Analysis of Large Screening Data Sets via Adaptively Grown Phylogenetic-Like Trees

C. A. Nicolaou,* S. Y. Tamura, B. P. Kelley, S. I. Bassett, and R. F. Nutt

Bioreason, Inc., 150 Washington Avenue, Suite 303, Santa Fe, New Mexico 87501

Received November 13, 2001

As the use of high-throughput screening systems becomes more routine in the drug discovery process, there is an increasing need for fast and reliable analysis of the massive amounts of the resulting data. At the forefront of the methods used is data reduction, often assisted by cluster analysis. Activity thresholds reduce the data set under investigation to manageable sizes while clustering enables the detection of natural groups in that reduced subset, thereby revealing families of compounds that exhibit increased activity toward a specific biological target. The above process, designed to handle primarily data sets of sizes much smaller than the ones currently produced by high-throughput screening systems, has become one of the main bottlenecks of the modern drug discovery process. In addition to being fragmented and heavily dependent on human experts, it also ignores all screening information related to compounds with activity less than the threshold chosen and thus, in the best case, can only hope to discover a subset of the knowledge available in the screening data sets. To address the deficiencies of the current screening data analysis process the authors have developed a new method that analyzes thoroughly large screening data sets. In this report we describe in detail this new approach and present its main differences with the methods currently in use. Further, we analyze a well-known, publicly available data set using the proposed method. Our experimental results show that the proposed method can improve significantly both the ease of extraction and amount of knowledge discovered from screening data sets.

1. INTRODUCTION

Currently, vast compound collections are available to the pharmaceutical companies. Compound libraries have been expanding due to mergers, acquisitions, and the synthetic explosion brought about by combinatorial chemistry. High-Throughput Screening (HTS) has enabled companies to screen these larger collections of compounds and thus has helped to move the bottleneck for lead discovery from the screening arena into the analysis arena. The shear volume of hits and the emphasis on automation has brought about the creation of lead discovery algorithms, which often utilize cluster analysis techniques[1−6] and classification methods[7,8] to assist in the data analysis.

Primary screening data analysis is mainly aimed at identifying lead compounds, either as individual molecules or grouped in chemical families. The identification initially takes place by examination of each compound and evaluation of the corresponding biological information. In addition to lead identification, the analysis is expected to provide the means for the compilation of lists of compounds for subsequent rounds of screening. These lists are commonly compiled from compounds already screened during the primary screen such as selected hits and potential false negatives. In addition, or alternatively, compound lists may be generated based on compound similarity of unscreened compounds to selected hits or identified leads.

Once interesting compounds and families are found a thorough analysis is performed to identify useful Structure−Activity Relationship (SAR) information that could reveal the structural explanation for the observed biological behavior. At the end of the process all information obtained is supplied to human experts who use it to prioritize the compounds of interest and reach a decision on the promise of each hit based on various parameters. These parameters include potential for optimization, ease of chemical synthesis, drug-like characteristics, and patentability.

To fulfill these goals, various lead discovery algorithms have been recently developed and tested. A number of these methods employ data reduction to limit the data set under investigation to a small subset of the original. Then they attempt to find families of compounds with biologically favorable characteristics using computational methods such as clustering.[1,4,5,9,10] A second category of approaches attempts to build predictive models from all available screening data by use of classification techniques such as the well-known recursive partitioning (RP) technique.[7,11,12] This category of algorithms employs a more extensive automated analysis in an effort to place the molecules into groups of particular interest and to derive SAR rules related to those groups. A variation of the above method aims to construct a universal chemical classifier that can then be used to classify any screening data set into meaningful families of compounds.[8]

A third category, also common, but only applicable to data sets of limited size, is visual inspection of chemical structures and biological activity. Multidimensional visualization packages,[13] also known as visual data mining tools, enable researchers to view screening data in an interactive environment and quickly recognize and discover hits of interest in the data set under investigation. This manual approach does indeed produce satisfactory results by allowing experts to explore and identify trends and relations in small to medium

* Corresponding author phone: (505)995 8188; fax: (505)995 8186; e-mail: nicolaou@bioreason.com.

10.1021/ci010244i CCC: $22.00
Published on Web 07/17/2002

data sets. However, it is the opinion of the authors that this approach does not constitute an automated screening data analysis method and is not really applicable to extremely large and complex data sets.

Finding lead compounds in a large screening data set is a difficult task. The huge amount of information impedes thorough and timely analysis, while the presence of noise requires algorithms robust with respect to the abundance of false positives and negatives. Further, the high noise content inherent in some types of screens, especially HTS, makes the detection of false positives and negatives a distinct goal.

2. CURRENT METHODOLOGY

The problem of lead identification in screening data sets is one of discovering and evaluating information useful to the drug discovery process that is hidden in complex relations throughout the data set. Information of this type can be obtained in a number of forms, mainly by isolating chemical families that exhibit higher than average activity against the target,[14] by discovering important SAR information[11,15−17] and by detecting significant pharmacophore points.[18] This information may be used to construct and refine a model, usually referred to as the pharmacophore model, which describes a mechanism of binding to a target. During its evolution the model is used repeatedly for a variety of purposes such as selecting compounds to rescreen, directing the synthesis of new compounds, and guiding patent searches.

Identification of chemical families can be performed in a two-dimensional (2D) or three-dimensional (3D) space.[19−21] 2D families are obtained by grouping compounds with similar 2D structure. Similarly, 3D families contain compounds that have similarities in their natural form, the 3D space. Commonly, chemical families are defined by a contiguous structure, the scaffold, or by a Markush structure[22] common to all the compounds in the family. Alternatively, high similarity of 2D or 3D molecular fragments may be considered satisfactory. Chemical families that exhibit favorable biological and/or chemicophysical characteristics are often chosen as candidates for follow-up studies. Further reduction of these candidates can be based on other characteristics such as structural novelty or the ability to easily synthesize related compounds.

SAR information extraction refers to the process of correlating structural features to measured activity response. Of particular interest is SAR information related to structural features responsible for markedly increased or decreased potency. It is worth noting that while potency-increasing features are mostly obtained from observations on the active compounds, potency decreasing features are discovered with the analysis of both active and inactive compounds. The isolation of such potency increasing or decreasing features can reveal important pharmacophore points necessary for activity against a target as well as other properties that the pharmacophore model must have.

Traditionally, scientists in this area have used a variety of approaches and algorithms to analyze chemical data. A necessary prerequisite to most approaches has been the representation of chemical structures in ways that facilitate algorithmic processing.[19−21] Most often chemical compound representations take the form of binary strings, e.g. MACCS keys,[23] Daylight fingerprints,[24] or BCI keys.[25] Such representations are suited for metric or measure comparisons, e.g. the Euclidean distance or the Tanimoto coefficient.[19] With the ability to represent and compare chemical structures using such measures, analytical computational algorithms specifically designed and implemented to tackle problems found in the drug discovery industry have flourished.

Among the most popular analytical methods are clustering algorithms. Recent work by Brown and Martin[14] focused on the use of hierarchical agglomerative clustering algorithms to group screening data sets into families. Wild and Blankley[26] and Engels et al.[4] employed the same type of algorithms in association with various chemical representation schemata, cluster level selection methods, and compound sampling techniques. Wagener et al.[27] have used a neural network clustering method, the self-organizing map (SOM), to the same end. These methods have been shown to be, in varying degrees, at least reasonable at grouping chemically active molecules together.

Among hierarchical agglomerative clustering algorithms—and possibly clustering methods in general—the Wards method[4,14,26,28,29] is most commonly used by the chemoinformatics community. Often the cluster hierarchy produced by the application of Wards is processed by the Kelley cluster level selection method[26] used to determine which level of the hierarchy contains the "natural" clusters for the data set. Recent publications have defined and investigated the cause and effect of the ties in proximity problem[30−32] which could cause a dependency of the cluster hierarchy produced to the order of the patterns supplied. However, a fix to hierarchical agglomerative clustering algorithms that resolves the ties in proximity problem has since been proposed.[33]

Despite the obvious advantages of clustering methods, a complete methodology including the precise clustering algorithm, distance or similarity measure, and type of representation, has not been defined with respect to data dependencies. Further, clustering algorithms have limitations regarding the number of patterns they can cluster in reasonable amounts of time. Clustering an entire HTS data set using any of the widely used algorithms takes days if not weeks even on today's powerful computers. To alleviate this problem a preprocessing step is often employed to partition the results into hits and nonhits based on the biological activity of the compounds. Thus, all subsequent screening data analysis can focus on the compounds in the active list, while the inactive compounds, which are the majority of the data set and may contain valuable SAR information, are completely ignored.

Classification methods have also attracted a substantial level of interest among the drug discovery community. Rusinko et al.[7] have used the Recursive Partitioning (RP) classification method to discover features of compounds in the screening data set that discriminate active from inactive compounds and formalize rules useful for SAR and pharmacophore point detection. According to the method, a set of descriptors is first provided, each indicating a structural feature that can be described as present or absent in a given molecule. RP is then used to divide the molecules of the entire data set into exactly two groups according to whether the molecules have a particular "best" descriptor in common. The "best" descriptor is the descriptor that would result in the largest possible difference in average potency between

those molecules containing the descriptor and those molecules not containing it. The method continues iteratively with respect to each subdivided group, dividing each group into two groups based on a next "best" descriptor selected from the group of descriptors. The result of this process is a tree structure in which some terminal nodes contain a majority of active molecules, while others contain a majority of inactive ones. Tracing the lineage of the structures defined by a terminal node can reveal molecular descriptors that may be related to increased or decreased potency. Molecules not yet empirically tested can be filtered through the tree structure generated by RP. A prediction for the activity of each filtered molecule can be made by simple examination of the characteristics of the tree nodes it was placed.

However, the use of RP to partition molecules on the basis of their structural and activity similarity can be limiting. When the sizes of the active and inactive compound sets are significantly unbalanced, as is often the case with screening data sets, building a classifier and hence a predictive model can be considerably difficult. A small amount of noise in the data will often prevent the descriptors from discriminating the very small class from the much larger class. In general, the noise level in screening data sets and especially HTS sets may be substantial as well as the numbers of false positive and negative responses. Thus, the use of RP on these data sets may be problematic because of faulty splitting decisions on the RP tree.

Furthermore, RP as well as all the common clustering methods used by the pharmaceutical industry are partitional. Consequently, if there is more than one set of descriptors, e.g. structural features in a molecule that are related to observed activity, RP and common clustering methods will be unable to identify all of them. MacCuish et al.[32] have explored extensively the multidomain nature of chemical compounds which is the cause of the above problem. For the purposes of this research a domain is defined to be a set of pattern features or descriptors; thus, multidomain molecules are those molecules that contain structural domains characteristic of multiple chemical families.[33]

A different classification approach to HTS data analysis, suggested by Roberts et al.,[8] attempts to classify compounds into a set of predefined chemical families defined by 2D molecular fragments. They describe a software package that reduces screening data analysis to the simpler task of classifying molecules into a vast collection of predefined chemical families. Again, the families are structured hierarchically, and potency enhancing and decreasing features can be extracted by simple inspection of the families. This method, characterized by its simple, exhaustive search approach, relies heavily on the predefined list of families and the molecular fragments that define them. The reliance of the method on that predefined list will impede the discovery of novel, unexpected pieces of knowledge. As a consequence, potency related structural features are going to be identified only if they happen to be significant enough to be contained in the molecular fragments that define chemical families. Small structural variations to a molecular fragment that result in a significant jump in activity may not be detected. In addition, new, small, unexpected, but pharmacologically interesting chemical families also risk not being identified using this method.

3. THE PHYLOGENETIC-LIKE TREE GROWING ALGORITHM

3.1. Goals of the Algorithm. The phylogenetic-like tree (PGLT) algorithm is a method for analyzing a data set of molecules to assist in identifying chemical classes of interest and sets of molecular features that correlate with a specified biological feature. The outcome of the algorithm is a PGLT data structure where each node contains a chemical fragment common to all molecules in the node, two sets of molecules corresponding to the active and inactive compounds of the node and links to its parent and children nodes. Note that the root node is the only node in the tree that does not have a chemical fragment characteristic of the molecules it contains and is not linked to its parent. Similarly, the leaf nodes are not linked to any children. Additional properties of the node and attributes of the molecules contained in it can be computed or supplied and stored upon request.

The tree structure resulting from the analysis of each data set is essentially a hierarchy of the chemical classes found in the active compounds in the set, where each class is represented by a node of the tree. In their final form the nodes contain both the active compounds that were used to define the class and the inactive compounds that were found to be members of it. Nodes close to the top of the tree tend to be larger, more general, and broadly defined, while their descendants, nodes closer to the bottom of the tree tend to be smaller, more precisely defined, and more homogeneous.

In its current implementation the algorithm uses as input both chemical structure and biological information about molecules in the data set. During a preprocessing step a descriptor of the chemical structure of each molecule in the data set is computed and associated with the biological information available for that molecule. By default, the set of chemical keys used by the algorithm to compute the descriptor of each molecule is a set of MACCS-like keys, a variation of the publicly available MACCS[23] defined by MDL. It is worth noting that a user may choose to supply and use any set of chemical keys. Since the chemical families detected by PGLT are based on 2D scaffolds, we prefer using 2D chemical key sets. Furthermore, 2D keys are faster to compute and according to a thorough report by Brown and Martin[14] they compare favorably to 3D keys when used for grouping compounds. As a result of this step when the PGLT algorithm commences each molecule in the data set under investigation is defined by both a chemical descriptor and a biological information value.

3.2. The Algorithm. The PGLT algorithm is of a hybrid nature employing various techniques ranging from neural networks and genetic algorithms to expert rules and chemical substructure searching. It is of a repetitive nature with six reoccurring steps in the main part of the algorithm and an initializing and a postprocessing step (see Figure 1). The entire algorithm is implemented by an integrated software package (see ref 34).

The process is initiated by the construction of the root node of the PGLT structure. This step also populates the corresponding list with all the active molecules in the data set under investigation. The root node is then supplied to the central, repetitive part of the algorithm. This part consists of the following steps:

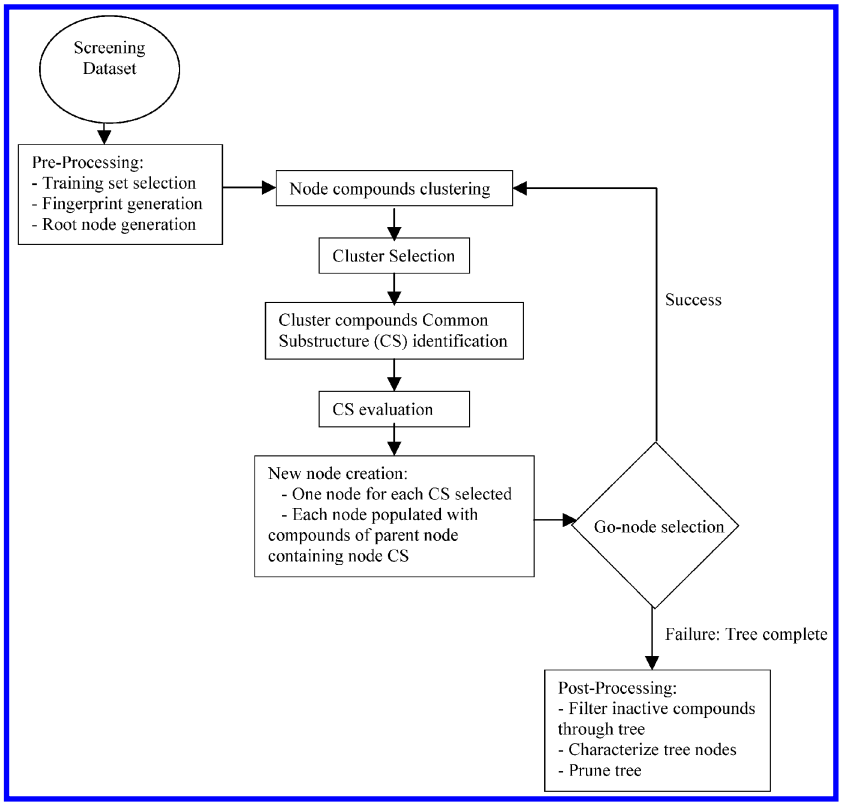

Figure 1. Flowchart of the PGLT algorithm including pre- and postprocessing steps.

a. Clustering: a clustering algorithm is used to group the molecules in the node based on the similarity of their chemical descriptors. Currently, the default clustering algorithm is a neural network based method, the Self-Organizing Map (SOM),[27,35] but other clustering algorithms can be used as well. In addition to SOM the software package implementing the algorithm[34] supports the use of Wards, Group-Average, and other hierarchical agglomerative clustering algorithms.

b. Clusters selection: the clustering results are processed and a set of "natural" clusters is selected. In this context, natural clusters are defined to be those clusters that represent sets of compounds characterized by high similarity. Depending on the clustering method used, a suitable "natural" cluster selection method should be applied. For example, if Wards clustering is used, a cluster level selection method, e.g. the Kelley method,[26] should be available. The need for cluster selection is a result of known shortcomings of clustering approaches when used in combination with molecular fingerprints (see ref 26).

c. Common Substructure Identification: For each natural cluster a common substructure capturing a potentially new similarity axis among the compounds of the cluster is learned. In the implementation used the common substructure is by default the Maximum Common Substructure (MCS) of the compounds in the cluster.[36] Alternatively, the Significant Common Substructure (SCS) could be used. To extract the SCS of a group of compounds all common substructures are found. Then, a set of individual atom weights could be used to compute the cumulative weight of each of the common substructures. The SCS would be defined to be the common substructure with the largest "weight" value.

d. Common Substructure Evaluation: A set of rules defined by expert human analysts, commonly referred to as expert rules, is used to evaluate each of the substructures and to eliminate all that do not constitute a significant gain in new knowledge. For example, a sample rule used for evaluation eliminates common substructures that have been discovered previously and exist as substructures of a node somewhere in the tree. Another rule eliminates substructures that are identical or subsets of the common substructure of the parent node.

e. New node creation: A set of new nodes is constructed, one for each newly found common substructure that avoided elimination in the previous step. For each node all molecules in the active compound list of the parent node are queried with its substructure. The molecules that contain that substructure form a new set placed in the new node as its active compound list. When all new nodes have been created, they are appended to the tree by constructing links to and from their parent.

f. Go-Node selection: All the leaf nodes of the PGLT data structure are examined and one of them is selected as a go-node. The selection of the go-node is based on the tree-growth method selected by the user and used for a specific analysis. The available tree growth methods include breadth-first, which grows the tree level by level and diversity-first, which chooses to follow the most diverse of the leaf nodes suitable for growing.

The process then iterates and performs the six steps defined above using the go-node as input. The process terminates when a predefined criterion has been met, e.g. the remaining leaf nodes contain too few compounds or have a low molecular diversity coefficient, or the tree has grown to a preset depth level.

When the PGLT is complete, a postprocessing step populates the inactive set of the root node with the inactive compounds of the data set under investigation. It then filters

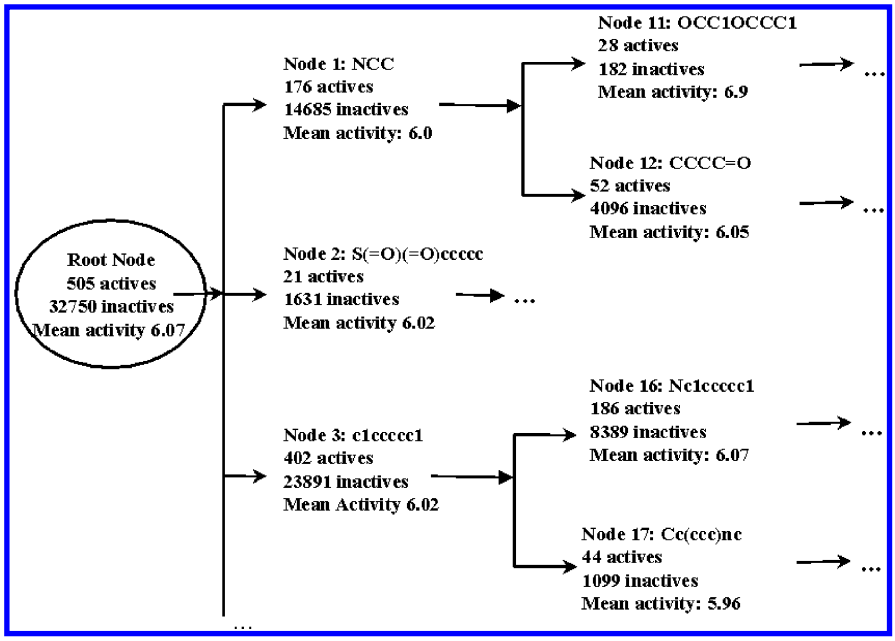

Figure 2. A fragment of a PGLT generated by processing an AIDS antiviral data set publicly available by NCI. Nodes of the tree fragment are shown to contain the substructure defining the node (in Smiles language notation), the number of active and inactive compounds, and the mean activity of the node. A typical PGLT node may contain additional properties. Note that children nodes contain their parent's substructure in addition to their own. Also note the mean activity change between pairs of parent-child nodes (e.g. nodes 1 and 11) likely to be explained by the difference in their defining structures.

those inactive compounds through the rest of the tree until it reaches the leaves. Filtering takes place by querying the inactive compounds of a parent with the substructure of a child node and placing all the matches in the inactive set of the child.

Further, the postprocessing step employs a hybrid system, making use of statistical methods and expert rules, described in more detail below, to extract knowledge from the PGLT data structure in an automated fashion. More analytically, the primary goal of this hybrid postprocessing system is to find which nodes of the complete PGLT represent structurally homogeneous chemical families. To achieve this goal the system initially employs statistical methods such as molecular similarity on the compounds of each node to assess whether they have sufficient structural similarities to represent a chemical family. The molecular similarity methods used can be based on the vector representations of the compounds[19] and on the ratio of the atom length of the MCS extracted from those compounds to the atom length of the compounds. In addition, a series of simple expert rules defined by humans is applied to further characterize the nodes. These rules are mostly related to the number of compounds contained in a node and the composition of those compound sets. The results obtained by the application of the statistical methods and the expert rules are then combined to derive conclusions on whether each node represents a structurally homogeneous chemical family. Nodes that are labeled chemical families form the basis of the analysis results to be processed further and eventually presented to the user. All other nodes are removed from the PGLT which subsequently contains only the set of potentially interesting results.

Once the PGLT has been pruned to include only the nodes representing structurally homogeneous families each node is further processed and populated with statistical values related to the activity or other biological attributes of the compounds in the node. These values include the mean activity, the standard deviation, and the percentage of actives in the node. Further, lineages of nodes of the PGLT are examined, and the relations among node attributes, such as average activity or percentage of active to inactive compounds, are evaluated to determine which nodes may contain SAR information. Note that the biological activity of compounds is not used to define nodes but is used to describe nodes discovered based on chemical structure alone. Figure 2 shows a part of a PGLT generated by processing an AIDS antiviral data set publicly available by NCI.[37]

3.3. Characteristics of the Algorithm. The PGLT algorithm performs an adaptive, data driven organization and analysis of large screening data sets. The entire process is not limited by dependencies on predefined chemical families, chemical descriptors, or specific computational methods. This enables the implementation to explore and learn from each different data set novel, often unique, pieces of information. Further, the algorithm considers the entire data set during the analysis. Active compounds are used to define the hierarchy of chemical classes and map the chemical space of interest for the specific assay; inactives are used to populate the classes and refine the knowledge extracted, especially potential SAR information.

An additional characteristic is related to the nature of the structurally homogeneous chemical families of compounds detected and used throughout by the algorithm. The definition of these families ensures that they share a substantial common substructure in addition to a high molecular similarity value. This characteristic has been designed to facilitate the interpretation of classes and to enable the correlation of significant substructures to observed biological properties such as activity. Note that the substructure common to a set of compounds contained in a node needs not be contiguous (see Figure 2, nodes 1 and 11). The nature of the algorithm enables the detection of an MCS at a child node that may be distinct from the MCS of its parent node.

Thus, the algorithm can detect a chemical family defined by two distinct structural fragments linked by a variety of substructures.

Finally, the algorithm accommodates the multidomain nature of chemical compounds. Compounds are not simply partitioned into a number of classes. Rather, the algorithm tries to detect all structural features with significant presence in the data set, constructs classes defined by those features, and populates the classes with all compounds containing them. This property of the algorithm aims at discovering under-represented chemical classes that might have been lost if some of their compounds were placed in larger, more prevalent classes. Another goal is to better populate all the classes detected.

4. EMPIRICAL STUDIES

4.1. Methodology. The set of experiments performed for this research focused on the analysis of a large chemical data set using the proposed algorithm and the presentation of the results obtained. Emphasis was placed on the quality of the chemical families detected, the SAR rules extracted, and the validation of the findings against known facts about the data set under investigation. For the purposes of this section we employed an enhanced version of the LeadPharmer suit of tools[34] to analyze the NCI-AIDS data set publicly available from the National Cancer Institute[37] with the PGLT algorithm. Further we used the DrugPharmer[34] decision support system to extract knowledge in the form of interesting chemical families and SAR rules.

The compounds in the data set were represented by BR-MACCS keys a set of structure based keys. BR-MACCS keys are a slightly modified version of the public MACCS keys.[23] The number of keys has been reduced from 166 to 157, and a small percentage of the original keys were modified. Molecular vector comparisons were performed using the Euclidean metric.

4.2. Data Set Description. The National Cancer Institute's HIV antiviral screen utilized a soluble formazan assay to measure the protection of human CEM cells from HIV-1 infection.[37,38] The results of the activities of the compounds tested in the assay were reported as EC50, the concentration required to produce 50% protection. For the purposes of this research all compounds with an activity of $<3.2 \times 10^{-6}$ M were considered active. In addition, the activity values were converted to -log EC_{50} for ease of SAR interpretation. In our analysis with the PGLT algorithm 505 active compounds that had a molecular weight of 500 or less were defined to be the training set. Thirty-two thousand seven hundred fifty compounds of lower activity were subsequently used to better populate the chemical classes defined by the training set.

4.3. Application of PGLT Algorithm to HTS Data Analysis. Our empirical studies were devoted to the analysis of the NCI-AIDS data set with the PGLT algorithm. The evaluation was performed by medicinal chemists and focused on the scientific validation of the results produced. To this end, we analyzed the results of the PGLT, evaluated the quality of the classes detected, and discussed SAR rules.

Chemical Families Detected. A total of 104 structurally homogeneous chemical classes containing 410 compounds were detected from the analysis of the 505 compounds in the training set. This relatively high number of classes, a typical characteristic of the PGLT algorithm, aims at achieving sufficient coverage of the classes present in the training set. The average number of compounds contained in each of the 104 classes was 15.4, while each of the 410 compounds was present in three classes on average. The 95 remaining compounds could not be grouped by the PGLT algorithm with any of the 104 classes and were thus presented as structural outliers or singleton classes. After filtering the test set of 32 750 low activity compounds 9359, or 28.6%, were placed in the identified classes. The vast majority of the 9359 compounds filtered in a very small number of classes.

As expected a range of known chemical families were featured prominently in the results' list (see Table 1). Among others, the PGLT algorithm discovered a well-known large class consisting of a collection of 46 nucleosides, including pyrimidine, dihydropyrimidine, and purine nucleosides (Figure 3). Since the scaffold describing this class contained only the sugar substructure, substitutions on the sugar with different types of heterocycles were allowed. The most active members of this class contained a pyrimidine nucleoside with a 3′-azido or 3′-fluoro moiety. The dihydropyrimidine nucleosides were less potent. The utility of comparing the filtered, inactive compounds to the active data set is illustrated in the finding of additional SAR within the nucleoside class as well as the two smaller classes described below. Many of the inactive compounds of lower activity that were contained in this class did not include a heterocycle attached to a sugar. Of the inactive compounds that did include a heterocycle many contained either sugars with a 3′-hydroxyl group or sugar mimics that contained 1,3-dioxolane ring.

Table 1. Class Attributes of Three Representative Classes

no. of actives[a]	learned substructures[b]	av activity[c]	SD[d]	no. of filtered[e]	% actives[f]
46	OCC(C)OCC	6.83	0.96	1354	3.3
7	Ccccccc; N1CSC(C1=O)=C	5.74	0.19	111	5.9
3	CCCn1c2c(nc1=S)cccc2	7.3	1.35	15	16.7

[a] Number of active compounds of class. [b] Maximum common substructure learned from all the active compounds in the class, in Smarts notation. [c] Average activity of the active compound in the class. [d] Standard deviation of the activity values of the active compounds in the class. [e] Number of inactive compounds filtered in the class, e.g. containing the learned substructure of the class. [f] Ratio of active compounds in the class to the total number of compounds (active + filtered) of the class.

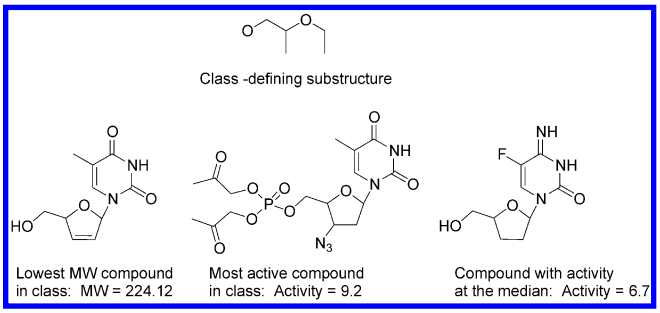

Figure 3. The nucleoside class defined by PGLT. This class consisted of 46 compounds containing the substructure shown above. Three of the 46 compounds are depicted: namely the lowest with respect to molecular weight (MW), the most active, and one with median activity.

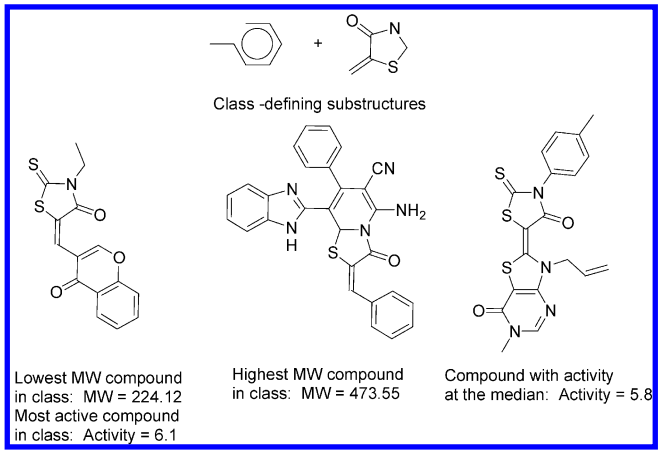

Figure 4. The 2-thioxo-thiazolidin-4-one class defined by PGLT. The class consisted of seven compounds containing the two noncontiguous substructures shown above. Three of the seven active compounds are depicted: the lowest with respect to molecular weight (MW) which is also the most active, the highest MW compound, and one with median activity.

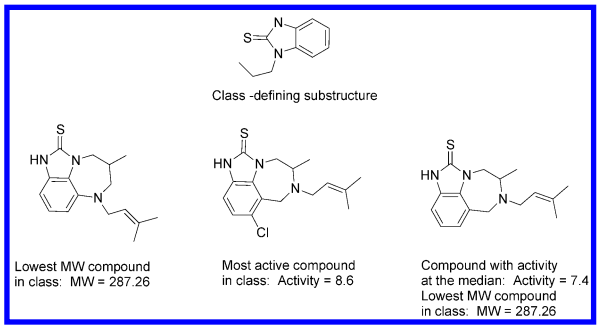

Figure 5. The thioxobenzimidazole class defined by PGLT. The class was defined by three active compounds containing the substructure shown above. All three of the active compounds are depicted. In an HTS setting, a smaller class such as this one is difficult to detect within the massive amounts of data produced.

Furthermore a number of small, under-represented in the data set classes were found. As an example the PGLT algorithm discovered a class of seven members described by two noncontiguous keys, namely a 2-thioxo-thiazolidin-4-one and an aryl methane key (Figure 4). This acceptable class contained compounds that the human eye may have difficulty combining. Within this class the 2-thioxo functionality is important for activity: several inactive compounds contained an oxo or imino functionality. An aromatic group was a preferred substituent on the nitrogen atom of the thiazolidinone, while a diamido-methylene substituent was a nitrogen substituent of many of the inactive compounds. A cyclic 2-substituted thiazolidine was permitted for activity. A cyclic structure from the exo-methylene onto itself led to active structures if it was a dihydrothiazole or heterocycle-fused dihydrothiazole, but inactive compounds contained a dihydroindolone cyclic substructure. In addition, a smaller class of three closely related members was described by a thioxobenzimidazole (Figure 5). This class contained inactive compounds without a propenyl side chain. The most inactive compounds also had larger and smaller side chains. A seven-member ring fused to the 6,5-ring system was preferred: inactive compounds contained an eight-member ring.

Multidomain Nature of PGLT. Multidomain compounds, taking advantage of the nature of PGLT, found their way into multiple clusters. This freedom of placement of multi-

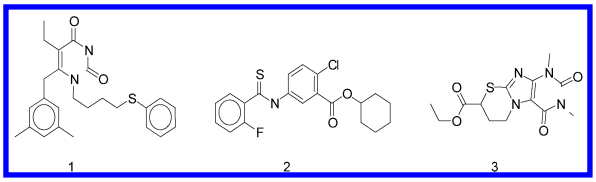

Figure 6. Compounds that illustrate the multidomain nature of the classification method.

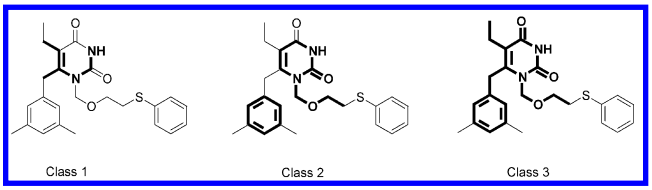

Figure 7. Substructure definition of classes to which compound **1** belongs. The bold part of each figure represents the substructure defining the class.

Table 2. Multidomain Compounds and the Attributes of Their Associated Classes

	class ID	no. of actives	no. of filtered	av activity	SD	% actives
compound **1**	1	32	416	7	1.08	7.14
	2	6	74	6.13	0.65	7.5
	3	5	0	6.14	0.72	100
compound **2**	4	12	52	6.98	0.55	18.75
	5	27	51	6.87	0.59	34.62
	6	30	2	6.87	0.55	93.75
compound **3**	7	34	477	6.92	0.99	6.65
	8	7	150	5.77	0.14	4.46

domain compounds, an essential feature of the PGLT analysis, had a significant impact in a number of ways. First, a number of small under-represented classes that could not be discovered by traditional, partitional algorithms were revealed.[33] Second, the allowed overlap resulted in an increased total number of compounds in a number of classes, and thus better representation of the chemical families described by those classes. Finally, it was advantageous to learn many possible classifications for a given compound in order to evaluate the effect of the learned structural elements on activity, either as a function of the active compounds (average activity of a class) or as a function of the total data set (percent of actives to the total number of compounds in a class). Each classification can be explored for relevance of the substructure keys to the biological activity observed in a specific assay. Three compounds were selected to illustrate the multidomain nature of the classification (Figure 6, Table 2).

Compound 1 was classified in three ways as shown in Figure 7. Class 1 was defined as a heteroaromatic. The average activity for this class was higher than the other two classifications (see Table 2). Class 2 was represented by fewer members of lower average activity. Class 3, a subclass of Class 2, had a similar membership and average activity compared to the parent class but distinguished the actives more accurately from the inactives. The key difference in the definition of Classes 2 and 3 was the larger contiguous substructure of Class 3. Class 1 represents a compound set with a high average activity, whereas the substructures of Class 3 best distinguished the actives relative to the inactives.

Compound 2 was contained in classes of similar average activity (Figure 8). However, there was a wide range of percent actives that distinguished these classes from each

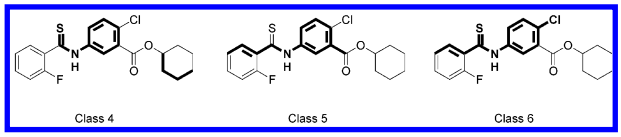

Figure 8. Substructure definition of classes to which compound **2** belongs.

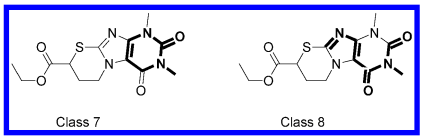

Figure 9. Substructure definition of classes to which compound **3** belongs.

other. Class 4, the class with the lowest percent active value, does not contain a second aryl system in its defining substructure. Classes 5 and 6, defined by a second aryl system, have a higher percent actives attribute. The second aryl ring must be an important feature for activity. Class 6, further defined as a thioamide, has a very high percent active value; this substructure was a feature that effectively distinguished the active compounds.

Compound 3 was classified in two ways (Figure 9). Class 7 not only has a higher average activity but also had a higher percent actives value. The pyrimidinone of Class 7 was a better descriptor for this compound than the extended aromatic system of Class 8.

Further SAR Rule Extraction. In addition to the identification of SAR information via the multidomain property of the PGLT classification shown above, potency related features were detected by use of other properties of the resulting tree structure. Of high value is the ability to detect SAR information related to potency decreasing substructures in addition to information related to potency increasing features. This is feasible since the algorithm proceeds to populate the chemical families detected using the active subset of the compounds with all inactive compounds in the data set that share the MCS associated with that family. In comparison, most current methods do not use the inactive part of a data set usually containing more than 98% of the total compounds screened. Thus, they neglect the majority of information produced by primary screening processes including all potency reducing SAR information present.

SAR across subclasses, e.g. children classes, may be discovered due to the greater structural definition that they tend to have over their corresponding parent classes. In the study of the anti-HIV data set, the increased substructure definition along with changes in activity has enabled the discovery of SAR within a class. As an example, the general class of nucleosides was considered. Figure 10 depicts substructures that lead to subclasses with increase, decrease, or insignificant change in average activity as shown in Table 3. Class 9, the parent class of all other classes in Figure 10, contained hydroxy-ether compounds with a variety of substitutions. This class also contained a large number of inactive compounds, including those that were not nucleosides. All of the smaller, better defined subclasses contained nucleoside substructures and therefore had a much higher percent actives value. Related classes 10 and 11 displayed a higher activity, with the additional 3′-fluoro substitution of class 10 giving rise to the most highly active compound set.

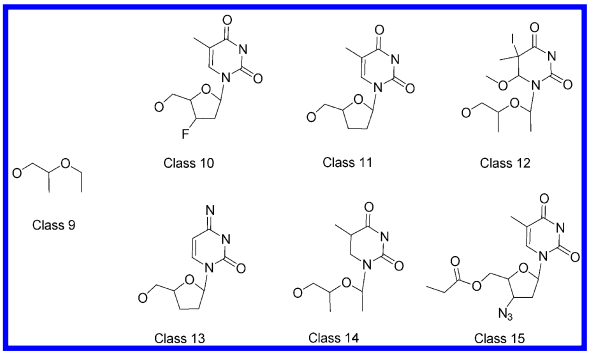

Figure 10. SAR derivation from class/subclass relationships: substructure definition of classes.

The pyrimidine nucleoside was found to be a potency enhancing substructure. Classes 12, 13, and 14 displayed lower average activity and thus are defined by potency lowering substructures. The iodo-dihydropyrimidine class 12 contained the least potent derivatives. Class 15 displayed the same average activity as the parent class 9. Although varying widely in average activity, classes 10, 12, and 15 were very well defined with respect to the active set, as exemplified by their respective percent actives values of 78, 83, and 81.

SAR can also be extracted directly from the nodes of the PGLT data structure. The change in substructure definition along with the concomitant change in average activity was used to find SAR from parent to child nodes. This relationship is found in the example parent class 16 and its children, classes 17−19 (Table 4). The chlorophenylthioamide class 16 was divided in three ways to give smaller classes with different average activities. Class 17, defined by a substructure containing a furan and ether, exhibited both a higher average activity and a higher percent actives value, features that indicate the substructure is potency enhancing. Alternatively, class 19 contained a potency decreasing oxime substructure when observing the activity trends within the active set. The inactive compounds in this class were also closely related oximes.

4.4. Discussion. A data set analysis with the PGLT algorithm produces clusters of similar compounds providing information such as the maximum common substructure and the homogeneity/diversity of the compounds in each cluster. A set of statistical and expert rules is then used to process each cluster's properties and reason about the level of medicinal interest of each cluster. The obtained interest—or potential—level of a cluster is then used to eliminate clusters with low information content and include in further analysis steps those that are likely to contain information that may lead to the identification of lead compounds or SAR information. A unique feature of the PGLT algorithm, namely the accommodation of the multidomain nature of compounds, ensures that multidomain compounds are classified in many appropriate ways. This feature enables the discovery of small, under-represented classes and allows all classes found by the algorithm to be populated with the appropriate subset of compounds present in the screening set. The final step in the PGLT analysis, that of filtering inactive compounds through the tree structure, assists the extraction of SAR information related to potency decreasing features; it is also useful in better defining and characterizing the chemical families present in the data set. In general, the NCI-AIDS data set analysis with the PGLT algorithm produced results

Table 3. SAR Derived from Class/Subclass Relationships

class ID	no. of active	av activity	SD	no. of inactive	activity change	% actives	common substructure
9	46	6.83	0.96	1354		3.3	OCC(C)OCC
10	7	7.69	1.20	2	0.86	78	Cc1c(nc(n(c1)C2OC(C(C2)F)CO)═O)═O
11	22	7.24	1.06	112	0.41	16	OCC1OC(CC1)n2c(═O)nc(c(c2)C)═O
12	5	6.26	0.27	1	−0.57	83	COC1C(C(NC(N1C(C)OC(CO)C)═O)═O)(C)I
13	4	6.52	0.24	20	−0.31	17	OCC1OC(CC1)n2c(nc(cc2)═N)═O
14	11	6.63	0.82	27	−0.20	29	C1C(C(═O)NC(N1C(OC(CO)C)C)═O)C
15	13	6.83	0.79	3	0.00	81	CCC(═O)OCC1C(CC(O1)n2c(nc(c(c2)C)═O)═O)N═N═N

Table 4. Activity along A Path: Method To Discover SAR within the Classification

class ID	no. of active	av activity	SD	no. of inactive	activity change	% actives	common substructure
16	29	6.87	0.58	3		91	Clc1ccc(NC═S)cc1C
17	11	7.12	0.63	0	0.25	100	Clc1ccc(NC(c2c(C)occ2)═S)cc1COC(C)(C)C
18	12	6.86	0.52	0	−0.01	100	Clc1ccc(NC(ccc)═S)cc1C(OC(C)C)═O
19	6	6.73	0.28	2	−0.14	75	Clc1ccc(NC(ccc)═S)cc1C═NOC(C)C

that effectively covered the entire pharmacologically interesting portion of the data set and had a high level of information content.

When compared to methods currently used for screening data analysis [see section 2] the PGLT approach has several distinguishing features. Unlike PGLT, analysis via clustering algorithms is limited to identifying clusters of similar compounds but not common substructures that would justify the grouping and facilitate discovery of SAR. Further, the quality of the clusters may be questionable since clustering algorithms are used in combination with molecular descriptors with known limitations,[30−32] whereas in PGLT each class must satisfy certain criteria such as the presence of a significant MCS and a high level of homogeneity. Classification algorithms depend heavily on the quality of the attribute measurement used for placing compounds into classes. In the case of large screening data sets, such as those produced by HTS systems, activity measurements are often noisy. Methods that exclusively use those measurements to guide discovery ignore that fact and assume lack—or minimal presence—of noise. In contrast, PGLT solely uses chemical structure to define its classes and only overlays biological measurements to characterize the classes. In addition, screening data analysis algorithms based on classification algorithms often characterize the detected classes with the descriptor(s) used to form the class.[7,11] While that is an improvement over clustering approaches it is limited by the initial choice of descriptors used to characterize the molecules and falls short from the PGLT approach of adaptively learning from each class the MCS of its compounds. An additional distinguishing feature is that PGLT accommodates the multidomain nature of compounds while classification and clustering approaches are mostly partitional and thus force compounds to be placed in a single cluster/class. Finally, global classifiers, such as the one proposed by Roberts et al.,[8] while sharing the same goals as PGLT use a different approach to achieve them. Where PGLT adaptively constructs the class hierarchy by grouping the compounds in the data set during each analysis this type of classifier uses a predefined class hierarchy to classify compounds. The resulting data structures while similar in form tend to be very different both in size and content; global classifiers present the user with a greater set of classes containing compounds that simply share the 2D molecular fragment defining the classes. This makes global classifiers an efficient exploratory tool capable of relating screening data results to known classes of pharmacological interest. PGLT produces a more concise set of classes characterized by high homogeneity of their compounds and by MCSs that reflect the commonalities of the compounds in the classes to the highest possible degree. This approach enables PGLT to find chemical classes present in the data set be they previously known or novel. As a result, the PGLT method is not only able to relate biological results to familiar chemical classes but also discover unexpected, interesting classes and 2D substructures that reveal potency related features specific to each analysis.

5. CONCLUSIONS

The PGLT algorithm is a method for adaptively learning what substructure(s) are responsible for classifications of molecules into families and for relating those substructures to the attributes of the molecules they describe. This learning method operates by grouping a set of molecules according to their molecular structure as characterized by a set of descriptors, identifying the groups that exhibit some desirable attribute (such as high homogeneity), and analyzing those groups to identify the maximum common substructure(s) among their molecules. The learned substructures then serve as a filter through which the molecules are passed. For each filter, a new "child" node based on the filter is formed, containing those molecules that include the corresponding substructure. By iteratively continuing this process of identifying substructures and using the substructures as filters to further group the molecules, the method constructs a multidomain tree structure where each node represents a collection of molecules containing the node substructure. Construction of the tree structure is completed when some predefined criteria are met; e.g. the homogeneity of all leaf nodes is above a threshold. The completed tree structure is then populated by the molecules in a test set, commonly the inactive part of the data set under investigation by a process of iteratively matching the molecules of a parent node with the substructures of each of its children. Further, the tree is pruned during a postprocessing step aimed at removing nodes that do not constitute a significant gain in novel knowledge.

The resulting multidomain tree structure enables hypotheses and inferences for the correlation between the substruc-

tures related to nodes and the characteristics of the compounds in the node. Thus, the proposed method not only determines how well a particular substructure defines a chemical family but also elucidates the relations between the node's substructures and the measured biological properties of the node's molecules. In addition, the nodes of the multidomain tree can be used as input to other techniques, including 3D molecular modeling, for further analysis. Applying those techniques to many small structurally homogeneous data sets, such as the nodes of the PGLT tree, is likely to produce high quality results faster than applying the same techniques on a larger, more diverse set such as the whole screening data set or the entire subset of the active compounds.

The empirical results presented in this research support the utility of the PGLT method in the task of analysis of screening data sets. A discussion of the results obtained dealt with the successful use of the maximum common substructure approach to capture and assess the relations of groups of molecules and further elaborated on the quality of the chemical families defined with the PGLT method. The experiments have also shown that the method is capable of forming structural families from a given data set and extract detailed SAR information in a convenient way. In addition, partly due to the multidomain nature of the tree structure produced, the method is robust to under- and over-represented structural families of compounds, a property highly desirable but not found in many of the methods already in use for screening data analysis.

Overall, the application of the PGLT method leads to a thorough identification of structural families and the substructure scaffolds that characterize them as well as to the extraction of SAR rules and related information. The PGLT method achieves those goals in a highly automated, robust, and easily interpretable manner. Key to this outcome is the exploitation and accommodation of chemical data multidimensionality and the use of an array of methods including chemical rule-based systems to guide the learning process from screening data sets.

ACKNOWLEDGMENT

The authors would like to thank Drs. Terence K. Brunck and Patricia Bacha of Bioreason, Inc. for insightful discussions, comments, and suggestions.

REFERENCES AND NOTES

(1) Barnard, J. M.; Downs, G. M. Clustering of chemical structures on the basis of two-dimensional similarity measures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644−649.

(2) Willett, P.; Winterman, V.; Bawden, D. Implementation of non−hierarchic cluster analysis methods in chemical information systems: Selection of compounds for biological testing and clustering of substructure search output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109−118.

(3) Shemetulskis, N. E.; Dunbar, J. B.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. H. Enhancing the diversity of a corporate database using chemical database clustering and analysis. *J. Comput-Aided Mol. Des.* **1995**, *9*, 407−416.

(4) Engels, M. F. M.; Thielemans, T.; Verbinnen, D.; Tollenaere, J. P.; Verbeeck, R. CerBeruS: A System Supporting the Sequential Screening Process. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 241−245.

(5) Stanton, D. T.; Morris, T. W.; Roychoudhury, S.; Parker, C. N. Application of Nearest-Neighbor and Cluster Analyses in Pharmaceutical Lead Discovery. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 21−27.

(6) Menard, P. R.; Lewis, R. A.; Mason, J. S.Rational Screening Set Design and Compound Selection: Cascaded Clustering. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 497−505.

(7) Rusinko, A., III; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.

(8) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E., Jr. LeadScope: Software for Exploring Large Sets of Screening Data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302−1314.

(9) Jain, A. K.; Dubes, R. C. *Algorithms for Clustering Data*; Prentice Hall Advanced Reference Series: Englewood Cliffs, NJ, 1998.

(10) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared near neighbors. *IEEE Trans. Comput.* **1973**, C-22(11), 1025−1034.

(11) Chen, X.; Rusinko A., III; Young, S. S.Recursive Partitioning Analysis of a Large Structure−Activity Data Set Using Three-Dimensional Descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1054−1062.

(12) Rhee, M. A. van; Stocker, J.; Printzenhoff, D.; Creech, C.; Wagoner, P. K.; Spear, K. L. Retrospective Analysis of an Experimental High-Throughput Screening Data Set by Recursive Partitioning. *J. Comb. Chem.* **2001**, *3*, 267−277.

(13) Ladd, B. Intuitive Data Analysis: The next high-throughput technology demands advanced information tools. *Modern Drug Discovery* **2000**, *3*(1), 46−52.

(14) Brown, R. D.; Martin, Y. C. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(15) Hansch, C.; Fujita, T. $\rho-\sigma-\pi$ analysis − A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616−1626.

(16) Frank R. B. Quantitative Structure−Activity Relationship Studies Using Gaussian Processes. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 830−835.

(17) Gao, H.; Williams, C.; Labute, P.; Bajorath, J. Binary Quantitative Structure−Activity Relationship (QSAR) Analysis of Estrogen Receptor Ligands. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 164−168.

(18) Chen, X.; Rusinko A., III; Tropsha, A.; Young, S. S. Automated Pharmacophore Identification for Large Chemical Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 887−896.

(19) Willet, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *39*, 983−996

(20) Morize, I.; Menard, P. R.; Mason, J. S.; Bauerschmidt, S. Chemistry space metrics in diversity analysis, library design, and compound selection. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1204−1213.

(21) Matter, H.; Potter, T. Comparing 3d pharmacophore triplets and 2d fingerprints for selecting diverse compounds subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211−1225.

(22) Fisanick, W. The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 1. Basic Concepts. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 145−154.

(23) MDL Information Systems, Inc., San Leandro, CA. Home page: http://www.mdli.com/.

(24) Daylight Chemical Information Systems, Inc., Mission Viejo, CA. Home page: http://www.daylight.com/.

(25) Barnard Chemical Information Ltd., Sheffield, U.K. Home page: http://www.bci1.demon.co.uk/.

(26) Wild, D. J.; Blankley, C. J. Comparison of 2d fingerprint types and hierarchy level selection methods for structural grouping using wards clustering. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 155−162.

(27) Wagener, M.; Sadowski, J.; Gasteiger, J. Assessing similarity and diversity of combinatorial libraries by spatial autocorrelation functions and neural networks. *Angew. Chem., Int. Ed. Engl.* **1995**, *34*(23), 2674−2677.

(28) Rhodes, N.; Willett, P.; Dunbar, J. B.; Humblet, C, H. Bit-string methods for selective compound acquisition. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 210−214.

(29) Downs, G. M.; Willett, P.; Fisanick, W. Similarity Searching and Clustering of Chemical-Structure Databases Using Molecular Property Data. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094−1102.

(30) Godden, J. W.; Xue, L.; Bajorath, J. Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and tanimoto coefficients. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 163−166.

(31) Flower, D. R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.

(32) MacCuish J. D.; Nicolaou C. A.; MacCuish N. J. Ties in proximity and clustering compounds. *J. Chem. Inf. Comput. Sci.* **2001**, *41*(1): 134−146.

(33) Nicolaou, C. A.; MacCuish, J. D.; Tamura, S. Y. A new multidomain clustering algorithm for lead discovery that exploits ties in proximities. In *Rational Approaches to Drug Design;* Proceedings from the 13th

European Symposium on Quantitative Structure−Activity Relationships, Dusseldorf, Aug. 27−Sept. 1, 2000; Prous Scientific.
(34) Bioreason, Inc., home page: http://www.bioreason.com/.
(35) Kohonen, T. *Self-Organizing Maps*, 2nd ed; Springer: Heidelberg, 1997.
(36) Wagener, M.; Gasteiger, J. The Determination of Maximum Common Substructures by a Genetic Algorithm: Application in Synthesis Design and for the Structural Analysis of Biological Activity. *Angew. Chem., Int. Ed. Engl.* **1994**, *33*, 1189−1192.
(37) Developmental Therapeutics Program, National Cancer Institute, Bethesda, MD, home page: http://dtp.nci.nih.gov/.
(38) Weislow, O. S.; Kiser, R.; Fine, D. L.; Bader, J.; Shoemaker, R. H.; Boyd, M. R. New soluble-formazan assay for HIV-1 cytopathic effects: application to high-flux screening of synthetic and natural products for AIDS-antiviral activity. *J. Natl. Cancer Inst.* **1989**, *81*, 577−86.

CI010244I