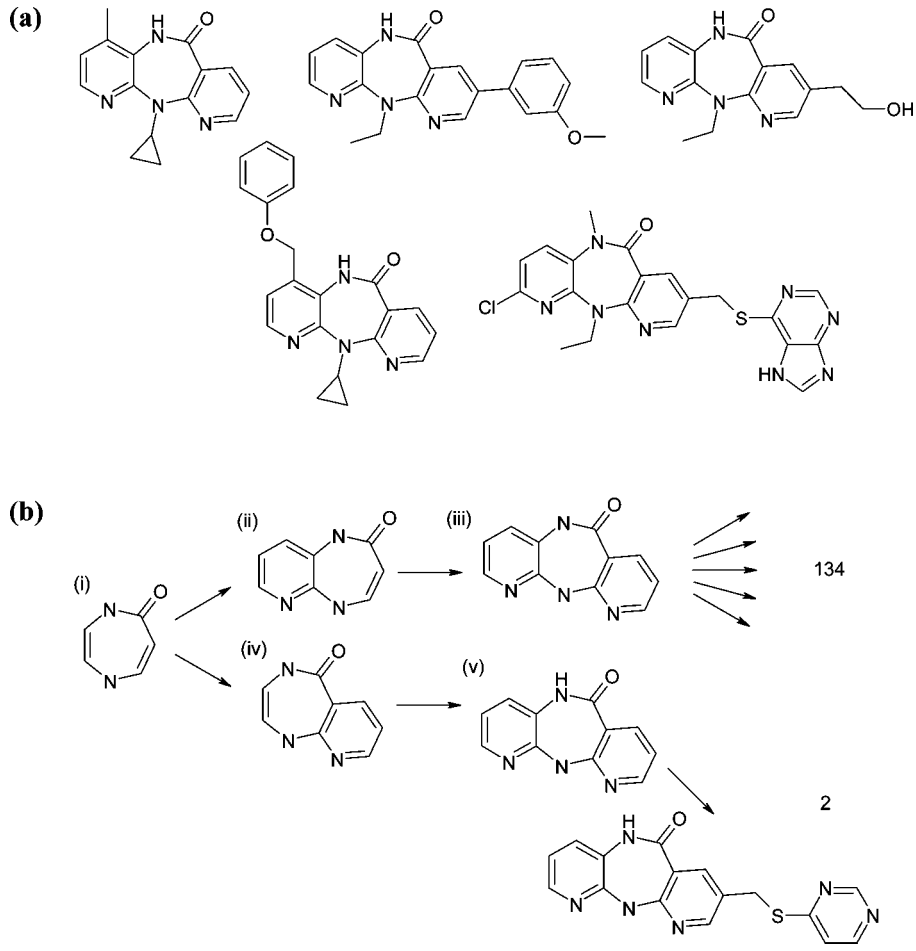


Figure 5. (a) Selected examples of molecules studied for possible HIV-1 reverse transcriptase inhibition. (b) The root of the fragmentation sequence and partial traversal of the tree.

scaffold, and this is evident by the way the root quickly branches into smaller groups. The molecules reduced to piperazine largely did so in a way which proceeds via the optimal scaffold: the first descendent shown has a group size of 38, which is much larger than any of the other sets and also has a fairly high atom count and lower symmetry than some of the other contending fragments. It is therefore selected as the first scaffold fragment. After molecules containing this substructure are removed, tentative scaffolds from the remaining fragments fall below the cutoff threshold due to a combination of low group size, high symmetry, and overly diverse possible substitution points.

Although the resulting single scaffold has 4 internal symmetry permutations, the final scaffold assignment is simple because all possible substructure mappings produce a degenerate collection of R-groups, i.e., the values of R_1 and R_2 are the same regardless of which degenerate substructure match is selected, and so duplicates are removed and the assignment does not require an extensive optimization step.

Case 3: Multiple Similar Common Scaffolds. A collection of 75 molecules with activity measurements against dihydrofolate reductase (DHFR) targets was assembled. Figure 7a shows the first two layers of the fragmentation tree, which reduces to two roots, one of which has two immediate descendents. During scaffold detection, the second-tier fragments are selected one at a time as the best scoring scaffold hypotheses, which completes the process.

Alignment of the scaffolds relative to one another is relatively straightforward because the chemical graphs without node or edge labels are equivalent, with 4-fold symmetry. Figure 7b indicates the manner in which the scaffolds are aligned to each other by decorating the substituent positions with R-groups. For any pairwise comparison of two scaffolds, the four possible ways to map them in a graph equivalent way quickly resolve themselves with the assistance of the ring shown on the left-hand side in each of the drawings, which matches element labels in all cases, then more so by the substituents. Parts c, d, and e of Figure 7 show some of the molecules based on these three scaffolds. It can be seen that the R_1 and R_2 positions are consistently placed, and share common values. The R_4 position is also consistently placed throughout the set, and some of the identities of the substituents are common between scaffolds, which further improves the rating of the match. All things considered, this is an example of when detection, alignment and assignment of multiple scaffolds is straightforward and easily resolved by a small amount of exploration.

Case 4: Ambiguous Common Scaffolds. A collection of 300 molecules with activity measured against the serotonin 5-HT_{2C} target was assembled. The three scaffolds determined from the set are shown in Figure 8, which match 159, 83, and 20 molecules, respectively, with 38 molecules remaining unmatched. In the diagram, the use of decorating R-groups, and the layout of the molecules, make the relative alignment of the scaffolds clear to the observer. However, the relation between these scaffolds is considerably less obvious to an algorithm. First of all, the central urea moiety, which is present in all scaffolds, can be matched in two distinct ways for any pairwise examination. When comparing scaffolds 1 and 3, the fact that one end of the urea group is cyclized into a 6:5 ring system

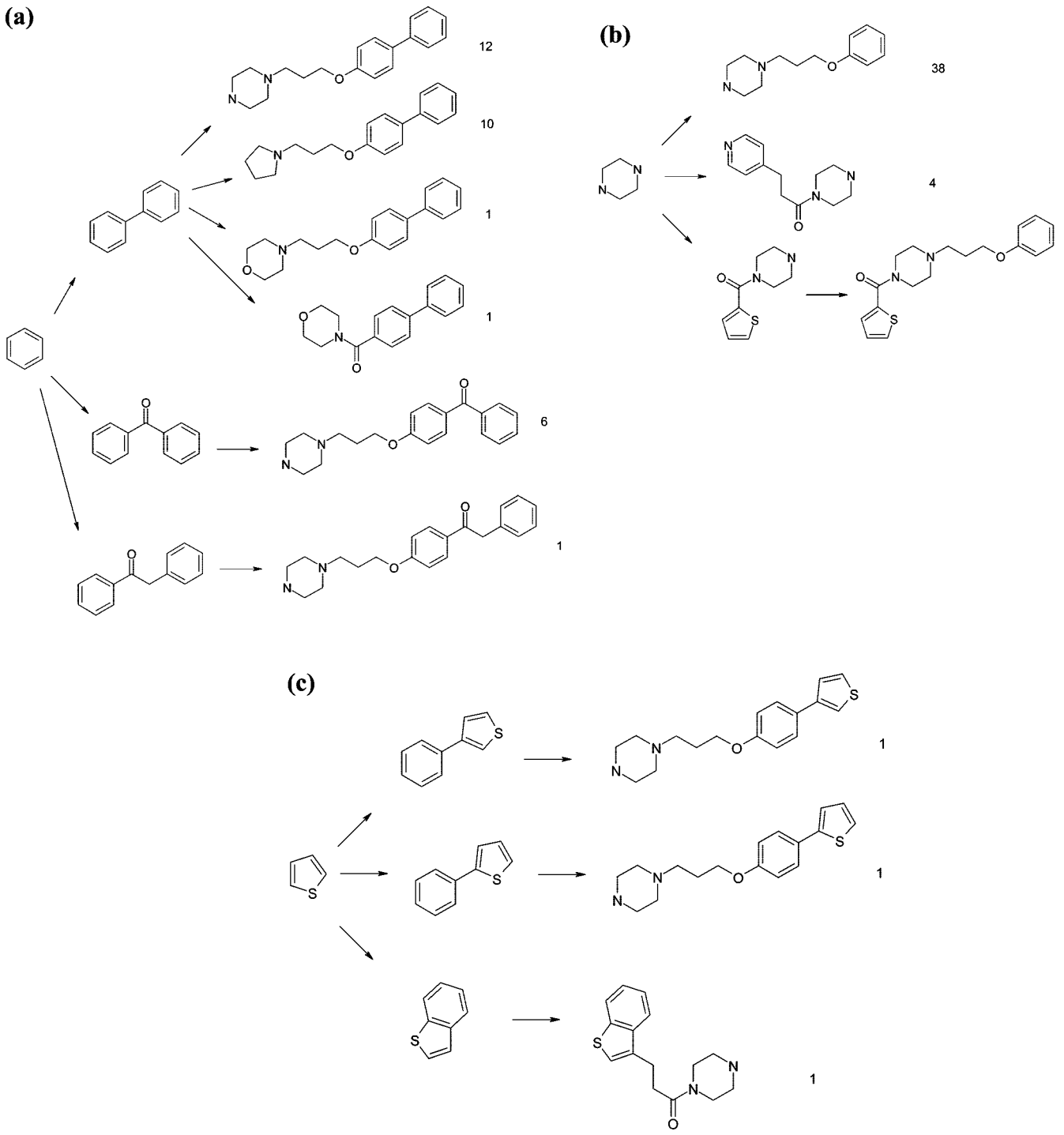


Figure 6. For a collection of molecules studied for possible HRH1 inhibition: (a) Subset of fragments which reduced to benzene (67 molecules, 31 shown). (b) Subset of fragments which reduced to piperazine (65 molecules, 43 shown). (c) Subset of fragments which reduced to thiophene.

should be sufficiently much of a clue, unless the priority of the mapping is chosen to be a different ring, such as the matching *m*-pyridyl substituents, in which case a quite different alignment would result. While an algorithm which maximizes graph compatibility would quite likely obtain the best alignment between scaffolds 1 and 3, when scaffold 2 is compared against either scaffold 1 or 3, it is not at all clear from the graph how these alignments should proceed. Given that matching the urea group contributes significantly to the score, the two ways in which this match can be achieved have comparable merits.

Figure 9 shows six representative examples of the source molecules. As it happens, this particular data set has a low variety of possible substituents at certain positions and the same substituents are repeatedly encountered across molecules that have different scaffolds. In the selection shown, it can be seen that the substituents denoted by R_4 in each of the scaffolds (which is degenerate with R_2 in the case of scaffold 2) feature $-CF_3$ at least once for all of the scaffolds, while $-Cl$ is also commonly observed. At position R_3, the $-OMe$ substituent is seen at least once in all cases, while a phenyl substitution at R_7 is also common. The presence of these common substituents influences the scoring function, which encourages the rings on either side of the lactam functional group to be aligned in the optimal way.

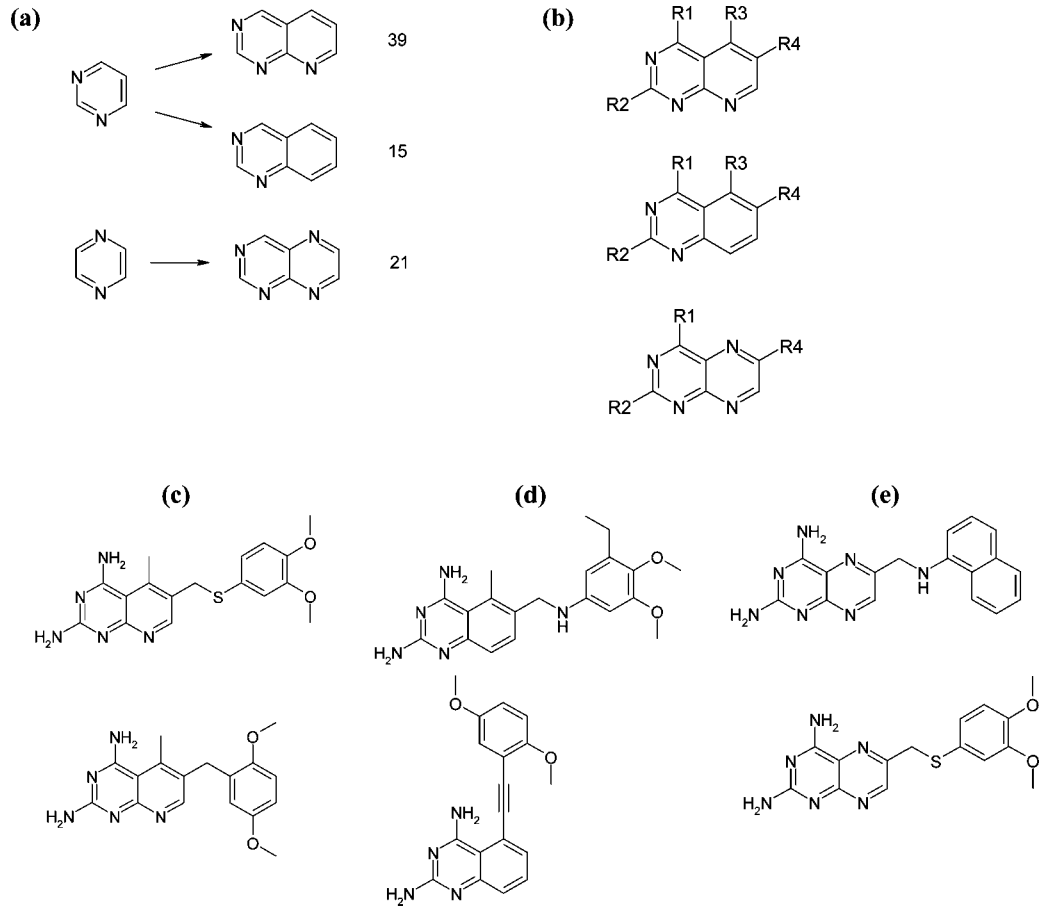


Figure 7. For a collection of molecules studied for possible DHFR inhibition. (a) The first two levels of the fragmentation tree. (b) The three resulting scaffolds, where the alignment is indicated by R-group labels. (c,d,e) Selected input molecules based on the three resulting scaffolds.

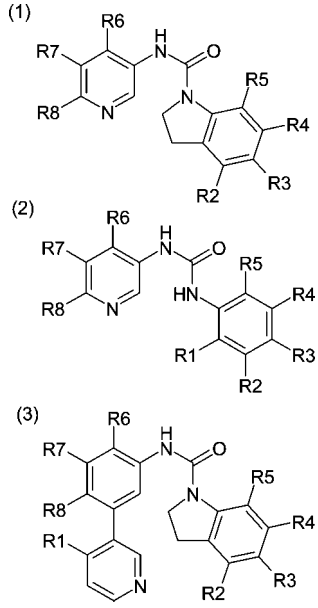


Figure 8. Scaffolds determined for molecules measured for activity against 5-HT_{2C}.

Note also that all of the nonfused rings in these scaffolds contain graph symmetry, which further increases the number of permutations that must be considered during the mapping process. It also ensures that the following scaffold assignment step requires a nontrivial analysis of the possible substructure matches.

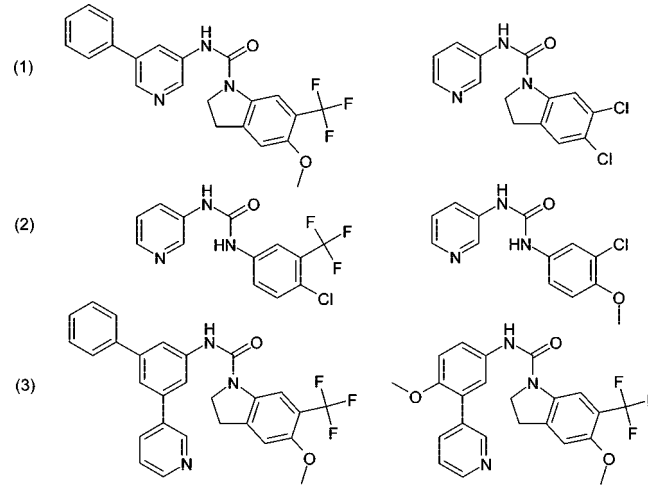
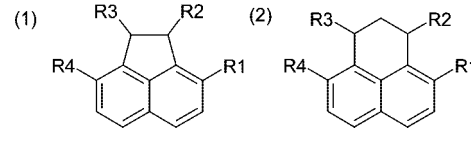


Figure 9. Representative examples of molecules measured for activity against 5-HT_{2C}.

Case 5: Symmetrical Common Scaffolds. A collection of 45 molecules with activity against melatonin receptor 1 (MT1) was assembled, all of which were based on the two scaffolds, shown below:



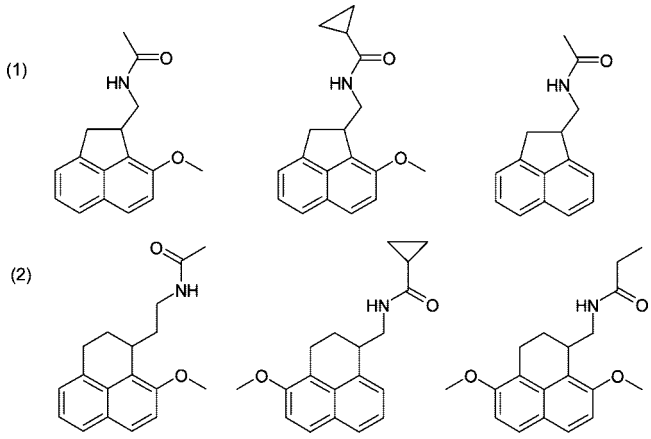


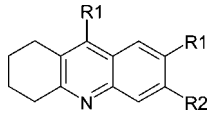
Figure 10. Representative molecules from the MT1 set.

Both scaffolds have the four possible substitution points marked as R_1 through R_4. Because both of these scaffolds have 2-fold symmetry, there are no self-energy terms in the initial assignment iteration, and the number of ways to assign the scaffolds to the input molecules is 2^{45}. In Figure 10, the first row shows three representative examples of the assignments made, drawn such that the layout orientation matches that of the scaffolds.

Note that all of the solutions chosen are such that R_2 has an amide substituent, which is true for all of the corresponding molecules. In no cases was the symmetric state chosen such that the amide substituent was placed at the location of R_3. Reinforcing this decision is the fact that all of the molecules with this scaffold have either no substituent on the aromatic part of the ring block or have a methoxy functional group in the β position closest to the amide, i.e., in the position of R_1, as assigned. The optimal solution could just have easily been such that R_3 and R_4 were substituted instead, due to symmetry. Mixing and matching the degenerate options, however, would result in a nonoptimal solution to the unary quadratic equation.

In the final analysis of the substituent patterns, the R_3 label is dropped because the optimal solution contains sufficiently low substituent diversity that this site is never occupied by anything other than hydrogen. A nonoptimal solution would have produced assignments which necessitated retention of all four possible substitution points.

Case 6: Overly Common Scaffolds. For a collection of 21 molecules with activity against acetylcholine esterase (AChE), all of them were found to have the scaffold shown below:



While the scaffold does not have permutational symmetry, and occurs in all of the scaffolds, it can be seen from Figure 11 that the scaffold is present in two fragments of some of the input molecules. Normally this would trigger the complex phase of the scaffold assignment algorithm, however, in this case, it happens that all of these nondegenerate assignments produce exactly equivalent implied substituents, so the molecules with degenerate potential assignments are trimmed down to a single unique assignment, and the procedure becomes trivial.

Case 7: Prespecified Scaffolds. The a priori scaffold discovery and analysis techniques described in this work are

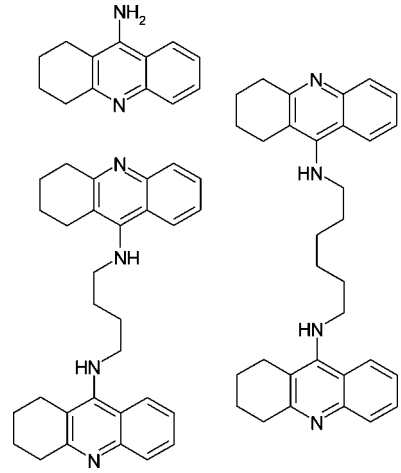
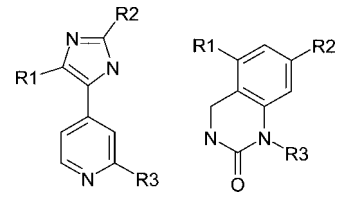


Figure 11. Several molecules from the AChE set.

based on finding optimal ways to maximize alignment of scaffolds by structural similarity and uniform distribution of substituents. Sometimes, however, information about the scaffolds is available from other sources, such as from the synthetic procedures used to produce the lead molecules, or alignments obtained from crystallographic data. The methods described in this work can be used when partial information is available, in which case the analysis proceeds to fill in any missing information.

Figure 12 shows two ligands, inhibitors of p38 MAP kinase, aligned such that the protein residues are superimposed, showing the relative binding modes of the ligands, which is not necessarily the alignment that would be guessed by comparing the chemical graphs. The pertinent scaffolds, manually derived from the structural data, are:



These user-supplied scaffolds are labeled with R-groups at the positions where auxiliary substituents were trimmed off. The label numbers have been chosen such that they have common meaning for both scaffolds.

When applied to a data set of 183 compounds measured for activity against p38 MAP kinase, the scaffold detection process identifies three additional substructures that cover the remainder of the data set. Figure 13 shows the complete set of scaffolds. The detected scaffolds (3, 4, and 5) are relatively easy to align to the presupplied scaffolds (1 and 2) because they are graph isomorphic, and the element substitution patterns clearly favor a preferred lock (e.g., 1 and 3 have multiple permutations, but the common pyridyl nitrogen exerts a significant bias). The scaffolds are shown with additional substitution points, R_4 and R_5, because the input data has additional substitution patterns that were not observed in the two ligands that were used to construct the original scaffold submissions.

This example illustrates how judicious use of a small amount of experimentally derived knowledge can be used to impart structural meaning on a database of ligands which was not otherwise readily derivable from the ligand structures alone.

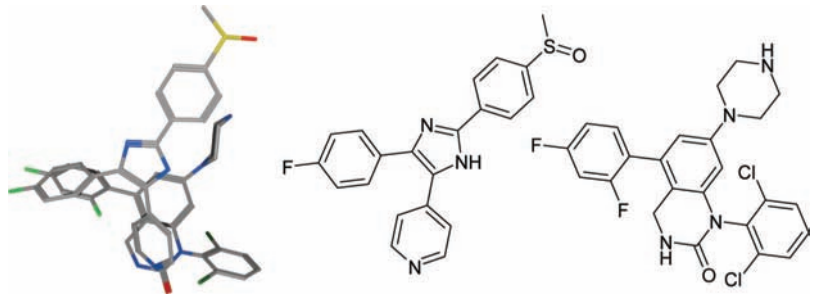


Figure 12. Structural alignment of two inhibitors of p38 MAP kinase (PDB codes 1A9U and 1M7Q).

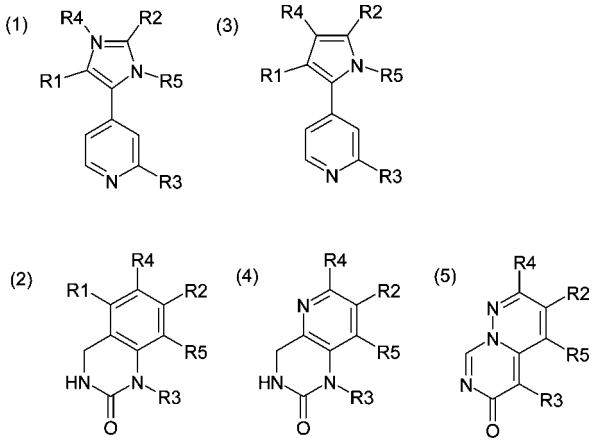
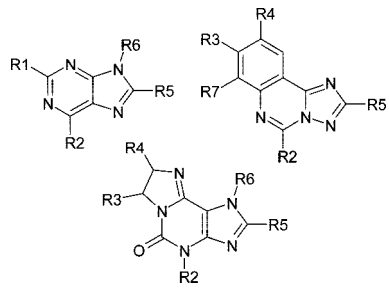


Figure 13. All scaffolds, including those determined automatically.

Case 8: Consistency. A collection of 38 structures tested for activity against the adenosine receptor 3 (ADORA3) as assembled, each of which was based on one of the following scaffolds:



From the original small set, all possible substituents at all connection points were obtained, of which there were 29 instances, some of them chains and others containing ring blocks, listed in the Supporting Information. A combinatorial library of 1000 molecules was generated by randomly combining substituents from any scaffold at any of the positions annotated with R-group labels.

The scaffold detection, alignment, and assignment algorithm was applied to the set of 1000 molecules, and the three scaffolds shown above were detected and aligned as shown above. The data set was then split randomly into two fractions of 500 molecules each and analyzed separately. In each case, the same scaffolds were detected, with the same topological alignment as shown above.

For cooperative data sets such as this one, where all molecules match one of a few scaffolds and there is a moderately high redundancy of R-groups at particular attachment positions, the algorithms that we describe are largely insensitive to size and permutation of input data because commonality provides the hints that are often vital to achieving the final scaffold assignment. For data sets which have a large proportion of molecules that are not based on one of a small number of analogous scaffolds, however, outlier scaffolds are often detected, the composition of which can vary as more data is added. Because outlier scaffolds, which are not part of any analogous series, can also interfere with the topological alignment, the number of molecules that do not belong in the series should be minimized. The results of a first-pass assignment of scaffolds is often an effective way to partition the data, and an iterative process of removing outliers, or selection of preferred scaffolds prior to resubmission, is an effective way of categorizing molecular structures about which little was previously known.

Conclusion

We have presented a method for detection, alignment, and assignment of scaffolds for databases of molecules which are related by one or more common scaffold substructures. The issues of multiple scaffolds, noncommon scaffolds, and symmetry are all addressed. The algorithms involved are fast and have been tuned such that steps which have rate-limiting scalability properties can be effectively reduced in size with minimal loss of efficacy. The method has practical utility for a priori application to databases where scaffold information is not already available, as well as instances where traditional Markush-style sketches of known scaffolds are supplied, in which case, the methods described in this work are useful for completing partial specifications and resolving symmetry and degeneracy issues.

Further work on this subject includes developing new ways to present structure−activity information once the scaffolds are determined, treatment of databases which have multiple *un*common scaffolds, which are not related and effectively partition the database, and methods for automatically examining other possible scaffold alignment information, such as may be available when some of the ligands in the set have accompanying structural data for known binding modes. An end-user application for presenting the results of this scaffold analysis in the form of an interactive report is presently under active development.

Supporting Information Available: The source data for each of the case studies is available as separate SMILES strings. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Agrafiotis, D. K.; Shemanarev, M.; Connolly, P. J.; Farnum, M.; Lobanov, V. S. SAR Maps: A New SAR Visualization Technique for Medicinal Chemists. *J. Med. Chem.* **2007**, *50*, 5926–5937.

(2) Kubinyi, H. Free Wilson Analysis. Theory, Applications and its Relationship to Hansch Analysis. *Quantum Struct.−Act. Relat.* **1988**, *7*, 121–133.

(3) Guan, H.; Laird, A. D.; Blake, R. A.; Tang, C.; Liang, C. Design and synthesis of aminopropyl tetrahydroindole-based indolin-2-ones as selective and potent inhibitors of Src and Yes tyrosine kinase. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 187–190.

(4) (a) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893. (b) Xu, J. A New Approach to Finding Natural Chemical Classes. *J. Med. Chem.* **2002**, *45*, 5311–5320. (c) Wilkens, S. J.; Janes, J.; Su, A. I. HierS: Hierarchical Scaffold Clustering Using Topological Chemical Graphs. *J. Med. Chem.* **2005**, *48*, 3182–3195. (d) Krier, M.; Bret, G.; Rognan, D. Assessing the Scaffold Diversity of Screening Libraries. *J. Chem. Inf. Model.* **2006**, *46*, 512–524. (e) Ertl, P.; Jelfs, S.; Mühlbacher, J.; Schuffenhauer, A.; Selzer, P. Quest for the Rings. In Silico Exploration of Ring Universe To Identify Novel Bioactive Heteroaromatic Scaffolds. *J. Med. Chem.* **2006**, *49*, 4568–4573. (f) Shelat, A. A.; Guy, R. K. Scaffold composition and biological relevance of screening libraries. *Nat. Chem. Biol.* **2007**, *3*, 442–446.

(5) (a) Raymond, J.; Gardiner, E.; Willett, P. RASCAL: Calculation of Graph Similarity using Maximum Common Edge Subgraphs. *Comput. J.* **2002**, *45*, 631–644. (b) Garc, G. C.; Ruiz, I. L.; Gómez-Nieto, M.Á. Step-by-Step Calculation of All Maximum Common Substructures through a Constraint Satisfaction Based Algorithm. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 30–41. (c) 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. (d) Barker, E. J.; Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Kitts, P.; Willett, P.; Gillet, V. J. Scaffold Hopping Using Clique Detection Applied to Reduced Graphs. *J. Chem. Inf. Model.* **2006**, *46*, 503–511.

(6) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree—Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, *47*, 47–58.

(7) Weininger, D. SMILES 1. Introduction and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(8) McGregor, M. J.; Pallai, P. V. Clustering of Large Databases of Compounds: Using the MDL "Keys" as Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443–448.

(9) In this work, values of *S* are between 0 and 1, these values representing the extremes of no common features in their substitution patterns to identical substitution patterns. Intermediate values can be obtained by similarity between the R-groups at each position, e.g., by the average of the Tanimoto coefficients, using structural fingerprints such as MACCS.

(10) Labute, P. Protonate3D: Assignment of Ionization States and Hydrogen Coordinates to Macromolecular Structures. *Proteins: Struct., Funct., Bioinf.* **2008**, in press.

(11) (a) Desmet, J.; de Maeyer, M.; Hazes, B.; Lasters, I. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **1992**, *356*, 539–542. (b) Pierce, N. A.; Spriet, J. A.; Desmet, J.; Mayo, S. L. Conformational splitting: A more powerful criterion for dead-end elimination. *Science* **2008**, *21*, 999–1009.

(12) (a) Clark, A. M.; Labute, P. 2D Depiction of Protein−Ligand Complexes. *J. Chem. Inf. Model.* **2007**, *47*, 1933–1944. (b) Clark, A. M.; Labute, P.; Santavy, M. 2D Structure Depiction. *J. Chem. Inf. Model* **2006**, *46*, 1107–1123.

(13) The source code for the algorithms described within, and those referred to, is packaged as part of MOE and may be examined and used under the terms of the MOE user license, which is available from the Chemical Computing Group, Inc., 1010 Sherbrooke Street West, Suite 910, Montréal, Québec, Canada; http://www.chemcomp.com (accessed Aug 18, 2008).

JM801098A