An Automated Method for Exploring Targeted Substructural Diversity within Sets of Chemical Structures

John W. Raymond* and Christopher E. Kibbey

Scientific Computing Group, Pfizer Global Research and Development, Ann Arbor Laboratories,
2800 Plymouth Road, Ann Arbor, Michigan 48105

Received June 1, 2005

Practicing medicinal chemists tend to treat a lead compound as an assemblage of its substructural parts. By iteratively confining their synthetic efforts in a localized fashion, they are able to systematically investigate how minor changes in certain portions of the molecule effect the properties of interest in the logical expectation that the observed beneficial changes will be cumulative. One disadvantage to this approach arises when large amounts of structure data begin to accumulate which is often the case in recent times due to such developments as high-throughput screening, virtual screening, and combinatorial chemistry. How then does one interactively mine this diverse data consistent with the desired substructural template, so those desirable structural features can be discovered and interpreted, especially when they may not occur in the most active compounds due to structural deficiencies in other portions of the molecule? In this paper, we present an algorithm to automate this process that has historically been performed in an ad-hoc and manual fashion. Using the proposed method, significantly larger numbers of compounds can be analyzed in this fashion, potentially discovering useful structural feature combinations that would not have otherwise been detected due to the sheer scale of modern structural and biological data collections.

1. INTRODUCTION

Substructure analysis in drug discovery is well established, and many distantly related computational methods predicated on the relationship between biological/physical properties and chemical structure have been developed.[1,2] These run the gamut from those intended for property prediction to those aimed at mining and organizing large amounts of data. The method proposed here focuses on the latter and provides a mechanism for medicinal and computational chemists to prioritize or browse chemical structure data in a manner consistent with a discovery project's current structural constraints and is essentially an abstraction of traditional R-group analysis.

The approach is conceptually simple. It involves performing a mapping of a target molecule to a template molecule that has been partitioned into substructures of interest and then segmenting the target molecule such that the resulting substructures are maximally similar to the template substructures. Using a literature example to illustrate the problem, Figure 1 depicts structure H-89, a protein kinase B (PKB) inhibitor discovered by Reuveni et al.[3] The researchers identified within H-89 a set of substructures which they dubbed "diversity domains". In this case, the diversity domains identified by Reuveni et al. were based on an inherent partitioning constrained by combinatorial synthesis and available reagents.

Based on their diversity domain template, Reuvini et al. then constructed combinatorial libraries introducing substructural diversity into each of the three specified domains. In this case, inspecting the level and content of structural diversity in the respective diversity domains in the enumerated products is relatively straightforward as it simply requires browsing cross-linked lists of reagents. Oftentimes, however, this is not a practicable solution as a specified diversity domain may comprise only a portion of a reagent or combination of reagents, or the synthetic pathway for a given molecule may not be known.

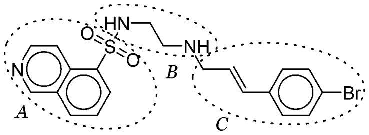


Figure 1. PKB inhibitor H-89 with the diversity domains identified.

The proposed substructure mapping approach operates by first having the user specify a diversity domain template as illustrated in Figure 1. The diversity domain substructures can be of any size, and there is no restriction on the number of diversity domains that can be adjacent to each other nor is there any restriction on the number of bonds connecting them (i.e., rings can be cleaved). The specified diversity domain template is then compared to a target set of chemical structures. These may be the result of a combinatorial library design enumeration, Markush structure enumeration, substructure search, virtual screen, or a collection of legacy project compounds that have been synthesized over time. Each compound in the target set is superimposed onto the template using an automated substructure mapping procedure, and then an analysis is performed to determine the most appropriate bonds to cut in the target molecule so as to maximize the degree of compatibility between the resulting disconnected substructures and their corresponding target diversity domains. Figure 2 illustrates an example substructure mapping of a target molecule to the diversity domain

* Corresponding author e-mail: John.Raymond@pfizer.com.

10.1021/ci0502247 CCC: $30.25 © 2005 American Chemical Society
Published on Web 08/16/2005

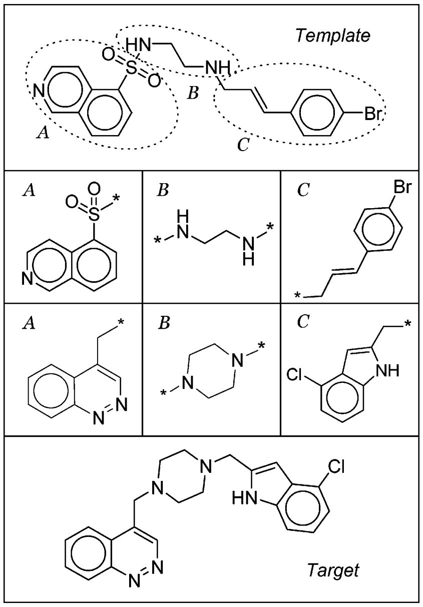


Figure 2. Example substructure mapping of a PKB inhibitor template to a target molecule.

template of Figure 1 and subsequent fragmentation. It is not difficult to envision that this process can become problematic depending upon the level of diversity within the target structures.

In practice, the natural partitioning of a molecular template at a given point in time for a project may be based on an observation of the substructural moieties related to biological activity, specificity, toxicology, physical properties, and intellectual property concerns in addition to synthetic issues. Moreover, the number and composition of the diversity domains are also subject to change throughout the course of an active project. This type of structural reductionism permeates much of the chemical literature. Table 1 depicts 10 example partitionings compiled from the medicinal chemistry literature. These examples illustrate some of the contextual diversity with respect to the structural partitioning of lead compound matter.

A review of the literature failed to discover any published methods that directly address this problem. Much of the published articles on focused substructure analyses attempt to establish various forms of structural commonality within groups of related molecules to discern a correlation between previously unknown structural features and a target property such as biological activity. Some of these are peripherally related to the current problem but remain quite distinct in application.

2. METHOD

Once a reference template is formulated which identifies the substructures to be considered during the segmentation process, the substructure mapping process can be initiated. The proposed mapping procedure consists of two primary stages. The first stage consists of a calculation to determine the structural commonalities between the template and the target structure, and the second stage establishes which bonds to cut in the target molecule to create an optimal substructure partitioning of the target molecule based on the detected commonalities. A pseudocode outline of the entire algorithm is presented in Chart 1. To facilitate browsing activities, the resulting substructures can then be output in lists indicating the substructure's parent ID and associated property values as well as the computed similarity to their corresponding diversity domains in the template structure.

Reference Frame (Stage 1). The first step in establishing a reference frame for the partitioning procedure is to indicate the desired diversity domains within the template molecule.

Table 1. Example Substructure Partitions from the Literature

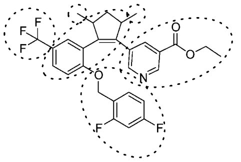
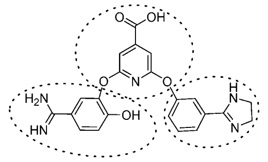
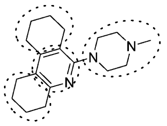
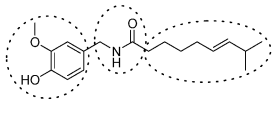
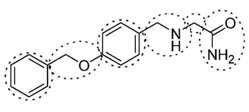
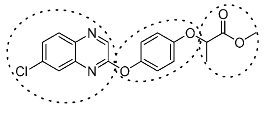
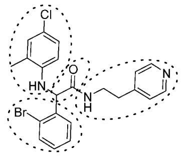
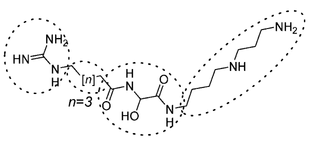
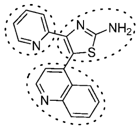
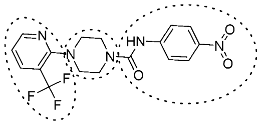
Structure	*Note*	*Structure*	*Note*
	Patent Markush structure enumeration [4]		SAR/binding site model [5]
	SAR/selectivity [6]		SAR/convenient partitioning [7-9]
	SAR/potency [10]		SAR/pharmacophore hypothesis [11]
	SAR/synthetic strategy [12]		SAR/design out chirality and hydrolytic sensitivity [13, 14]
	SAR/binding site model [15]		SAR/combi-synthesis [16]

Chart 1

```
COMMENTS
Set notation follows that of Brassard and Bratley.[30]
($\tilde{a}_{ij} \in \mathbf{M}$) is true if the clique solution M contains a phantom node or
false otherwise.

Program Molecular Partition
 Assign the template diversity domains and corresponding weights
 Identify cores in $G_2$ if present in $G_1$ by subgraph isomorphism
 Compute constrained MCES $G_{12}$ between $G_1$ and $G_2$
 $\mathbf{B}_0 \leftarrow$ initial set of allowable bonds to cut
 Call Bond Cut routine
 Output optimum partition
End

Subroutine Bond Cut
Begin
 $K \leftarrow$ number of diversity domains in $G_1$
 $W^{best} \leftarrow 0$      /* best vertex assignment score */
 $R^{best} \leftarrow \infty$      /* best subgraph radius score */
 $S^{best} \leftarrow \infty$      /* best subgraph size score */
 $C \leftarrow 0$         /* current number of cuts */
 $\mathbf{b} \leftarrow \varnothing$         /* set of selected bonds */
 $\mathbf{b}^{best} \leftarrow \varnothing$      /* best cut-set */
 While $(\exists x)[x \in \mathbf{B}_C]$
   $C \leftarrow C + 1$
   $b_C \leftarrow x \mid x \in \mathbf{B}_{C-1}$
   $\mathbf{b} \leftarrow \mathbf{b} \cup b_C$
   $\mathbf{B}_C \leftarrow \mathbf{B}_{C-1} \backslash b_C$
   $K_s \leftarrow$ the number of components in $G_2(\mathbf{V},\mathbf{E}\backslash\mathbf{b})$
   If $K_s > K$ Then BackTrack
   Else
     $W,R,S \leftarrow$ Upper Bound Scores for $G_2(\mathbf{V},\mathbf{E}\backslash\mathbf{b})$
     If Upper Bound Check = true then
       If $(\tilde{a}_{ij} \in \mathbf{M})$=false then
         /* Record best score */
         $W^{best} \leftarrow W$; $R^{best} \leftarrow R$; $S^{best} \leftarrow S$
         $\mathbf{b}^{best} \leftarrow \mathbf{b}$
         BackTrack
       End if
     Else BackTrack
     End if
   End if
 End while
 Return $\mathbf{b}^{best}$
End

Subroutine BackTrack
Begin
 $\mathbf{b} \leftarrow \mathbf{b} \backslash b_C$
 $\mathbf{B}_{C-1} \leftarrow \mathbf{B}_{C-1} \backslash b_C$
 $C \leftarrow C - 1$
End

Subroutine Upper Bound Scores
Begin
 Compute W and subgraph matching using maximum weight clique
 If $W = W^{best}$ and $(\tilde{a}_{ij} \in \mathbf{M})$=false then
   Compute R and S
 End if
 Return W,R,S
End

Subroutine Upper Bound Check
Begin
 If $W > W^{best}$ then return true
 Else if $W = W^{best}$ then
   If $(\tilde{a}_{ij} \in \mathbf{M})$=true then return true
   Else if $R < R^{best}$ then return true
   Else if $R = R^{best}$ then
     If $S < S^{best}$ then return true
   End if
 End if
 Return false
End
```

This consists of a simple labeling of the atoms of the template molecule such that each atom in the template molecule is assigned a label under the constraint that each atom contained in a diversity domain has the same label but no two atoms in different diversity domains have the same label. The diversity domains corresponding to the published partitioning of the antihistamine "superstructure" of Cammarata and Menon[17] are depicted in Figure 3(*i*).

Following this, the structural commonality existing between the template and target structures must first be deduced to serve as a reference frame for fragment correspondence as well as a constraint on the bond cutting procedure. Once a reference frame is established, each atom in the target molecule is assigned the same label as its corresponding atom in the template molecule in accordance with a mapping of the largest common substructure(s) between the template and target structures. If an atom in the target molecule is not included in the largest common substructure, it is assigned a null label as illustrated in Figure 3(*ii*). This process is accomplished by using an algorithm to determine the largest substructure(s) common to both the template and target structures with size being defined as the number of bonded atom pairs.

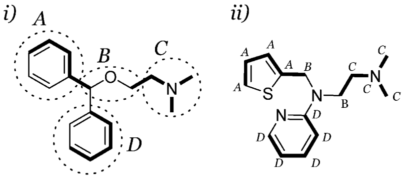


Figure 3. Example maximum common edge subgraph (MCES) between two chemical structures highlighted in bold.

In a graph theoretic context, the atoms of a 2D representation of a molecule are referred to as vertices, and the bonds between atoms are referred to as edges. This form of representation is referred to as a chemical graph.[18] The common substructure consisting of the most bonded atom pairs is more formally referred to as the maximum common edge subgraph (MCES). The MCES need not be connected. The MCES problem belongs to the class of graph theoretic problems called isomorphism algorithms. A survey of the MCES problem and associated algorithms is provided in the literature.[19] Figure 3 illustrates an example MCES calculation between two antihistamines where the MCES is highlighted in bold in both structures.

Typing Constraints. One of the drawbacks of using an MCES to establish the commonality between two chemical graphs is the inherent disparity between a chemical and graph theoretical description of commonality. An MCES description of chemical similarity can occasionally suffer from two principal limitations. The first shortcoming is due to atom and bond type compatibility constraints. Using a strict adherence to atom and bond typing in the MCES calculation procedure can sometimes produce results that are not optimal with respect to a chemical interpretation of similarity. For instance in a particularly diverse set of target structures, a cyclohexane ring may be an acceptable replacement for a benzene ring; however, a strict adherence to bond typing in the MCES calculation would detect essentially no commonality between the two rings. To address this issue, we have accommodated varying degrees of bond type relaxation. This includes options for (1) ignoring all bond typing, (2) ignoring bond typing when both bond pairs occur in rings, and (3) increasing bond type specificity by distinguishing between pairs of bonds of the same type when one is in a ring and the other is not. Bond type relaxation option 2 has been found to work well as the default alternative.

Connectivity Constraints. Additionally, the occasional incongruity between the concept of local and global similarity also has the potential to become problematic during an automated bond cutting procedure. This is best described visually. Figure 4 depicts two structures and their corresponding diversity domains. While there is a high degree of local similarity between the corresponding diversity domains,

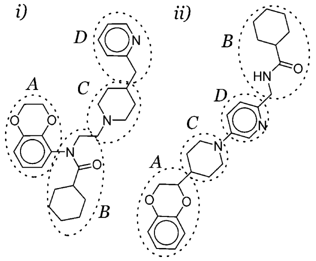


Figure 4. Example of structures that are locally similar but globally dissimilar with respect to the specified diversity domains.

the overall similarity between the two molecules is small since the position of the diversity domains in each molecule is markedly different. For instance, diversity domains *A* and *B* in structure (*i*) of Figure 4 are one bond length apart, whereas they are nine bond lengths apart in structure (*ii*). In this case, a mapping of the diversity domains from structure (*ii*) to structure (*i*) based solely on diversity domain similarity would have little meaning from a chemical perspective as it does not maintain the integrity of the connectivity between diversity domains in their respective parent structures.

Enforcing a global connectivity constraint so that the relative orientation of diversity domains is preserved between a template and target molecule is accomplished using shortest path criteria. A shortest path between two atoms is a chain of bonds connecting the two atoms consisting of the fewest number of bonds and can be easily computed using established algorithms.[20] Prior to calculating the MCES, the shortest paths between all pairs of atoms in each molecular structure is calculated using an all shortest paths algorithm. Then an MCES is computed under the constraint that all pairs of atoms present in the MCES must correspond to pairs of atoms in each structure being compared such that the difference in shortest path lengths between the pairs of atoms are within a threshold number of bond lengths. In Figure 4, the nitrogen atom in the pyridine ring of diversity domain *D* and the carbonyl oxygen of diversity domain *B* in structure (*i*) are eleven bond lengths apart, whereas they are five bond lengths apart in structure (*ii*). Therefore, if the shortest path threshold is three or less, then the two nitrogen atoms would not simultaneously occur in any calculated MCES. Our experiments have found that a default shortest path threshold of *three* works well in most simulations.

The MCES-based labeling procedure described here utilizes the published RASCAL algorithm;[21,22] however, any sufficiently efficient MCES algorithm that allows for the shortest path constraints can be substituted. Since the MCES mapping provides the atom-to-atom and bond-to-bond correspondences between the template and target structures, transferring the template labels to the target structure becomes a trivial step (see Figure 3). This MCES-based approach for establishing a reference for the bond cutting procedure has been shown to be very robust even when the structural commonality between the template and target molecules is limited, especially when compared to the degree of commonality necessary to be effective when used as a virtual screening similarity measure.[23] The reason for this is that the structural commonalities are used solely in an intramolecular context to establish the relative positioning of the prospective diversity domains in the target molecule. The absolute level of commonality between the two structures is not relevant as described in the next section.

Substructure Core Constraints. Much of the novelty of the proposed fragmentation method lies in its ability to map sets of diverse substructures while still maintaining global connectivity. In this sense it is an abstraction of conventional R-group analysis, whereby various R-group functionalities are clipped based upon a mapping of a specified core structure onto a target molecule; however, it dispenses with the requirement that the target molecule must possess a core substructure.

In some instances, the ability to specify that certain portions of the molecule are to remain constant and to only consider those target structures containing the specified fixed substructures may be desirable. These situations can be readily accommodated within the proposed algorithmic framework. To accomplish this, the user need only indicate which substructure diversity domains in the template structure are to remain constant, and then a simple substructure search routine is executed prior to the MCES calculation. A subgraph isomorphism algorithm such as the Ullmann method[24,25] is used to enumerate all instances of the specified core substructures in each target structure. If the subgraph isomorphism detection was successful, the resulting substructure mappings are then used to constrain which atoms in the target structure can be mapped to the core template substructure atoms during the MCES calculation. This not only enforces that the core substructures be present in the resulting MCES but can also improve the efficiency of the MCES calculation.

Bond Cutting (Stage 2). The bond cutting procedure determines which bonds in the target molecule should be cleaved so that the resulting substructures are maximally similar to their corresponding diversity domain segments in the template molecule. In a mathematical context, it is related to the *k*-cut problem from graph theory. Given a graph $G = (V,E)$ with nonnegative edge weights, the *k*-cut problem is a partitioning of the vertices of G into k nonempty components by deleting a subset of edges $E^{\mathrm{cut}} \in E$ such that the total weight of E^{cut} is a minimum.[26] If no edge weights are specified, then each edge is assigned a value of unity. While a review of the *k*-cut literature retrieved several algorithmic variants,[26−29] the authors were unable to find a published methodology that directly addressed the current problem.

The problem posed here differs from the classic *k*-cut problem because it is necessary to impose constraints on the cutting procedure with respect to the diversity domains specified in the template. The aforementioned MCES procedure provides the mechanism for establishing the association between the target atoms and the template diversity domain atoms. The bond cutting problem posed here is to partition a target chemical graph into K_s components, where $K_s \leq K$, using an optimal number of bond cuts such that $f(\boldsymbol{S}^1,\boldsymbol{S}^2)$ is maximized. $f(\boldsymbol{S}^1,\boldsymbol{S}^2)$ is an objective function evaluating the level of compatibility between the set of calculated diversity domains ($\boldsymbol{S}^2$) in the target molecule and the reference diversity domains in the template molecule ($\boldsymbol{S}^1$). K is the number of diversity domains specified in the template molecular graph G_1. During the bond cutting procedure, the objective function $f(\boldsymbol{S}^1,\boldsymbol{S}^2)$ also serves to upper-bound the enumeration process increasing the efficiency of the algorithm. The objective function consists of the calculation of three graph-based descriptors, W, R, and S, used in combination to evaluate the relative quality of a partial or

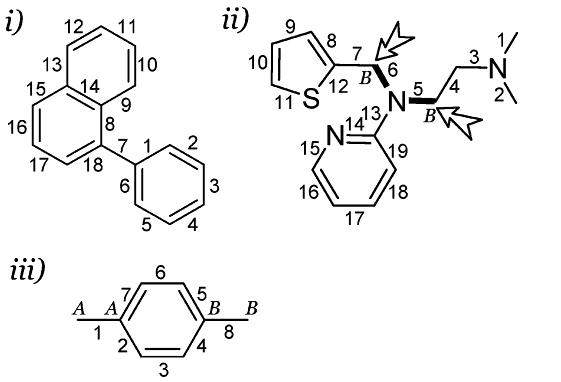


Figure 5. Illustration of the steps in the permissible bond-cut procedure: (*i*) aromatic ring heuristic, (*ii*) "buried" bond heuristic, and (*iii*) correction heuristic.

complete solution. This procedure is discussed in detail in the Upper Bounding section.

There may arise situations where it is important to stress the structural consistency of some diversity domains with respect to the others. A simple example demonstrating this situation is a molecule that is divided into three fragments, *A*, *B*, and *C*, where *B* is a variable linker between two pharmacologically important binding regions. In this case, it is desirable to have the potential variations in the size of the linker in the target molecules to be confined to the set of *B* fragments rather than be equally distributed among the *A*, *B*, and *C* fragments. This issue is addressed by allowing the user to assign weights (ω_i) to each template diversity domain (S_i^1), so that those fragments which are weighted the lowest will be forced to accept the most diversity with respect to fragment size incongruities. The default procedure is to treat the significance of the template diversity domains, $\boldsymbol{S}^1$, uniformly, and this works surprisingly well—even in cases with considerable variability in corresponding fragment sizes.

Permissible Cuts. The proposed algorithm is a simple depth first search[30] (DFS) that exhaustively enumerates all permissible bond cut possibilities. The efficiency of the algorithm is attributable to constraints placed on the initial set of permissible bond cuts and on the calculation of the upper bound during the DFS enumeration process. It is obvious from the formulation of the problem that when fewer bonds are considered for possible cleaving efficiency will increase since the number of possible bond cut combinations increases exponentially with the number of allowable bond cuts; however, the list of permissible bond cuts must not be so stringent so as to negatively effect the quality of the calculated partition.

The sequential strategy employed here for specifying the set of bonds that are not allowed to be cut in the enumerative bond cutting procedure, $\boldsymbol{B}^{\text{fix}}$, is as follows:

(1) All bonds are initially placed in the set of bonds that are permitted to be cut, $\boldsymbol{B}$.

(2) Any fused ring bond or unfused ring bond, b_{ij}, with endpoint atoms *i* and *j* within an aromatic ring where b_{ij} is not incident on an adjacent fused ring bond is removed from $\boldsymbol{B}$ to $\boldsymbol{B}^{\text{fix}}$. In Figure 5(*i*), bonds {7,8,9,13,15} belong to set $\boldsymbol{B}$, and bonds {1,2,3,4,5,6,10,11,12,14,16,17,18} belong to set $\boldsymbol{B}^{\text{fix}}$ at this stage in the decision criteria.

(3) "Buried" bonds are removed from set $\boldsymbol{B}$ to set $\boldsymbol{B}^{\text{fix}}$. "Buried" bonds are identified by computing a shortest path between all pairs of atoms, *i* and *j*, in the target molecule with identical projected template atom labels (excluding the null label). If every atom on the computed shortest path has the same label as atoms *i* and *j* or has a null label, then each bond on the shortest path is considered to be a "buried" bond and is moved from the set $\boldsymbol{B}$ to set $\boldsymbol{B}^{\text{fix}}$; otherwise, no action is taken for the given *ij* shortest path. As an example, Figure 5(*ii*) depicts the structure in Figure 3(*ii*) along with projected template atom labels for two atoms under consideration. Each bond has also been assigned a sequential numerical label. The shortest path highlighted in bold between the pair of atoms (both labeled *B* from Figure 3(*ii*)) marked with arrows consists of two bonds incident on the middle tertiary amine (i.e., 5 and 6). Since there is only one atom in this particular shortest path (i.e., the nitrogen) and it has a null label, *Step 3* stipulates that bonds 5 and 6 be moved to set $\boldsymbol{B}^{\text{fix}}$ if they are not already present in the set. In fact in this simple example, implementing *Steps 1*−*3* for this example will result in only three bonds remaining in the set $\boldsymbol{B}$ (i.e., 4,7,13).

(4) In the final step of the permissible bond cut procedure, a check is made to determine whether the combination of *Steps 2* and *3* were overly restrictive. This special case is depicted in Figure 5(*iii*) and occurs when the bridge between two identified diversity domains is an aromatic ring. From *Step 2*, it is clear that bonds {2,3,4,5,6,7} are in the set ($\boldsymbol{B}^{\text{fix}}$) of bonds which cannot be cut, and from *Step 3*, bonds {1,8} are also in $\boldsymbol{B}^{\text{fix}}$. However, there is a clear demarcation between the projected template atom labels. To account for this, a final pass is made through all the bonds in $\boldsymbol{B}^{\text{fix}}$, and any bond, b_{ij}, in $\boldsymbol{B}^{\text{fix}}$ that is incident on an atom, *j*, in an aromatic ring where neither of the atoms bonded to atom *j* present in the ring have a template projection label identical to that of atom *j* is moved back to set $\boldsymbol{B}$. Therefore, both bonds 1 and 8 are placed back into set $\boldsymbol{B}$.

Upper Bounding. Further algorithmic increases in efficiency are provided by the upper bounding procedure. This procedure helps prevent the further exploration of bond cut combinations that cannot result in an improved solution with respect to the current best solution. The upper bound is calculated using the partition objective function $f(\boldsymbol{S}^1,\boldsymbol{S}^2)$. The objective function $f(\boldsymbol{S}^1,\boldsymbol{S}^2)$ is a relative measure of the quality of the mapping between the template diversity domain partition and the partition induced by cutting the currently selected bonds in the target molecule.

Through several configurations and test simulations, the upper bound cost function found to result in the most intuitive partitions consists of three dependent bound tests: a vertex compatibility score (*W*), a subgraph radius score (*R*), and a subgraph size score (*S*). All three bound tests are predicated on a subgraph to subgraph mapping procedure that is performed at every state-space instance in the DFS tree (i.e., at every combination of possible bond cuts). At any given depth, *d*, in the DFS tree, the currently selected set of "cut" bonds, $\boldsymbol{b}$, will separate the target graph G_2 into K_s subgraphs. The K_s subgraphs can be identified by deleting the *d* bonds in set $\boldsymbol{b}$ from G_2 and implementing a connected component algorithm.[31,32] If $1 \leq K_s \leq K$ where *K* is the number of specified diversity domains in the template graph G_1, then a subgraph to subgraph correspondence between the set of computed G_2 subgraphs, $\boldsymbol{S}^2$, to the specified set of G_1 diversity domains, $\boldsymbol{S}^1$, is performed; otherwise, a backtrack instance is initiated.

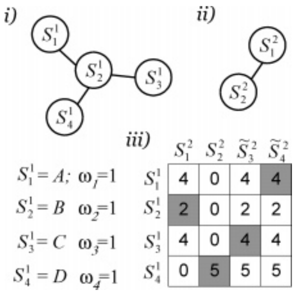

i)
S_1^1
S_2^1
S_3^1
S_4^1
ii)
S_1^2
S_2^2
iii)
S_1^2 S_2^2 $\tilde{S}_3^2$ $\tilde{S}_4^2$
$S_1^1 = A;$ $\omega_1=1$ S_1^1 4 0 4 4
$S_2^1 = B$ $\omega_2=1$ S_2^1 2 0 2 2
$S_3^1 = C$ $\omega_3=1$ S_3^1 4 0 4 4
$S_4^1 = D$ $\omega_4=1$ S_4^1 0 5 5 5

Figure 6. Example $\boldsymbol{S}^1$ to $\boldsymbol{S}^2$ subgraph correspondence calculation: (*i*) reduced template graph of Figure 3(*i*), (*ii*) reduced target graph of Figure 3(*ii*) at depth = 1, and (*iii*) optimal $\boldsymbol{S}^1$ to $\boldsymbol{S}^2$ mapping identified by shaded boxes in the profit matrix $\boldsymbol{A}$.

The subgraph to subgraph correspondence is simply an assignment of the diversity domains in $\boldsymbol{S}^1$ to the current set of computed subgraphs, $\boldsymbol{S}^2$, at a given depth in the DFS tree. A linear assignment approach[33] was initially tested as a possible solution, but it was found to suffer from two primary defects. The first is that it ignores connectivity between the subgraphs. This is ameliorated to a large extent by the shortest path threshold criteria used in the initial MCES calculation; however, occasional problems with an imperfect MCES mapping led to the adoption of a more robust $\boldsymbol{S}^1$ to $\boldsymbol{S}^2$ correspondence scheme.

A linear assignment approach also creates difficulties when the two structures are structurally diverse enough that the MCES label projection is unable to project any labels for a specified template diversity domain onto any part of the target molecule or when the shortest path constraint in the MCES calculation failed to enforce global connectivity. This occurs, for instance, in situations such as when the template structure is designated as having three diversity domains (*A*, *B*, and *C*), but the target structure only has *A* and *B* fragments or is dissimilar enough that the corresponding *C* fragment in the target structure that should be matched to the *C* fragment in the template structure displays no exact substructural commonality.

Subgraph Correspondence. A maximum weight clique algorithm has proven the most effective for establishing the $\boldsymbol{S}^1$ to $\boldsymbol{S}^2$ correspondence. A maximum weight clique is a set of mutually adjacent vertices in a graph for which the sum of weights associated with each vertex is a maximum. In our implementation, we have modified the Wood maximum clique algorithm[34] which was originally developed for the unweighted case; however, this level of customization is not necessary to implement the proposed method, and most published maximum weight clique algorithms should perform satisfactorily, especially the partition or vertex coloring based algorithms.[35−37]

As a simplification, each subgraph $S_i^1 \in \boldsymbol{S}^1$ and $S_j^2 \in \boldsymbol{S}^2$ can be condensed to form another type of graph referred to as a reduced graph[38] where each vertex in the reduced graph corresponds to a subgraph in its parent graph. Figure 6 formulates an example calculation of the $\boldsymbol{S}^1$ to $\boldsymbol{S}^2$ correspondence procedure using the template partitioning and target structure of Figure 3 (parts (*i*) and (*ii*), respectively). In this example, it is assumed that the current depth in the DFS tree is $d = 1$, and the selected cut-set, $\boldsymbol{b}$, contains bond 13 (as numbered in Figure 5(*ii*)). Figure 6(*i*),(*ii*) depicts the reduced graphs for the template and target structures, respectively. Cleaving bond 13 in the set $\boldsymbol{b}$ separates the structure in Figure 5(*ii*) into two substructures represented by the reduced graph nodes S_1^2 and S_2^2. Reduced graph node S_1^2 corresponds to the substructure containing the two amines (i.e., bonds {1,2,3,4,5,6,7,8,9,10,11,12}), and S_2^2 corresponds to the pyridine ring fragment (i.e., bonds {14,15,16,17,18,19}).

The first step in this process is to construct a $K \times K$ profit matrix, $\boldsymbol{A}$, detailing the compatibility of pairing a template subgraph S_i^1 node to a target subgraph S_j^2 node as illustrated in Figure 6(*iii*). Each element in the profit matrix is defined as $a_{i,j} = \omega_i \cdot n(S_i^1, S_j^2)$ where ω_i is the relative weighting factor assigned to diversity domain S_i^1 and $n(S_i^1, S_j^2)$ is the number of vertices in S_j^2 that have the same label as the vertices in S_i^1. For instance, S_2^2 consists of 5 vertices with a *D* label and 1 with a null label; therefore, $a_{4,2} = 5$, $a_{1,2} = 0$, $a_{2,2} = 0$, and $a_{3,2} = 0$ assuming $\omega_i = 1$ for all S_i^1. In this instance, the current cut-set $\boldsymbol{b}$ only separates G_2 into $K_s = 2$ subgraphs, S_1^2 and S_2^2; therefore, the profit matrix must be padded to facilitate the maximum weight clique upper-bound calculation. If $K_s < K$, then K-K_s phantom node columns must be created as placeholders in the profit matrix. The value assigned to each phantom node, $\tilde{a}_{i,j}$, where $K_s < j \le K$, is defined by $\tilde{a}_{i,j} = \max\ (a_{i,k}) | 1 \le k \le K_s$, as demonstrated in Figure 6(*iii*).

For clique detection purposes, each element in the profit matrix $\boldsymbol{A}$ becomes a vertex in the compatibility graph, G_c. Adjacency in G_c is determined using shortest path criteria. Let SP_{xy}^1 and SP_{xy}^2 denote the shortest path length from vertex x to vertex y in the template and target graphs, respectively. Without loss of generality, it is assumed here that for two vertices $a_{i,j}$ and $a_{m,n}$ that $j < n$. Then an edge exists in G_c between two vertices $a_{i,j}$ and $a_{m,n}$ where $i \neq m$ and $j \neq n$ if (1) $SP_{im}^1 = SP_{jn}^2 | 1 \le j \le K_s$ and $1 \le n \le K_s$, or if (2) $SP_{im}^1 \ge SP_{\min}^2 + 1 | 1 \le j \le K_s$ and $K_s < n \le K$ where $SP_{\min}^2 = \min\ (SP_{jy}^2 | 1 \le y \le K_s$ and $a_{i,j} > 0$ and $a_{m,y} > 0)$, or if (3) $SP_{im}^1 \ge SP_{\min}^2 + 2 | K_s < j \le K$ and $K_s < n \le K$ where $SP_{\min}^2 = \min\ (SP_{xy}^2 | 1 \le x \le K_s$ and $1 \le y \le K_s$ and $a_{i,x} > 0$ and $a_{m,y} > 0)$.

The clique in graph G_c corresponding to an optimal $\boldsymbol{S}^1$ to $\boldsymbol{S}^2$ mapping can then be determined using a maximum weight clique algorithm. The maximum weight clique algorithm returns a set $\boldsymbol{M}$ of vertices (i.e., $a_{i,j}$) in G_c such that the sum of weights for all $a_{i,j}$ elements in $\boldsymbol{M}$ is the maximum possible. Note that $|\boldsymbol{M}| \le K$ since it is possible that there is not a corresponding substructure in the target molecule for every diversity domain in the template molecule. Once the subgraph correspondence has been established, the upper-bound calculation is performed to determine whether the partitioning induced by the current cut-set has either the potential to or has resulted in the most optimal partitioning discovered thus far.

Upper Bound Evaluation. The three stage upper bound is evaluated at each forward DFS step, and if it returns a value of '*false*', a back-track is instanced. The upper bound calculation is derived from the three sequentially dependent scoring functions described below:

(1) Vertex Compatibility Score (W). Following the $\boldsymbol{S}^1$ to $\boldsymbol{S}^2$ mapping, the vertex compatibility score is determined by summing the vertex weights of the clique solution $\boldsymbol{M}$. Therefore, $W = \sum a_{i,j} | a_{i,j} \in$ M. If $W > W^{\text{best}}$ where W^{best} is

the best vertex compatibility score thus far detected or if $W = W^{\mathrm{best}}$ and $\boldsymbol{M}$ contains at least one phantom node (i.e., $\tilde{a}_{i,j}$), then the objective function returns a value of '*true*'; else if $W = W^{\mathrm{best}}$, then the subgraph radius score (R) is evaluated. Otherwise a value of '*false*' is returned.

(2) Subgraph Radius Score (R). This criterion anticipates the likelihood that multiple target molecule partitions may exist that maximize the value of W. Depending upon the diversity of the target molecules, there may be significant portions of the target molecule that do not have an exact match in the template molecule as determined by an MCES but still must be considered in the optimal partition. The subgraph radius score is defined as $R = \sum_{i=1}^{K}\omega_i\cdot|rad(S_i^1) - rad(S_{\mu(i)}^2)|$ where $rad(S_i^1)$ is the *radius* of subgraph (i.e., diversity domain) S_i^1 in the template graph G_1 and $S_{\mu(i)}^2$ is the target molecule subgraph that corresponds to S_i^1 as determined in the maximum weight clique mapping $\boldsymbol{M}$. If $K_s < K$ and there is no corresponding $S_{\mu(i)}^2$ for a given S_i^1, $rad(S_{\mu(i)}^2) = 0$. The *radius* of a (sub)graph is defined as the minimum *eccentricity* in the sub(graph). The *eccentricity* is the longest of all shortest paths between a vertex and all other vertices in the sub(graph).[39]

This score helps select for subgraph partitions that best approximate the relative size distribution of the template diversity domain partition while also considering connectivity within each subgraph. If the current value of R is less than the best radii difference score thus far detected (R^{best}), then the objective function returns a value of '*true*'; else if $R = R^{\mathrm{best}}$, then the subgraph size score (S) is evaluated. Otherwise a value of '*false*' is returned.

(3) Subgraph Size Score (S). Occasionally, multiple partitionings may exist that have both the same W and R score. To further discriminate the optimal partitioning in these instances, the subgraph size score is used to further enforce the appropriate size distribution. S is defined as $S = \sum_{i=1}^{K}\omega_i\cdot||S_i^1| - |S_{\mu(i)}^2||$ where $|S_i^1|$ is the cardinality (i.e., number of atoms) of subgraph S_i^1. If $K_s < K$ and there is no corresponding $S_{\mu(i)}^2$ for a given S_i^1, $|S_{\mu(i)}^2| = 0$.

If the current value of S is less than the best subgraph size score thus far detected (S^{best}), then the objective function returns a value of '*true*'; else a value of '*false*' is returned.

3. APPLICATION

We have introduced a conceptually simple and convenient method for automatically partitioning sets of potentially diverse compounds consistent with a user-specified template. It has proven to be remarkably robust with regard to structural diversity in the target molecule set and has been used successfully in medicinal chemistry projects to discover interesting substructural feature combinations that would not have been detected otherwise.

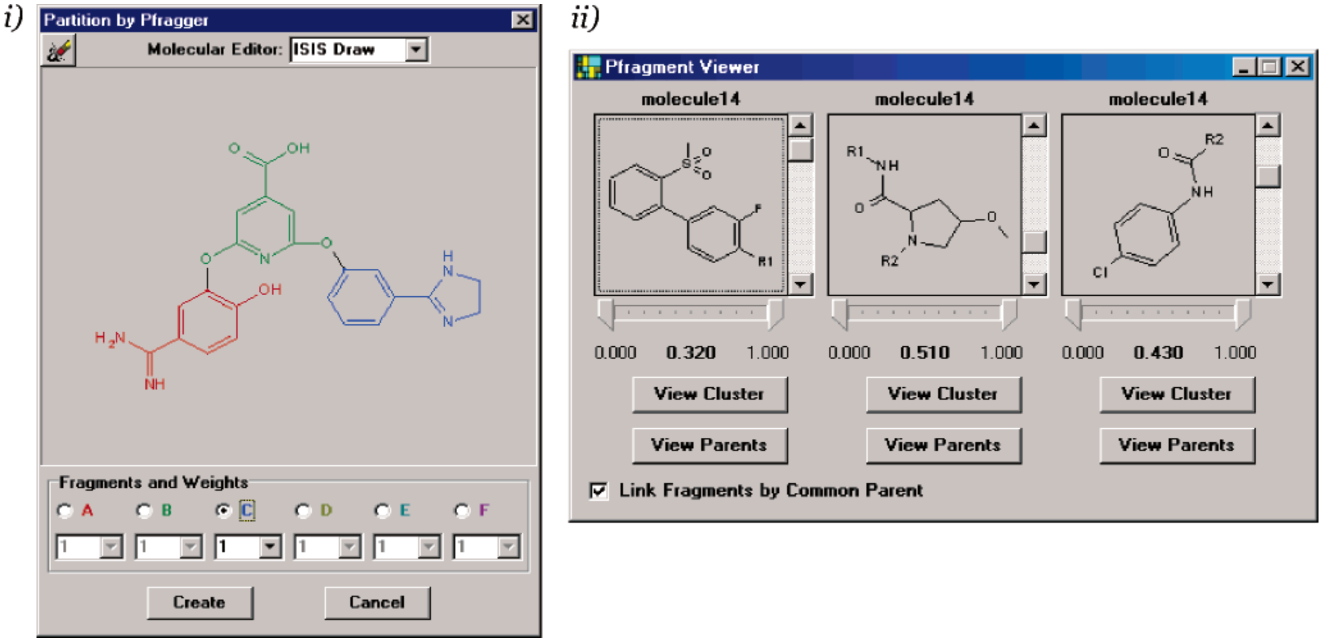
The algorithm is efficient enough to be used in real-time analysis and has been used on data sets as large as 2500 structures. We have found this to be a reasonable limit since this process is based on the expectation that the collection of molecules will be similar enough to the template that the resulting partitions will be meaningful. Since the proposed method has proven efficient enough for its intended purpose and efficiency is dependent upon several factors such as molecule size, structural diversity, and the number of specified diversity domains, a systematic investigation of efficiency has not been performed. We have found that an approximate time ~0.05 s per structure is typical in our system configuration (Windows 2000, Visual C++ 6.0, 3 GHz).

Currently the proposed algorithm has been included into two end-user applications. The first is a simple browsing/filtering tool called Pfragger Lite designed to output subsets of compounds fulfilling specific substructural similarity and property requirements.

This application allows the user to graphically specify a template, fragment a set of molecular structures in SD format using the previously described algorithm, and then filter the structures based upon the similarity of the generated fragments from the data set to their corresponding diversity domain templates as well as biological or physicochemical properties associated with the whole molecules. This is accomplished via a set of range sliders present in the graphical user interface (GUI) that allow the user to independently set different similarity criteria for each diversity domain fragment. In this fashion, the user is able to pose questions such as "What active molecules exist (using the property range slider) in my data set that have substructures very similar to my template diversity domains *A* and *B* but allow for significant diversity in the *C* substructure (using the respective fragment similarity range sliders)?". The user can view and scroll through the subset of structures meeting these criteria as well as export the subset as an SD file. In addition, the application allows the user to enumerate a virtual combinatorial library based on the substructures that meet the filtering criteria. These virtual compounds could then be browsed, docked, or scored using a QSAR model.

The Pfragger algorithm has also been incorporated into Molecular Property Explorer (MPX)[40] as a data mining tool to allow more sophisticated fragment-based analysis. MPX uses a reciprocal nearest neighbors (RNN) algorithm to cluster a data set on the basis of 2D chemical fingerprints or biological and physicochemical properties associated with molecules in the data set. MPX presents hierarchical clusters to the user in the form of interactive tree-maps and heatmaps. The combination of the Pfragger algorithm and MPX allows a researcher to partition the structures in a data set into fragments based on a diversity domain template and then visualize the influence these fragments may have on the biological and physicochemical properties associated with the parent structures. Furthermore, the interactive nature of the tree-map and heatmap visualizations allow one to quickly identify the parent structure and its related fragments associated with a particular fragment of interest.

Figure 7 shows both the input GUI and the results dialogue for the implementation of Pfragger within MPX using the diversity domains of the template structure from Table 1[5] applied to a set of 674 FXa inhibitors from the MDDR database. The Pfragger GUI allows the user to specify the template diversity domains by color using a lasso tool. The template structure of Figure 7(*i*) has been divided into three diversity domains labeled *A*, *B*, and *C*, respectively. The user may also specify the relative weighting for each diversity domain. In addition, the user may specify one or more diversity domains as a "core," in which case the diversity domain must have an exact match in the target structure. The range slider below each set of fragment structures



Molecular Editor: ISIS Draw

Fragments and Weights

○ A ○ B ● C ○ D ○ E ○ F

1 1 1 1 1 1

0.000 0.320 1.000

0.000 0.510 1.000

0.000 0.430 1.000

☑ Link Fragments by Common Parent

Figure 7. (*i*) Pfragger GUI and (*ii*) results dialogue implemented within MPX.

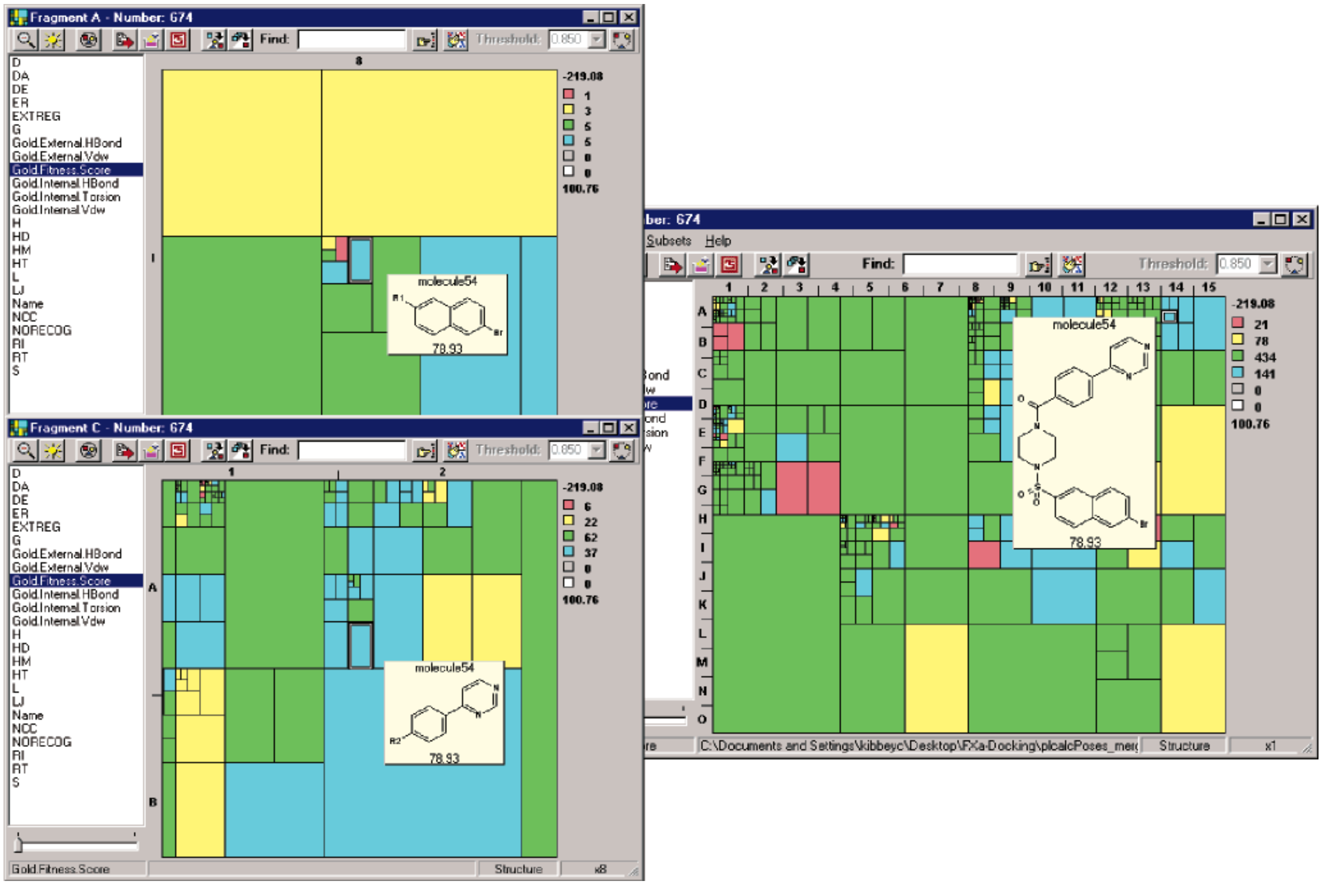


Figure 8. Diversity domain analysis within MPX highlighting diversity domain sets *A* and *C* for molecule 54.

controls the similarity thresholds for the diversity domains from the target structures to the corresponding diversity domain in the template by allowing the user to set maximum and minimum similarity values. The range sliders are used to constrain the number of fragment structures available within each diversity domain set. The “View Cluster” button generates a tree-map visualizing the clustering of the subsetted fragments, while the “View Parents” button highlights the corresponding parent structures of the subsetted fragments in the main MPX tree-map.

The screen shot in Figure 8 illustrates the visualization of clustered fragments from diversity domain sets *A* and *C* within MPX. The tree-maps corresponding to diversity domains *A* and *C* are depicted along with the main tree-map of the set of clustered parent structures. The tree-maps are colored by the binned value of the docking fitness score

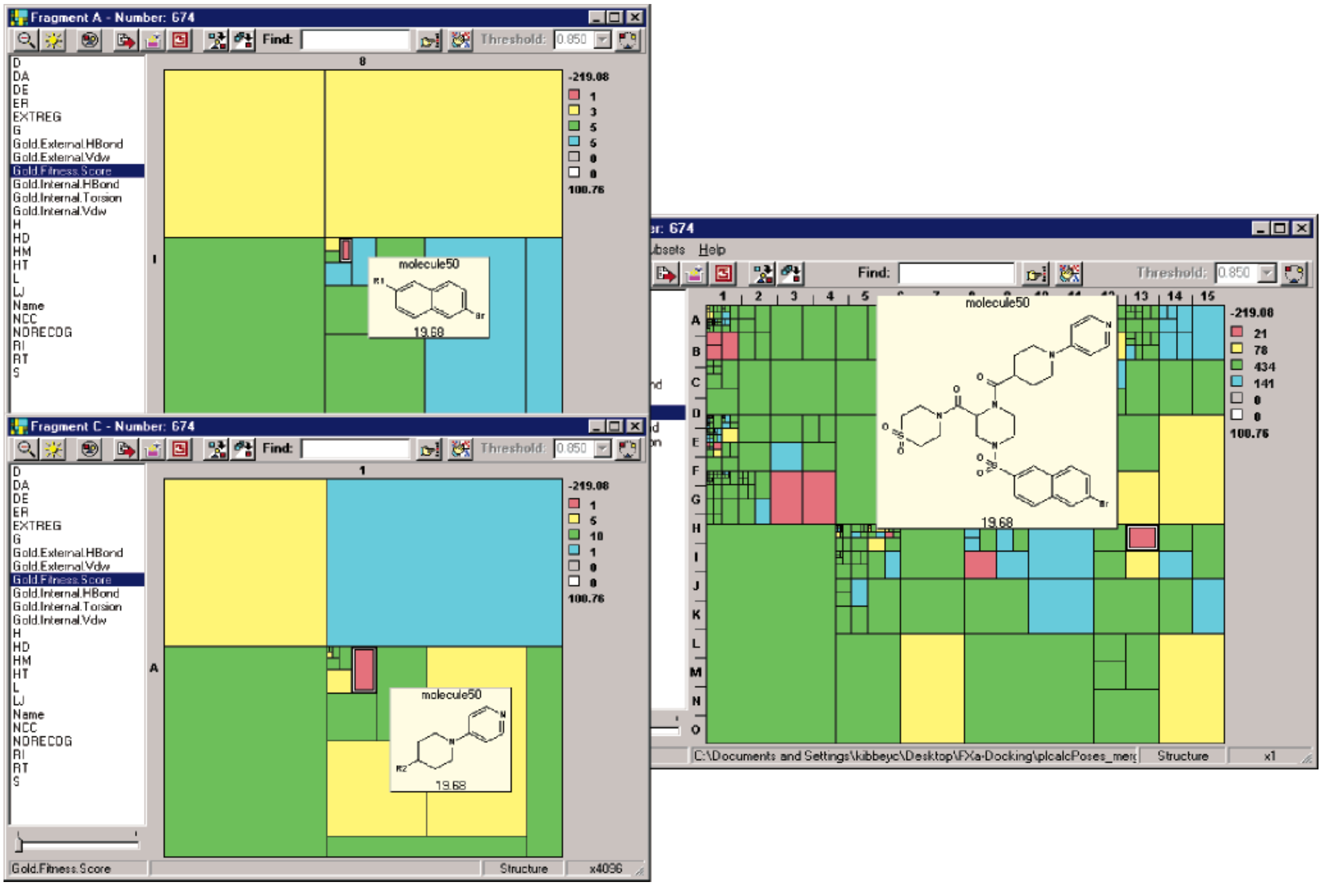


Figure 9. Diversity domain analysis within MPX highlighting diversity domain sets *A* and *C* of molecule 50.

computed for each parent structure. The fitness scores for the parent structures range from −219.88 to 100.76 and are divided among four bins as follows: −219.88 to 25 (red), 25 to 50 (yellow), 50 to 75 (green), and 75 to 100.76 (blue). The upper-right quadrant in the tree-map of the parent structures consists primarily of compounds with high docking score. The rectangle corresponding to molecule 54 has been selected in the tree-map of the parent structures, and the MPX software has highlighted the corresponding fragments in the tree-maps corresponding to diversity domain sets *A* and *C*. The nearest neighbors to the fragment of molecule 54 in the tree-map of diversity domain set *C* also correspond to parent structures with high docking score. However, the nearest neighbors to the fragment of molecule 54 in the tree-map of diversity domain set *A* correspond to parent structures with both low and high docking score. The subcluster shown in the tree-map of diversity domain set *A* (see Figure 8) consists of the same fragment from molecule 54 present in multiple parent structures of the data set. The observation that the same fragment is present in molecules with both low and high docking score suggests that another diversity domain within these structures may account for the differences in docking to the Factor Xa model.

The interactive design of the tree-maps within MPX provides a convenient means of exploring the influence of Pfragger derived diversity domain sets on various properties within the data set. For example, Figure 9 illustrates the selection of the fragment adjacent to that selected in diversity domain set *A* of Figure 8. MPX highlights the corresponding parent structure (molecule 50) and its related fragment in diversity domain set *C*. While molecules 50 and 54 share the fragment identified in diversity domain set *A* of Figures 8 and 9, they differ with respect to the fragment of diversity domain set *C* and in their substitution on the central piperazine ring (diversity domain set *B*). It is important to note that the fragment highlighted in diversity domain set *C* of Figure 9 also appears in parent structures whose docking score is significantly greater than that of molecule 50. The combination of substitution on the central piperazine ring and the presence of the piperidine-pyridine fragment in diversity domain set *C* appears to contribute to the poor docking score for molecule 50 relative to that of its (structural) nearest neighbors. Performing a similar analysis using the entire data set without the aid of MPX and Pfragger would likely be much more difficult.

4. CONCLUSION

In this paper, we have introduced a convenient and fully automated method for analyzing sets of molecular structures according to predefined substructural templates. It is a graph-based approach incorporating graph theoretic algorithms such as the maximum common edge subgraph (MCES), subgraph isomorphism, maximum weight clique, and the *k*-cut. The method has proven to be very robust on diverse sets of structures and efficient enough to be used on real-world data sets in a project setting. In addition, we have described two custom designed, in-house applications which incorporate the proposed method.

REFERENCES AND NOTES

(1) Merlot, C.; Domine, D.; Church, D. J. Fragment Analysis in Small Molecule Discovery. *Curr. Opin. Drug Discuss. Dev.* **2002**, *5*, 391−399.

(2) Merlot, C.; Domine, D.; Cleva, C.; Church, D. J. Chemical Substructures in Drug Discovery. *Drug Discovery Today* **2003**, *8*, 594−602.

(3) Reuveni, H. et al. Toward a PKB Inhibitor: Modification of a Selective PKA Inhibitor by Rational Design. *Biochemistry* **2002**, *41*, 10304−10314.

(4) Giblin, G. et al. (2-((2-Alkoxy)-Phenyl)-Cyclopent-1-Enyl) Aromatic Carbo and Hetercyclic Acid and Derivatives, WO 03/084917 A1, 2003.

(5) Maignan, S.; Mikol, V. The Use of 3D Structural Data in the Design of Specific Factor Xa Inhibitors. *Curr. Topics Med. Chem.* **2001**, *1*, 161−174.

(6) Cappelli, A. et al. Novel Potent and Selective Central 5-HT3 Receptor Ligands Provided with Different Intrinsic Efficacy. 1. Mapping the Central 5-HT3 Receptor Binding Site by Arylpiperazine Derivatives. *J. Med. Chem.* **1998**, *41*, 728−741.

(7) Walpole, C. S. J. et al. Analogues of Capsaicin with Agonist Activity as Novel Analgesic Agents; Structure−Activity Studies. 1. The Aromatic “A-Region”. *J. Med. Chem.* **1993**, *36*, 2362−2372.

(8) Walpole, C. S. J. et al. Analogues of Capsaicin with Agonist Activity as Novel Analgesic Agents; Structure−Activity Studies. 2. The Amide Bond “B-Region”. *J. Med. Chem.* **1993**, *36*, 2373−2380.

(9) Walpole, C. S. J. et al. Analogues of Capsaicin with Agonist Activity as Novel Analgesic Agents; Structure−Activity Studies. 3. The Hydrophobic Side-Chain “C-Region”. *J. Med. Chem.* **1993**, *36*, 2381−2389.

(10) Pevarello, P. et al. Synthesis and Anticonvulsant Activity of a New Class of 2-[(Arylalkyl)amino]alkanamide Derivatives. *J. Med. Chem.* **1998**, *41*, 579−590.

(11) Hazeldine, S. T. et al. Design, Synthesis and Biological Evaluation of Analogues of the Antitumor Agent, 2-{4-[7-Chloro-2-quinoxalinyl)oxy]phenoxy}proprionic Acid (XK469). *J. Med. Chem.* **2001**, *44*, 1758−1776.

(12) Molteni, V. et al. N-Phenylphenylglycines as Novel Corticotropin Releasing Factor Receptor Antagonists. *J. Med. Chem.* **2004**, *47*, 2426−2429.

(13) Lebreton, L.; Annat, J.; Derrepas, P.; Dutartre, P.; Renaut, P. Structure-Immunosuppresive Activity Relationships of New Analogues of 15-Deoxyspergualin. 1. Structure Modifications of the Hydroxyglycine Moiety. *J. Med. Chem.* **1999**, *42*, 277−290.

(14) Lebreton, L. et al. Structure-Immunosuppressive Activity Relationships of New Analogues of 15-Deoxyspergualin. 2. Structural Modifications of the Spermidine Moiety. *J. Med. Chem.* **1999**, *42*, 4749−4763.

(15) Gellibert, F. et al. Identification of 1,5-Naphthyridine Derivatives as a Novel Series of Potent and Selective TGF-b Type I Receptor Inhibitors. *J. Med. Chem.* **2004**, *47*, 4494−4506.

(16) Swanson, D. M. et al. Identification and Biological Evaluation of 4-(3-Trifluoromethylpyridin-2-yl) piperazine-1-carboxylic Acid (5-Trifluoromethylpyridin-2-yl)amide, a High TRPV1 (VR1) Vanilloid Receptor Antagonist. *J. Med. Chem.* **2005**, *48*, 1857−1872.

(17) Cammarata, A.; Menon, G. K. Pattern Recognition. Classification of Therapeutic Agents According to Pharmacophores. *J. Med. Chem.* **1976**, *19*, 739−748.

(18) Trinajstic, N. *Chemical Graph Theory*; CRC Press: 1992.

(19) Raymond, J.; Willett, P. Maximum Common Subgraph Isomorphism Algorithms for the Matching Of Chemical Structures. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 521−533.

(20) Syslo, M.; Deo, N.; Kowalik, J. Shortest Path Problems. In *Discrete Optimization Algorithms*; Prentice-Hall: 1983; pp 227−253.

(21) Raymond, J.; Gardiner, E.; Willett, P. Heuristics for Rapid Similarity Searching of Chemical Graphs Using a Maximum Common Edge Subgraph Algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 305−316.

(22) Raymond, J.; Gardiner, E.; Willett, P. RASCAL: Calculation of Graph Similarity Using Maximum Common Edge Subgraphs. *Comput. J.* **2002**, *45*, 631−644.

(23) Raymond, J.; Willett, P. Effectiveness of Graph-Based and Fingerprint-Based Similarity Measures for Virtual Screening of 2D Chemical Structure Databases. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 59−71.

(24) Ullmann, J. R. An Algorithm for Subgraph Isomorphism. *J. ACM* **1976**, *23*, 31−42.

(25) Brint, A.; Willet, P. Pharmacophoric Pattern Matching in Files of 3D Chemical Structures: Comparison of Geometric Searching Algorithms. *J. Mol. Graph.* **1987**, *5*, 49−56.

(26) Goldschmidt, O.; Hochbaum, D. S. A Polynomial Algorithm the k-Cut Problem for Fixed k. *Math. Oper. Res.* **1994**, *19*, 24−37.

(27) Yeh, W. C. A Simple Branch-and-Bound Algorithm for the k-Cut Problem for Applications with k Target Vertices, e.g., VLSI Design. *Int. J. Adv. Manuf. Technol.* **2002**, *20*, 63−71.

(28) Karger, D. R.; Stein, C. A New Approach to the Minimum Cut Problem. *J. ACM* **1996**, *43*, 601−640.

(29) Karypis, G.; Kumar, V. A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs. *SIAM J. Sci. Comput.* **1998**, *20*, 359−392.

(30) Brassard, G., Bratley, P. *Algorithmics*: *Theory and Practice*; Prentice Hall: 1988.

(31) Allburn, E. Graph Decomposition: Imposing Order on Chaos. *Dr. Dobbs J.* **1991**, *16*, 88, 90−2, 94−6, 118−20, 122, 124.

(32) Recuero, A. Algorithms for Path Searching and for Graph Connectivity Analysis. *Adv. Eng. Software* **1995**, *23*, 27−35.

(33) Carpaneto, G.; Martello, S.; Toth, P. Algorithms and Codes for the Assignment Problem. *Ann. Oper. Res.* **1988**, *13*, 193−223.

(34) Wood, D. An Algorithm for Finding a Maximum Clique in a Graph. *Oper. Res. Lett.* **1997**, *21*, 211−217.

(35) Babel, L. A Fast Algorithm for the Maximum Weight Clique Problem. *Computing* **1994**, *52*, 31−38.

(36) Pardalos, P.; Xue, J. The Maximum Clique Problem. *J. Global Optim.* **1994**, *4*, 301−328.

(37) Kumlander, D. A New Exact Algorithm for the Maximum-Weight Clique Problem Based on a Heuristic Vertex-Coloring and a Backtrack Search. In *Applied Mathematical Programming and Modelling*; Brunel University: West London, U.K., 2004.

(38) 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.

(39) Harary, F. *Graph Theory*; Addison-Wesley: 1994.

(40) Kibbey, C.; Calvet, A. Molecular Property eXplorer: A Novel Approach to Visualizing SAR Using Tree-Maps and Heatmaps. *J. Chem. Inf. Model.* **2005**, *45*, 523−532.

CI0502247