

# Molecular Substructure Mining Approaches for Computer-Aided Drug Discovery: A Review

Christos A. Nicolaou, and Constantinos S. Pattichis, *Senior Member, IEEE*

**Abstract**—Substructure mining is a well-established technique used frequently in drug discovery. Its aim is to discover and characterize interesting 2D substructures present in chemical datasets. The popularity of the approach owes a lot to the success of the structure-activity relationship practice, which states that biological properties of molecules are a result of molecular structure, and to expert medicinal chemists who tend to view, organize and treat chemical compounds as a collection of their substructural parts. Several substructure mining algorithms have been developed over the years to accommodate the needs of an ever changing drug discovery process. This paper reviews the most important of these algorithms and highlights some of their applications. Emphasis is placed on the recent developments in the field.

## I. INTRODUCTION

The concept of pharmacophore, the specific arrangement of molecular features forming a necessary (but not sufficient) condition for biological activity [1], is central to drug design. Approaches aiming to recognize pharmacophores capitalize on the measured biological property values of a set of compounds to isolate, describe and use biologically significant chemical features. In practice, these methods mainly focus on information provided by sets of compounds known to bind strongly to the desired pharmaceutical target [2]. In order to refine the set of significant chemical features and improve the quality of the pharmacophore model sometimes the methods also take into account information provided by structurally similar compounds that fail to show activity [3], [4].

Pharmacophore models and representations take various forms [1]. A number of approaches attempt to construct a model corresponding to the complete pharmacophore in 2D or 3D. Such approaches usually start by standardizing the chemical representation of compounds in a pre-processing step and determining the chemical properties of the various sites of each molecule. Following they proceed to identify substructures frequently occurring in a known ligand set. Other approaches adaptively calculate 2D or 3D descriptors and construct more abstract representations of known ligands. Since the representations are “learned” from the ligands it is assumed that some of the descriptors capture the

pharmacophore and therefore can be used as a general description of it for a variety of purposes. However, descriptor-based approaches are plagued by several limitations the most important being the loss of information associated with the reduction of a 2D topological graph (or 3D-conformation) to a numerical vector, often a bit-vector also known as fingerprint [3], [5], [6].

Approaches based on substructure mining focus on the identification of sizeable 2D structural commonalities among ligands. Depending on the availability of supporting biological information these commonalities may be referred to as scaffolds, privileged substructures or 2D pharmacophores. Compared to 2D or 3D descriptors these approaches have the advantage of preserving topological information related to chemical structure. Compared to 3D pharmacophore representations they have the clear advantage of simplicity and speed. This is mainly due to the dependence of 3D methods on the generation of multiple conformations for the molecules under investigation and the requirement for reliable superposition of the molecules [7]. In addition, and contrary to popular expectation, there is no clear performance advantage of 3D representations over their 2D counterparts. For example, Brown and Martin report that in the context of ligand-receptor binding 2D descriptors outperform 3D ones [8]. Similarly, in a study by Hessler *et al.* a MTree 2D pharmacophore model compared favorably with a 3D model in a virtual screening experiment [7].

Once generated scaffolds can be used in a multitude of ways. Prime among them has historically been the task of biochemical screening data organization and interpretation [3], [4], [6], [9]–[11]. Equally popular are techniques employing scaffolds for virtual screening [2], [7], [12], [13]. These techniques use scaffolds to select a subset of compounds from a larger set e.g. via substructure search. More recently privileged substructures have been used for molecular library design and ligand design [7], [14]. Both approaches use scaffolds as templates to either select sets of molecules covering the pharmacologically interesting space defined by the scaffold (library design) or for virtually synthesizing compounds with high likelihood of exhibiting increased biological activity (ligand design).

This paper reviews the topic of substructure mining from sets of chemical compounds and summarizes recent trends. Section II briefly describes some basic background information on graph theory and lists necessary definitions. Section III reviews some milestone techniques as well as the

Manuscript received October 9, 2006.

C. A. Nicolaou is with the University of Cyprus, Computer Science Department, Nicosia, Cyprus. (phone: +357-2289-2685; fax: +357-2289-2701; e-mail: cnicolaou@cs.ucy.ac.cy).

C. S. Pattichis is with the University of Cyprus, Computer Science Department, Nicosia, Cyprus.

present state of the art methods for substructure mining. The final section summarizes the methods described and describes our future work plan.

## II. BACKGROUND INFORMATION

### A. Graph Theory Fundamentals

A graph  $G = (V, E)$  consists of a set of vertices  $V(G)$  and a set of edges  $E(G)$ . In the case of labeled graphs both vertices and edges have identifiers, i.e. each vertex and edge has a label drawn from a predefined set of vertex labels  $L_V$  and edge labels  $L_E$ . Note that vertices and edges need not have unique labels, as is the case in molecular graphs where, for example, multiple vertices in any drug-like molecule have the C (carbon) label. Graphs can be directed or undirected. In directed graphs edges are ordered pairs of the vertices they connect where, in undirected, edges simply list the pair of vertices they connect. A vertex  $V_i$  is said to be incident with an edge if one of the two endpoints of the edge is  $V_i$  while an edge  $E_a$  is incident with an edge  $E_b$  if they have a vertex in common. Two vertices  $V_i, V_j$  of graph  $G$  are connected, or adjacent, if there is an edge  $E_{ij} = (V_i, V_j) \in E(G)$ . If there is a path  $P = (E_1, E_2, \dots, E_n)$  between every pair of vertices in a graph  $G$ , then  $G$  is a connected graph.

A graph  $S = (V_S, E_S)$  is a subgraph of  $G = (V, E)$  if and only if  $V_S \subseteq V$  and  $E_S \subseteq E$ . If  $E_S$  contains all edges in  $E$  connecting the vertices in  $V_S$  then  $S$  is an induced subgraph of  $G$ . An additional property of induced subgraphs is that it can be shown that there is a one-to-one mapping between the edges in  $E_S$  and all edges in  $E$  incident on vertices in  $V_S$  when  $V_S$  is mapped on  $V$ . A clique is a special case of an induced subgraph where all its vertices are incident on each other. A maximum clique of a graph  $G$  is its largest clique.

The problem of determining whether two graphs are identical is known as graph isomorphism [15]. In graph theoretic terms two graphs  $G_1 = (V_1, E_1)$  and  $G_2 = (V_2, E_2)$  are isomorphic if there is a mapping from  $V_1$  to  $V_2$  such that there exists a mapping for each edge in  $E_1$  to an edge in  $E_2$ . A common induced subgraph between  $G_1$  and  $G_2$  is a graph  $CS$  that is an induced subgraph of both  $G_1$  and  $G_2$ . The largest induced subgraph between  $G_1$  and  $G_2$  is known as the Maximum Common Induced Subgraph (MCIS). A related concept is that of Maximum Common Edge Subgraph (MCES) also known as Maximum Overlapping Set (MOS). An MCES is a subgraph consisting of the largest number of edges common to both  $G_1$  and  $G_2$  [5].

It is worth pointing out that the MCIS and MCES between two graphs may consist of several disconnected subgraphs as seen in fig. 1. The largest contiguous common substructure is known as the Maximum Common Substructure (MCS). Informally, the MCS of two graphs  $G_1$  and  $G_2$  is the largest possible graph that is isomorphic to subgraphs of  $G_1$  and  $G_2$ .

Several algorithms have been proposed in the literature dealing with the problem of graph isomorphism. Among

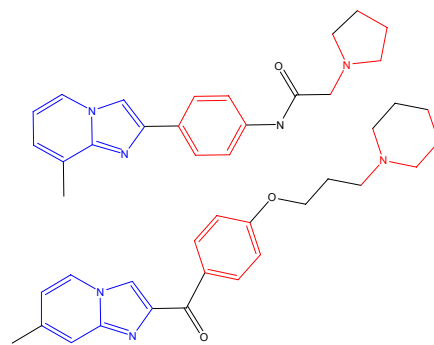


Figure 1: A pair of compounds from SM05 and their corresponding MCS (in blue) and MOS (in blue and red).

them the category of clique finding algorithms is one of the most popular [5], [15]. These methods rely on the calculation of a *new graph*, the compatibility graph  $CG$ , via a modular product operation on graphs  $G_1$  and  $G_2$ . Following is the determination of the maximum clique in the  $CG$  by applying one of the many maximum clique algorithms available. The maximum clique of the  $CG$  has been shown to be equivalent to the MCIS of the two input graphs [16]. The modular product of the graphs  $G_1$  and  $G_2$ , denoted as  $G_1 \diamond G_2$ , is defined in (1).

$$V(G_1 \diamond G_2) = V(G_1) \times V(G_2) \quad (1)$$

where two vertices  $(u_i, v_i)$  and  $(u_j, v_j)$  are adjacent if:

$$(u_i, v_i) \in E(G_1) \text{ and } (u_j, v_j) \in E(G_2) \text{ or } (u_i, v_i) \notin E(G_1) \text{ and } (u_j, v_j) \notin E(G_2)$$

Note that the vertices of  $CG$  consist of pairs of vertices, one vertex from each input graph. For a more detailed explanation please look at [5], [16].

### B. Chemical Scaffolds, Privileged Substructures and 2D Pharmacophores: Definitions

Chemical structures are typically represented as labeled, undirected graphs where atoms correspond to vertices, and chemical bonds are represented by edges. In this context, molecular fragments, or substructures, are induced subgraphs of molecular graphs. Scaffolds are molecular fragments defined in association to well-defined sets of compounds. A scaffold derived from a compound set is a substructure present in all the compounds of the set. A common requirement imposed on scaffolds is that they are sufficiently large so as to be useful in characterizing the compounds in their corresponding sets. Often, a compound set scaffold is thought of as the MCS or MCIS of that set. It is worth noting that the problem of MCIS calculation in molecular graphs can also be solved using modular product based approaches provided that a provision is made to distinguish labeled vertices (atoms) and edges (bonds). More informally scaffolds can be thought of as the common, distinguishing “core” of a compound set. Scaffold-based analysis of a large compound set refers to the process of identifying frequently occurring substructures in the compounds of the set -the scaffolds- and the subsets

associated with them. Privileged substructures are scaffolds positively correlated with favorable behavior, i.e. scaffolds present preferentially in compounds with a desired biological profile. 2D pharmacophores are those special privileged substructures that capture the key molecular features necessary for biological activity.

### III. SUBSTRUCTURE MINING APPROACHES

Chemical substructure mining amounts to processing numerous undirected, labeled graphs and discovering common subgraphs of substantial size. In practice the methods take as input a set of known ligands and use some algorithm to detect substructures frequently occurring in large subsets of the ligands. There are two general categories of such methods: the first uses a predefined list of candidate scaffolds composed of a selection of substructures. Substructure searching is used to identify and select all candidate scaffolds found in numerous compounds in the set under investigation. In contrast, the methods of the second category adaptively learn the substructures from the compounds in the set. Each category of methods has some advantages but also some drawbacks. However, most recent approaches proposed belong to the second category since it has the clear advantage of detecting scaffolds specific, and possibly unique, to each input compound set supplied. In the following sections we review several scaffold identification methods. The initial sections describe some of the early, simpler approaches that served as the stepping-stones for later developments. More emphasis is placed on the presentation of the more recent methods reviewed in the later sections.

#### A. Stigmata

One of the earliest attempts in the area of frequent substructure identification has been the Stigmata algorithm designed to find structural commonalities in chemically diverse datasets [17]. The key feature of the algorithm is the generation of a “modal fingerprint” of the input chemical dataset. Initially Stigmata calculates molecular fingerprints for the chemical structures supplied. A bit of the modal fingerprint is set on if the corresponding descriptor key can be found in at least a certain portion of the entire collection of molecular fingerprints. The threshold value for turning a bit on or off is user defined, usually ranging between 0.5 (half the chemical structures contain the key corresponding to the bit) and 1 (all the chemical structures containing the key corresponding to the bit). Post-processing of the modal fingerprint reveals the bits and their corresponding fragments that are common to the input chemical structures. A visual interface is typically employed to highlight the structural commonalities found. A limitation of the method is that it can only identify structural commonalities encoded as structural descriptors. This makes the method overly dependent on the choice of molecular fingerprint method used to describe molecules.

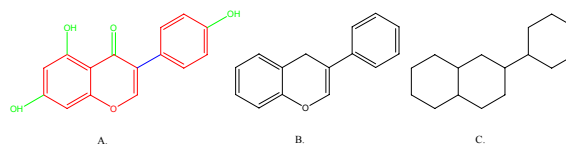


Figure 2. Frameworks analysis of genistein (A). In a first step the molecule is dissected into 3 units: rings (red), linkers (blue) and side-chains (green). Frameworks are the unions of rings and linkers. Two types of frameworks were proposed: Frameworks with exact atoms (B) and purely graph-based frameworks (C).

#### B. Frameworks Analysis

Another early approach is the so-called frameworks analysis proposed by Bemis and Murcko [12]. The method is designed to provide a “high-level overview” of gross structural features of substantial size, the so-called frameworks, present in a set of molecules. According to the method each molecule is dissected into three units: rings, linkers and side-chains. Rings are the cycles of the molecular graph while linkers are any substructures connecting two rings. Side-chains are the leftovers, e.g. any substructures that are not part of a ring or a linker. Frameworks are defined to be the union of rings and linkers [12]. The authors proposed two types of frameworks depending on whether atom or purely graph-based representations are used (fig. 2).

The process generates a set of frameworks, one for each molecule. Duplicate frameworks are removed from the set and a count of the number of occurrences of each framework is kept. Bemis and Murcko originally used their method for discovering frameworks in a set of drugs with the intent of using them for the purposes of similarity searching and library design.

An extension to the frameworks analysis method has been developed since then aiming to reduce the number of frameworks and generalize the resulting substructures. To achieve this, the frameworks are subjected to a Maximum Common Substructure (MCS) analysis and the MCS found is used to remove the molecules containing it from the dataset under investigation. The process of MCS extraction and molecule removal is repeated on the remaining compounds from the original dataset until a user-defined portion of the dataset has been removed [14]. This process reduces significantly the number of characteristic substructures needed to describe a set of molecules.

A different variation of framework analysis, Scaffold-Based Classification (SCA), has been proposed to classify compound libraries using scaffolds [9]. The novelty of SCA is mainly found in the introduction of the concept of scaffold complexity and the technique used to calculate the similarity of a compound to a scaffold. Unlike traditional frameworks analysis where scaffolds are used for substructure searching, SCA characterizes scaffolds further and uses those characteristics to sort them and form compound classes around them. This process eventually leads to increased

generality of scaffolds and thus fewer, larger and more diverse classes [9].

### C. PhyloGenetic-Like Trees

The extraction of MCS has also been employed in various other algorithms designed to identify privileged substructures for specific pharmaceutical projects. Nicolaou *et al.* [3] proposed the PhyloGenetic-Like Tree (PGLT) algorithm designed to identify chemical classes characterized by a chemical substructure from diverse screening datasets. The algorithm produces a dendrogram data structure where each node contains a set of molecules and a chemical fragment common to the molecules. The construction of the dendrogram is achieved via the following process: First, the root node containing the compounds to be analyzed is formed. Then the compounds in the node are clustered, the natural/interesting clusters are determined and an MCS extraction process is applied on each of those clusters. The MCS's are then evaluated using some expert rules and those deemed interesting are used to create new nodes that are connected to the parent node. The new nodes are populated with the compounds of the parent node that contain the MCS characteristic of the node. The process iterates on each node that is sufficiently large. The end result is a hierarchy of chemical substructures found in the input dataset with smaller, more general substructures close to the root of the tree and larger, more specific substructures closer to the leaves. In its original implementation the algorithm was applied on the active portion of a screening dataset to identify chemical substructures frequently occurring among active compounds. The discovered substructures were matched with inactive compounds in order to assess whether they were indeed privileged, i.e. related to the activity observed or simply chance correlations.

PGLT was probably the first algorithm for discovering privileged substructures cited in the literature where clustering and the MCS process played key roles. The process was indeed able to define a hierarchy of chemical substructures frequently occurring in a dataset and populate this dendrogram with volumes of screening data. However, the algorithm had poor performance and required lengthy times for constructing the dendrogram on large numbers of data. This was mainly due to the requirements of the MCS generation step and the repeated usage of clustering. PGLT has been used successfully in several occasions, among them by Bacha *et al.* [18] for mutagenicity rule extraction and Rodgers *et al.* [10] for the identification of chemical substructures related to bitterness.

### D. Distill

An alternative approach that also forms a hierarchical organization of compounds using common substructures is that employed in the commercial package Distill [19], [20]. As in the PGLT algorithm each node in the constructed dendrogram represents a substructure and the compounds containing that substructure. Contrary to the PGLT Distill is

bottom-up and as such it begins by forming leaf nodes out of all the compounds in the dataset to be analyzed. The process then computes the pair-wise similarity of all the nodes, merges the two closest ones to form a new node and iterates until only one node is left or a certain predefined criterion, such as node size, is met. Similarity between nodes is based on the actual MCS fragment extracted from the characteristic substructures of the two nodes. Note that the MCS extraction algorithm takes as input the characteristic substructures of each node and not the compounds of the node. This "shortcut" enables the algorithm to only extract MCS's from pairs of substructures and avoid dealing with larger datasets at each node. The associated improvement in performance and savings in computational resources are compromised since the characteristic substructure of each node may not be the true MCS of the compounds in the node. However, even with this compromise the algorithm can only process a few hundred compounds in reasonable time since the requirement for performing all the pair-wise MCS extractions, on which similarity calculations are based, is both substantial and of order  $O(N^2)$ . Distill has been used by Shen [11] in creating the HAD system, a Hits Data Analysis decision support system designed to provide a structural overview of classes in a screening dataset and the associated activity statistics.

### E. ClassPharmer

More recently a set of commercially available software tools for the identification of privileged substructures have been reported. These tools have been designed to break large sets of chemical compounds in groups sharing substantially large substructures. These characteristic substructures may be derived using MCS extraction [4], or via other approaches capable of detecting a Significant Common Substructure (SCS) [21] in a set of structurally similar compounds. ClassPharmer [4] is a non-hierarchical method designed to detect chemical families of compounds in a dataset defined by large common substructures. The tool uses graph-based analysis to derive molecular fragments that capture commonalities in a given ligand training set [14]. Essentially, the method detects approximations of all the potential MCS's that could be derived from a chemical dataset and then employs an optimization procedure to select the subset of approximations/substructures that could be used to form clusters of compounds that satisfy a predefined minimum intra-cluster similarity criterion. Special care is made for selecting the subset of substructures that would result in the least number of singleton classes [4]. An MCS extraction process is used on each class to refine the initial approximations used to form the classification. The resultant classes of compounds and their representative MCS's can be further analyzed through importation of activity data and extraction of structure-activity conclusions. Thus, privileged substructures in the analyzed dataset can be highlighted based on the biological properties of the compounds in the

associated class.

ChemTK [21] uses a similar approach for the classification of molecular datasets using SCS fragments. According to the method the classification is performed in two steps. The first step involves searching the full set of molecules for all unique structural fragments that are present. Various types of fragments can be identified including ring systems (single and fused rings, as well as connected groups of these) and branched fragments (structures created by combining one or more unbranched fragments). The second step is to form classes using the list of identified structures. Each fragment identified in the first step is considered as a potential new class, the members of which include each molecule having the particular fragment as part of its molecular structure. The final set of classes is a subset of this full collection, chosen so as both to minimize the number of singletons and to limit the level of redundancy in the final classes [21]. An optional MCS extraction step for each class is available to ensure the best possible chemical structure characterization of the classes.

The limiting step of approximation-based approaches is the number of fragments generated that form the collection of potential classes. The number is proportional to the complexity of the molecules in the data and can become very large even for data sets of modest size. Even so, and assuming that the approximations used are truly representative of the MCS's the method is capable of processing chemical datasets of substantial size (ca 100k) given sufficient time on modern workstations [21].

#### F. RASCAL/SPINFEX

A different category of graph-based clustering methods has focused on defining and using graph-based similarity measures and implementing algorithms efficient enough to allow application on sizeable chemical datasets. A representative clustering algorithm of this type is Spinifex [6] based on the RASCAL graph-based similarity measure [5].

In its simplest form graph-based similarity of two compounds may be some measure of their calculated MCS. For example, the Distill application described above uses the size of the MCS of two compounds as their actual similarity value. Typical MCS-based similarity requires that a pair of compounds share a single substantially large contiguous substructure in order to have a high similarity value. While this is intuitive there are cases where it leads to misleading results. For example, such method fails to capture the true similarity of the compounds in fig. 1 since the various identical components of the molecules are connected with different linkers [6]. To address this limitation of MCS-based similarity Rapid Similarity CALCulation (RASCAL) proposed the usage of the Maximum Overlapping Set (MOS) [5] that may consist of several, disconnected substructures, and thus, capture the global graph-based similarity of two molecules.

RASCAL consists of two major components, graph matching where the MOS is identified, and similarity calculation. The graph matching procedure is based on the reduction of the MOS problem to the maximum clique problem. To achieve this the algorithm transforms the pair of input molecular graphs to their corresponding line -or edge- graphs. A line graph  $L(G_1)$  is a graph whose vertices correspond to the edges of graph  $G_1$  and whose edges correspond to  $G_1$ 's vertices. This transformation is followed by the calculation of the modular product of the line graphs and the detection of the maximum clique in the resulting compatibility graph. The isomorphism of the two line graphs, indicated by the cliques in the compatibility graph, has been proven to correspond to the MOS of the original graphs  $G_1$  and  $G_2$  [5]. The above principle does not hold for certain well-defined graph shapes and so RASCAL detects these problematic shapes and treats them differently. Discussing them is out of the scope of this paper and the interested reader is referred to the literature for more information. RASCAL uses several heuristics to improve its performance. Among them is "screening" used to exclude from expensive graph matching calculation pairs of molecules that cannot satisfy a user-defined similarity threshold. Also used are heuristics that simplify the modular graph by deleting nodes or edges based on the symmetry of certain molecular shapes and the resulting redundancy [16]. Upon calculation of the MOS between two molecules RASCAL uses (2) [5] for the calculation of their similarity value:

$$S = \frac{2(NV(MOS) + NE(MOS))}{(NV(G1) + NE(G1))(NV(G2) + NE(G2))} \quad (2)$$

where NV is the number of vertices and NE is the number of edges of a chemical structure.

Spinifex [6] uses RASCAL with an enhanced version of the similarity measure to calculate all pair-wise similarities between compounds in a dataset. The MOS implementation in Spinifex requires both atom and bond matching of the chemical graphs. Further, during a preprocessing step, the algorithm performs selected atom-typing, i.e. groups specific sets of atoms so that they are considered the same atom as in the case of halogens (F, Cl, Br, I), or differentiates occurrences of a single atom into two groups based on their expressed chemical properties as in the case of nitrogen which may (or may not) behave as a hydrogen bond donor. All calculated similarity values are stored in a proximity matrix. Following the calculation of the matrix a variety of clustering algorithms can be applied. The authors propose the usage of a hierarchical method, the Group-Average, which they have found to perform better in a series of tests.

Comparisons using a test set indicated that the Spinifex approach outperformed clustering using MCS or fingerprint-based similarity measures with respect to the number and the purity of the clusters produced as well as the number of singleton compounds. However, in practical terms the method is troubled by performance issues and can only be

applied to small to medium size datasets, ca. 3000 compounds. This is most notably due to the quadratic requirement imposed by the need to calculate a proximity matrix using MOS-based similarity.

### G. Feature Trees and MTrees

Rarey and Dixon [22] have described FeatureTrees, an interesting version of a 2D-based topological descriptor with potential applications in substructure mining and scaffold-based applications. Feature trees are essentially 2D representations of molecules designed to address the loss of information associated with linear representations of molecules at the cost of increased descriptor complexity. A feature tree is a graph whose nodes represent hydrophobic centers and functional groups of a molecule while its edges are modeled after the way these groups are linked together [13], [22]. The first step for creating a feature tree of a molecule is to identify the rings of the molecule. Single nodes represent ring centers and edges are drawn to connect them if the original rings were fused (e.g. shared atoms/edges). If after ring center nodes are added cycles can be found, the possible result of ring systems, those cycles are collapsed into single nodes representing their center. In this manner the new graph is guaranteed to be acyclic and thus simpler to manipulate. The next step of the algorithm is to traverse the molecule and add new nodes to the graph for each atom that has more than two bonds. Terminal atoms form a single node with the atom they are connected to. Edges are used to connect nodes that represent atom(s) connected in the molecule.

The second part of feature tree construction deals with the definition of features and the labeling of nodes accordingly. At a first pass the algorithm traverses the acyclic graph and calculates steric and chemical features for each node. Steric features aim to describe the size of the node and include the number of atoms and an approximated van der Waals volume of the fragment. The chemical features of a node focus on its chemical properties that could be used for protein-ligand interactions. Chemical features are computed for a fixed number of interaction types and are stored in a data structure called the "interaction profile" of the node. In essence, the profile is an array where the  $i$ -th entry describes the ability of the node to form an interaction of type  $i$  [22].

The resulting data structure is a simplified abstraction of molecules that preserves their topology and adds potential pharmacophore point information. Similarity between feature trees is calculated using specifically designed shape matching algorithms that exploit the acyclic nature of graphs and take into account the interaction profile of the nodes. Feature trees have been shown to be rather successful in grouping compounds having the same pharmacophore structure. Other uses include the prediction from a database of possible new lead molecules, often with substantially different molecular structure but sharing the same topology of pharmacophore points [23].

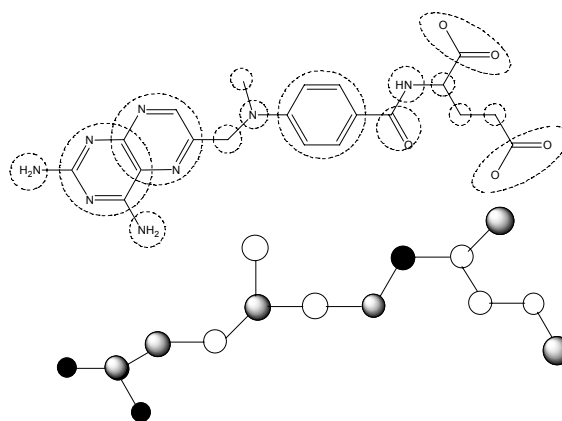


Figure 3.: Methotrexate and its corresponding feature tree. Feature trees collapse groups of atoms, such as rings, into single nodes. The nodes are categorized as hydrophobic (empty circles), hydrogen-bond donors (black), hydrogen bond acceptors (gray). The edges in a feature tree connect fragments of the original molecule sharing atoms or bonds. Adapted from [22].

Recently, an algorithm for combining the information from pairs of feature trees into a new tree has been proposed [7]. The process is based on chemically reasonable matching of corresponding functional groups as encoded by the feature trees. The nodes of the new tree represent the matches containing the features of the mapped subtrees while the edges are formed following the topologies of the input feature trees [7]. When presented with a set of input molecules the algorithm first converts them into their corresponding feature trees and then extracts a topological model by comparing those trees in a pair-wise fashion. The latter process begins by selecting a pair of feature trees, combining them to form a new tree and then repeating the process using as input the new tree with another feature tree of the initial set. Each resulting new tree is combined with one of the remaining input feature trees until all have been taken into account. The resulting model, called an MTree, is in fact a form of a fuzzy 2D pharmacophore represented by a tree data structure where the set of nodes capture local commonalities of the input feature trees and the set of edges their general -or global- topology. The novelty of the method can be found in its ability to extract MTree models from sets of structurally diverse molecules as opposed to the SCS/MCS approaches where the input molecules must share substantially large *identical* substructures. This is due to the intermediate abstraction step provided by the conversion of molecules to feature trees. Additionally weights can be introduced so that highly similar fragments in the input set result in MTree nodes of greater importance (i.e. increased weight). This property of MTrees enables the application of similarity measures where nodes with increased weight contribute more than other nodes. This is especially useful for the main proposed application for MTree models which is that of virtual screening [7], e.g. the comparison of an MTree model with a set of compounds, one at a time, and the selection of the most similar for further processing.



TABLE I  
SUBSTRUCTURE MINING METHODS

Name	Description	Points to note
Stigmata	Structural Descriptor-based; scaffolds are descriptors found in more than a user defined percentage of compounds	Very fast; can only find substructures present in descriptor set.
<i>Frame Works</i>	Decomposes compounds in rings, linkers, side-chains; scaffolds formed by union of rings and linkers; significance assessed by frequency count	Very fast; limited, strict definition of scaffolds; interesting extensions to rings cannot be detected.
<i>PGLT</i>	Combining descriptor-based compound clustering with MCS extraction; iterative process, generating a hierarchy of scaffolds populated with compounds	Slow process, low throughput of compounds; thorough scaffold identification and related classes
<i>Distill</i>	Hierarchical, agglomerative clustering process using the size of pair-wise MCS as similarity value.	Slow process, low throughput of compounds; MCS at higher hierarchy levels not guaranteed
<i>Class Pharmer</i>	Approximates MCS's in a compound dataset using easily/quickly derived substructures, eg. frameworks or ring systems; optimization method to form classes around scaffolds	Medium-to-high throughput; dependency on the quality of the approximate MCS's and the optimization method used.
Spinifex	Combining graph-based similarity calculated using thorough MOS technique and hierarchical agglomerative clustering.	Slow, process, low throughput; thorough clustering process.
F/MTrees	Transform molecules into abstract, simpler, acyclic graphs; pairwise matching of graphs to produce model	Increased generality; maybe too coarse for some datasets
<i>FSG</i>	Apriori principle; bottom-up approach starting from one and two edge graphs, combining them to form larger frequent subgraphs; finds complete set of substructures present in dataset.	Dependency on user supplied attributes $\sigma$ (support). Depending on $\sigma$ results may miss important under-represented fragment

#### H. FSG and Other A Priori Based Methods

An alternative approach for detecting chemical scaffolds based on the Apriori principle [24] has been proposed by several research groups [25]-[30]. Apriori, originally applied for finding frequent itemsets in market-basket datasets, states that if a given itemset does not satisfy a frequency threshold then no superset itemset can satisfy the threshold either [24]. The method follows a bottom-up approach, from simpler to more complex itemsets gradually eliminating those that do not meet the threshold criteria.

A representative of these methods is Frequent SubGraph discovery (FSG), that takes as input a set of graphs  $D$  and a minimum support  $\sigma$  and finds all connected subgraphs that occur in at least  $\sigma\%$  of the graphs in  $D$  [27]. FSG's level-by-level structure, starting from a complete enumeration of simple graphs and proceeding to candidate edge graphs of

larger size is borrowed from Apriori's complete enumeration of all frequent itemsets and the addition of single items one at a time [26].

Analytically, FSG starts by enumerating small frequent subgraphs consisting of one and two edges and proceeds to find larger, candidate subgraphs by joining previously discovered smaller frequent subgraphs [26]. The size of the subgraphs is grown by adding one edge at a time. The set of candidate subgraphs is pruned by removing patterns with less than adequate frequency. FSG achieves high computational performance and is reported to analyze 200,000 compounds in one hour at  $\sigma = 1\%$ . This rate has been made feasible using sophisticated algorithms for canonical labeling of graphs to uniquely identify the various generated subgraphs without having to resort to computationally expensive graph- and subgraph-isomorphism computations [26]. To the same end, FSG employs various optimizations for the generation of candidate subgraphs and during subgraph frequency counting. One consideration with the use of FSG is the value of  $\sigma$ . A low value is more likely to capture all the important substructures present in the dataset at the expense of generating very large numbers of results. Setting the value of  $\sigma$  high reduces the number of substructures found but risks losing some of the potentially important ones.

An extension to the FSG algorithm called Frequent Subgraph-based Classification (FSC) has recently been proposed with the goal to provide graph-based classification of sets of chemical compounds [27]. One of the key ideas of the algorithm is to decouple the substructure discovery process from the classification model construction step.

FSC has three distinct steps: (i) feature generation, (ii) feature selection, and (iii) classification model construction. Feature generation takes place using the FSG algorithm. Feature selection aims to reduce the size of the substructure pool produced by the feature generation step while preserving all the information present. The selection scheme used is based on a sequential covering algorithm, an iterative approach that selects a feature and removes all compounds containing the feature at each step. The feature selected in each iteration is the one exhibiting the highest degree of accuracy [27]. Upon termination of the process the selected features are used to generate descriptor vectors for each of the compounds in the dataset  $D$ . The vectors produced are used for the construction of a classification model based on Support Vector Machines (SVM) although any of a number of classification methods can be used.

In comparative tests the FSC method has shown its superiority over descriptor-based classification methods of chemical compounds using MACCS keys [31] and Daylight fingerprints [32] for molecular descriptor generation and SVM for classification. FSC was also shown to be more accurate than algorithms based on heuristic substructure discovery. This improved performance can be attributed to the usage of FSG and thus the complete set of frequent

substructures present in a compound set [27].

#### IV. CONCLUSION

In this paper we have reviewed several substructure mining methods currently in use in the drug discovery field following closely the evolution of this research domain. Table I summarizes the key aspects of the methods described. We have also referred to several usage examples that describe successful applications of discovered substructures for the organization and interpretation of chemical data as well as for compound selection and library design. Our future work will include exploring the combinations of abstractions of molecules such as feature trees with apriori-based methods in order to achieve increased performance and generality. Additionally we will be looking into the less explored area of scaffold-based ligand design.

#### REFERENCES

- [1] J. S. Mason, A. C. Good, and E. J. Martin, "3-D Pharmacophores in Drug Discovery," *Current Pharmaceutical Design*, vol. 7, pp. 567-597, 2001.
- [2] P. D. Lyne, "Structure-based Virtual Screening: An Overview," *Drug Discovery Today*, vol. 7, no. 20, pp. 1047-1055, 2002.
- [3] C. A. Nicolaou, S. Y. Tamura, B. P. Kelley, S. I. Bassett and R. F. Nutt, "Analysis of Large Screening Data Sets via Adaptively Grown Phylogenetic-Like Trees," *J. Chem. Inf. Comput. Sci.*, vol. 42, no. 5, pp. 1069-1079, 2002.
- [4] Bioreason Inc, Santa Fe, NM, USA, <http://www.bioreason.com>
- [5] J. W. Raymond, E. J. Gardiner, and P. Willet, "RASCAL: Calculation of Graph Similarity using Maximum Common Edge Subgraphs," *The Computer Journal*, vol. 45, no. 6, pp. 631-644, 2002.
- [6] M. Stahl, H. Mauser, M. Tsui and N. R. Taylor, "A Robust Clustering Method for Chemical Structures," *J. Med. Chem.*, vol. 48, pp. 4358-4366, 2005.
- [7] G. Hessler, M. Zimmermann, H. Matter, A. Evers, T. Naumann, T. Lengauer, and M. Rarey, "Multiple-Ligand-Based Virtual Screening: Methods and Application of the MTree Approach," *J. Med. Chem.*, vol. 48, no. 21, pp. 6575-6584, 2005.
- [8] R. D. Brown, and Y. C. Martin, "The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand-Receptor Binding," *J. Chem. Inf. Comput. Sci.*, vol. 37, pp. 1-9, 1997.
- [9] J. Xu, "A New Approach to Finding Natural Chemical Classes," *J. Med. Chem.*, vol. 45, pp. 5311-5320, 2002.
- [10] S. Rodgers, J. Busch, H. Peters and E. Christ-Hazelhof, "Building a Tree of Knowledge: Analysis of Bitter Molecules," *Chemical Senses*, vol. 30, no. 7, pp. 547-557, 2005.
- [11] J. Shen, "HAD: An Automated Database Tool for Analyzing Screening Hits in Drug Discovery," *J. Chem. Inf. Comput. Sci.*, vol. 43, pp. 1668-1672, 2003.
- [12] G. W. Bemis, and M. A. Murcko, "The Properties of Known Drugs 1. Molecular Frameworks," *J. Med. Chem.*, vol. 39, pp. 2887-2893, 1996.
- [13] M. Rarey, C. Lemmen, H. Matter, "Algorithmic Engines in Virtual Screening," in *Chemoinformatics in Drug Discovery*, vol. 23, T. I. Oprea, Ed. Weinheim: Wiley-VCH Verlag, 2004, pp. 59-115.
- [14] D. Schnur, B. R. Beno, A. Good, and A. Tebben, "Approaches to Target Class Combinatorial Library Design," in *Chemoinformatics, Concepts, Methods and Tools for Drug Discovery*, vol. 275, J. Bajorath, Ed. Totowa, NJ: Humana Press Inc, 2004, pp. 355-378.
- [15] A. R. Leach, V. J. Gillet, *An Introduction to Chemoinformatics*, Dordrecht, The Netherlands: Springer, 2003.
- [16] J. W. Raymond, E. J. Gardiner, and P. Willet, "Heuristics for Similarity Searching of Chemical Graphs Using a Maximum Common Edge Subgraph Algorithm," *J. Chem. Inf. Comput. Sci.* vol. 42, pp. 305-316, 2002.
- [17] N. E. Shemetulskis, D. Weininger, C. J. Blankley, J. J. Yang., and C. Hubmlet, "Stigmata: An Algorithm To Determine Structural Commonalities in Diverse Datasets," *J. Chem. Inf. Comput. Sci.*, vol. 36, no. 4, pp. 862-871, 1996.
- [18] P. A. Bacha, H. S. Gruver, B. K. Den Hartog, S. Y. Tamura and R. F. Nutt, "Rule Extraction from a Mutagenicity Data Set Using Adaptively Grown Phylogenetic-like Trees," *J. Chem. Inf. Comput. Sci.*, vol. 42, no. 5, pp. 1104-1111, 2002.
- [19] Tripos Inc, St. Louis, MO, USA, <http://www.tripos.com>
- [20] P. Gedeck, and P. Willett, "Visual and Computational Analysis of Structure-Activity Relationships in High-Throughput Screening Data," *Curr. Op. Chem. Bio.*, vol. 5, pp. 389-395, 2001.
- [21] SageInformatics LLC, Santa Fe, NM, USA <http://www.sageinformatics.com> (acquired by Simulation Plus)
- [22] M. Rarey, J. S. Dixon, "Feature Trees: A New Molecular Similarity Measure Based on Tree Matching," *J. Comp.-Aid. Mol. Des.*, vol. 12, no. 5, pp. 471-490, 1998.
- [23] BioSolveIT GmbH, Sankt Augustin, Germany, <http://www.biosolveit.de>
- [24] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," In *Proc. of 20th Int. Conf. VLDB*, Santiago, Chile, 1994, pp. 487-499.
- [25] A. Inokuchi, A. Washio and H. Motoda, "An A Priori-Based Algorithm for Mining Frequent Substructures from Graph Data," In *Proc. 4th Eur. Conf. on Principles of Knowledge Discovery and Data Mining (PKDD'00)*, Lyon, France, Sept. 2000, pp. 13-23.
- [26] M. Kuramochi, and G. Karypis, "An Efficient Algorithm for Discovering Frequent Subgraphs," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 9, pp. 1038-1051, 2004.
- [27] M. Deshpande, M. Kuramochi, N. Wale, and G. Karypis, "Frequent Sub-structure Based Approaches for Classifying Chemical Compounds," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 8, pp. 1036-1050, 2005.
- [28] C. Borgelt, and M. R. Berthold, "Mining Molecular Fragments: Finding Relevant Substructures of Molecules," In *Proc. of Int. Conf. Data Mining (ICDM)*, Maebashi, Japan, 2002, pp. 51-58.
- [29] S. Nijssen and J. N. Kok, "A Quickstart in Frequent Structure Mining can make a Difference," In *Proc. 10th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, 2004, pp. 647-652.
- [30] J. Kazius, S. Nijssen, J. Kok, T. Back, and A. P. Ijzerman, "Substructure Mining Using Elaborate Chemical Representation," *J. Chem. Inf. Model.*, vol. 46, pp. 597-605, 2006.
- [31] MDL Information Systems, San Leandro, CA, USA. <http://www.mdli.com/>
- [32] Daylight Inc, Mission Viejo, CA, USA <http://www.daylight.com>