

## Data Analysis of High-Throughput Screening Results: Application of Multidomain Clustering to the NCI Anti-HIV Data Set

Susan Y. Tamura, Patricia A. Bacha,\* Heather S. Gruver, and Ruth F. Nutt

Bioreason, Inc., 150 Washington Avenue, Suite 220, Santa Fe, New Mexico 87501

Received November 21, 2001

The routine use of high-throughput screening (HTS) systems in the drug discovery process has resulted in an increasing need for fast, reliable analysis of massive amounts of data. A new automated multidomain clustering method that thoroughly analyzes screening data sets is used to examine both the active and the inactive compounds in a well-known, publicly available data set based on primary screening. Large and small compound sets that defined both chemical families and potential pharmacophore points were discovered. The detection of structure–activity relationships (SAR), aided by the unique classification method, is described in this article.

### Introduction

With the advent of automated biological *in vitro* bioassays and automated chemical synthesis, embodied in high-throughput screening (HTS) and combinatorial chemistry, the complete analysis of large sets of diverse molecules, their structural motifs, and corresponding levels of activity has become an emerging problem. The scientists involved in HTS have presented the computational chemists with a set of new problems: effectively managing, condensing, and utilizing the body of knowledge contained in the massive amounts of data.

The focus of primary screening data analysis is to identify lead compounds. The identification of individual compounds or compound sets initially takes place by examination of each compound and evaluation of the corresponding biological information. A thorough analysis is performed to identify useful structure–activity relationship (SAR) information. At the end of the process, the results of the automated analysis are supplied to human experts who prioritize the compounds of interest and reach a decision on the individual promise of the identified leads based on various parameters. The problem of lead identification in HTS data sets is the discovery and evaluation of information within the data set that may be beneficial for drug discovery. Information of this type can be obtained in a variety of forms, mainly by isolating chemical families that exhibit higher than average activity against the target, by discovering important SAR<sup>1–4</sup> information, and by detecting significant pharmacophore points.<sup>5</sup>

Identification of chemical families can be performed in a two-dimensional (2D) or three-dimensional (3D) space.<sup>6–8</sup> Two-dimensional families are obtained by categorizing compounds with similar 2D structures or substructures. Similarly, 3D families contain compounds that have similarities in 3D space. Commonly, chemical families are defined by a contiguous structure, the scaffold, or by a Markush structure<sup>9</sup> common to all of the compounds in the family. Chemical classes that

exhibit favorable characteristics are often chosen as candidates for follow-up studies.

One goal of HTS data analysis is to reduce the data to a manageable level. Recent advances have been made in quantitative structure–activity relational expert system (QSAR-ES), as illustrated by the expert system MCASE,<sup>10,11</sup> with capabilities of learning from and organizing data. A structural feature hierarchy organized from a predefined library of structural templates has been employed by LeadScope to allow chemists to explore large sets of chemical compounds, their properties, and their biological activities.<sup>12</sup>

The human immunodeficiency virus (HIV) is the critical viral vector that causes AIDS, the acquired immunodeficiency syndrome. The AIDS antiviral screen, led by the National Cancer Institute's (NCI) Developmental Therapeutic Program, has enabled the testing of thousands of chemicals for their ability to inhibit the HIV virus in a cell-based assay. These data, which have been previously examined by Klopman and Tu using MCASE<sup>10</sup> and Roberts et al. using LeadScope, were analyzed using the LeadPharmer/DrugPharmer/TreeView suite of programs. (These programs are available from Bioreason, Inc., 150 Washington Ave., Santa Fe, NM, 87501; URL address: <http://www.bioreason.com>.) Although HIV-1 virus inhibitors have been studied extensively using molecular modeling and QSAR approaches, these methods are based on a small number (less than 110) of congeneric molecules acting via a single mechanism of action.<sup>13</sup>

Three drug discovery processes, exemplified in the study of the anti-HIV-1 data set, will be addressed in this paper: automated class identification, data-mining queries for class prioritization, and extraction of SAR information contained in the data. Moreover, a decision support tool to selectively choose compounds for future screening based on 2D substructure-based virtual screening is a valuable extension of the data management. The LeadPharmer program along with its user interfaces, DrugPharmer and TreeViewer, is an automated reasoning system that organizes the data into discreet classes through the analysis of 2D structural elements from an

\* To whom correspondence should be addressed. Tel.: 505 995-8188 ext. 208. Fax: 505 995-8186. E-mail: [bacha@bioreason.com](mailto:bacha@bioreason.com).

active data set. The automated analysis is designed to prioritize all of the lead classes, to gain information about SAR models, and to identify the structural domains responsible for activity.

## Materials and Methods

**NCI Anti-HIV-1 Database.** The NCI's HIV antiviral screen utilized a soluble formazan assay<sup>14</sup> to measure the ability of compounds to protect human CEM cells from HIV-1-induced cell death (URL: <http://dtp.nci.nih.gov>). In the primary screening set of results, the activities of the compounds tested in the assay were described in three categories: confirmed active for compounds that provided 100% protection, confirmed moderately active for compounds that provided more than 50% protection, and confirmed inactive for the remaining compounds or compounds that were toxic to the CEM cells and therefore appeared to not provide protection. For the purposes of this analysis, the categories were assigned numeric values of 0 (inactive), 1 (moderately active), or 2 (confirmed active) and the average activity of a class was used as a measure of its distribution.

The compounds used in our analysis included 385 confirmed active, 996 confirmed moderately active, and 38 054 confirmed inactive. All compounds were standardized in a data normalization process to ensure their consistent and correct representation.<sup>15</sup> This automated process involved (i) conversion of an SD file or SMILES string to a canonical SMILES that did not retain chiral information, (ii) removal of salts and solvents, (iii) removal of compounds containing metals, (iv) protonation or deprotonation of ions, and (v) removal of invalid structures. In addition, duplicate structures with lower activity were removed to enable the explanation of SAR results. All structures and their corresponding biological activities were registered in a MySQL database in preparation for data analysis.

**Automated Analysis.** The algorithmic process used to grow and postprocess phylogenetic-like trees (PGLT) is described in detail in Nicolaou et al.<sup>16</sup> Briefly, this process is based upon a hybrid algorithm employing various techniques ranging from neural networks and genetic algorithms to expert rules and chemical substructure searching. It is of repetitive nature with several reoccurring steps in the main part of the algorithm and an initializing and a postprocessing step. The PGLT algorithm does not use any information of biological activity of the compounds and, hence, is unsupervised.

The initial step of the process constructs the root node of the PGLT and populates it with all of the active molecules in the HTS data set under investigation. The root node is subjected to the key, repetitive part of the algorithm that consists of the following steps: (i) a clustering algorithm is used to group the molecules based upon the similarity of their chemical descriptors using a neural network-based self-organizing map method; (ii) the clustering results are processed, and a set of "natural" clusters are selected; (iii) the maximum common substructure for each cluster is learned capturing a potentially new similarity axis among the compounds of the cluster;<sup>17,18</sup> (iv) the common substructures are evaluated by a set of expert rules and eliminated if they do not represent a significant gain in new knowledge or are identical to either the parent node or the other subsets of the parent node; (v) new nodes are created, one for each newly found common substructure, and all compounds that are in a parent node are queried with this substructure with matching compounds added to the new nodes; (vi) new nodes are linked to the parent node; and (vii) go nodes are selected from all leaf nodes based upon a set of rules that define tree growth. The process then iterates and performs these steps using the go node as input. The process terminates when one of the predefined criteria have been met as follows: (i) the terminal nodes have a low molecular diversity coefficient, (ii) the terminal nodes are below a specified size, or (iii) the tree has built to a preset depth level.

An important feature is the ability of the algorithm to accommodate the multidomain nature of chemical compounds.

Compounds are distributed into a diverse set of nodes. The algorithm attempts to detect all structural features with significant presence in the data set and constructs nodes defined by those features regardless of the population of other nodes. All of the nodes that are detected are completely populated based on the maximum common substructure. Many substructures may define a given class. This property of the algorithm is aimed at discovering all structural types as well as discovering underrepresented chemical classes.




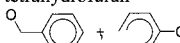
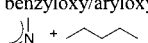

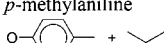
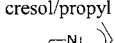
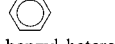
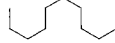
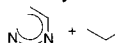
When the PGLT is complete, several postprocessing steps are employed to condense and enhance the information contained therein. Again, a hybrid system employing both expert rules and statistical methods is used to extract further information from the PGLT analysis. In the first step, molecular similarity methods based on both fingerprint and graph-based representations of the compounds in a node are used to examine whether or not the compounds in a node are structurally similar enough to represent a chemical class. Nodes for which the compounds have sufficient chemical similarity are marked as homogeneous. The second step annotates all of the unique homogeneous nodes as classes or subclasses based upon a set of expert rules related to the number of compounds in the class and the composition of those compound sets. Subclasses may be derived from more than one class and are annotated with the ID of each class to which it is related. All compounds that are not defined by the classes are defined as singletons. This method of classification completely categorizes all of the active molecules in classes or as singletons. The third step provides the comparison of the active compounds of the data set under investigation with the corresponding inactive compounds. The inclusion of the entire inactive data set for a complete analysis is important for class characterization and SAR determination. Inactive compounds are added to the root node and then filtered using common substructure queries for each node. All inactive compounds that match such queries are placed into the node that was used to generate the query. The automated generation of R-tables for all classes, subclasses, and singletons is the fourth step. Last, the nodes representing structurally homogeneous families are further processed and populated with statistical values related to the activity of the compounds in the node (mean, standard deviation, and percent actives). Furthermore, substructures defining the node hierarchy are examined and the relationships among node attributes are evaluated to determine which nodes may contain SAR information.

## Results

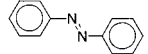
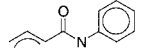

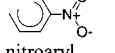
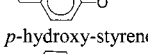
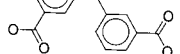
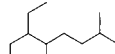
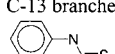
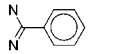
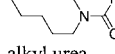
**Classification.** The advantages of building a classification structure based on the PGLT algorithm are exemplified in the following set of results. A phylogenetic tree was built containing 177 structural classes defined by 1–3 learned substructures. These substructures could be contiguous or noncontiguous. Of the 1381 compounds, 93.8% was classified (Tables 1–3). The multidomain nature of the classification was evident in the following statistic: the average number of classes into which a compound fell was 5.2.<sup>19</sup> Overall, 130 classes had overlapping membership while 47 classes had unique membership.

A range of known chemical families were featured prominently in the list of classes. Among others, the algorithm discovered a well-known large class (class 1) consisting of a collection of 183 compounds containing a tetrahydrofuran (THF) substructure. The majority of members of this class consisted of the nucleosides, including pyrimidine, dihydropyrimidine, and purine nucleosides (Figure 1). Active compounds that contained the THF without the heterocycle were not as well-represented. An overlapping class (class 2) was defined by the pyrimidine substructure. The majority of the

**Table 1.** Selected Large Classes

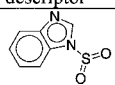

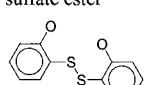
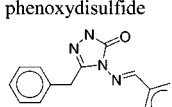
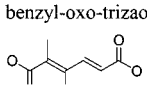
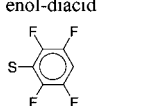
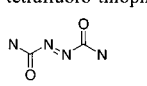
descriptor	no. actives	avg. activity	s.d.	avg. MW
 phenyl sulfone	311	1.35	0.48	629
 pyrimidine	205	1.49	0.50	460
 tetrahydrofuran	183	1.62	0.49	530
 benzyloxy/aryloxy	170	1.32	0.47	627
 heteroaryl/pentyl	184	1.49	0.50	530
 <i>p</i> -methylaniline	157	1.34	0.48	666
 cresol/propyl	131	1.26	0.44	648
 benzyl-heteroaryl	120	1.30	0.46	659
 C-11 hydrocarbon	130	1.48	0.50	711
 heteroaryl/amine	119	1.59	0.49	561
 hydroxy-ether	119	1.50	0.50	667

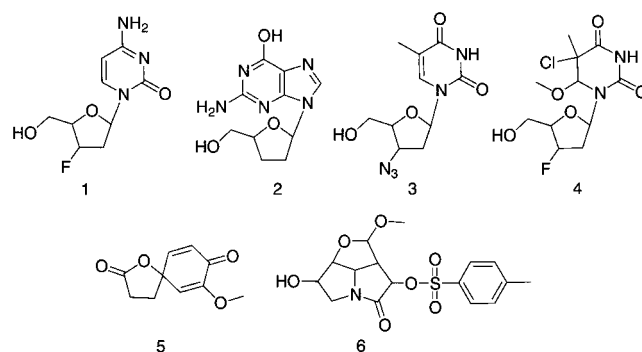
**Table 2.** Selected Well-Represented Classes

descriptor	no. actives	avg. activity	s.d.	avg. MW
 diphenyl azo	114	1.53	0.50	790
 phenyl arylamide	106	1.30	0.46	689
 2,3-dihydroxybutane	103	1.35	0.48	763
 nitroaryl	73	1.23	0.43	435
 <i>p</i> -hydroxy-styrene	72	1.22	0.42	664
 diphenylmethane dicarboxylate	69	1.25	0.43	676
 C-13 branched hydrocarbon	60	1.27	0.45	758
 aniline thiocarbonyl	53	1.81	0.40	397
 phenyl amidine	26	1.15	0.37	563
 alkyl urea	19	1.89	0.32	394

membership from class 2 contained the pyrimidine nucleoside framework; however, there were representa-

**Table 3.** Selected Underrepresented Classes with Low Number of Corresponding Inactives

descriptor	no. actives	avg. activity	s.d.	avg. MW
 sulfonyl-benzimidazole	2	1.5	67	386
 sulfate ester	2	1.0	29	568
 phenoxydisulfide	2	1.0	22	389
 benzyl-oxo-triazole	3	1.0	38	301
 enol-diacid	3	1.0	75	241
 tetrafluoro-thiophenol	3	1.0	75	203
 azo-dicarboxamide	3	1.0	50	181

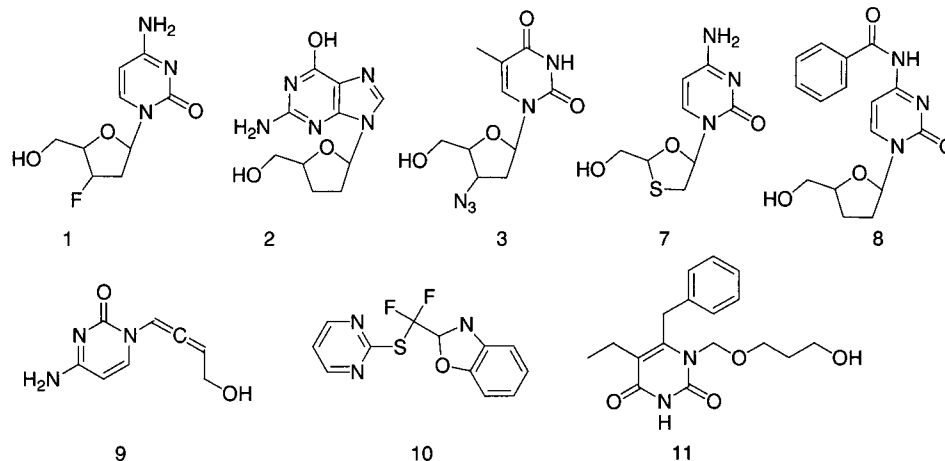


**Figure 1.** Members of class 2 defined by a THF substructure. Compounds 1–4 contain a heterocycle attached to a sugar substructure. Compounds 5 and 6 contain the THF without the additional heterocycle but are nonetheless unique members of this class.

tives containing pyrimidines without a sugar and pyrimidine nucleoside mimics (Figure 2). The class characteristics for classes 1 and 2 are shown in Table 4.

In the classification, 86 compounds were not defined and were labeled singletons. These singletons are based on the active set alone: 19 singletons had corresponding inactive structures associated with them that contained the entire structure of the singleton as a substructure (Table 5). Interestingly, three singletons had an activity value of 2 (confirmed active), while the majority of 83 singletons had an activity value of 1 (moderately active).

**Class Prioritization.** The automated method is used to classify the active compounds according to structural motifs without regard to biological activity. Incorporation of the biological information after the classification is complete may lead to further understanding of the data set including highlighting classes based on either



**Figure 2.** Members of class 2 defined by a pyrimidine substructure. Compounds **1–3**, **7**, and **8** contain a pyrimidine directly attached to a cyclic ether substructure. Compounds **9–11** contain the pyrimidine without a sugar substructure, but are nonetheless unique members of this class.

**Table 4.** Class Characteristics of Classes 1 and 2

	descriptor	no. actives	avg. activity	SD	no. inactives	% actives
class 1	pyrimidine	205	1.49	0.5	3382	5.72
class 2	THF	183	1.62	0.49	1345	11.98

**Table 5.** Selected Singletons with Corresponding Inactives

structure of singleton	activity	no. inactives
	1.0	22
	1.0	17
	1.0	16
	1.0	8
	1.0	7
	1.0	4
	2.0	2

a high activity (Table 6) or a high range of activity (Table 7). Classes with a higher range of activity may be classes that contain SAR. These classes that display SAR should have a higher potential for optimization. The SAR can be derived from the automated classification (see below) or by expert visualization of the compounds within a class.

Classes that uniquely describe the active compounds relative to the inactive compounds are also of high

**Table 6.** Selected Classes with High Median Activity

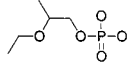
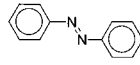
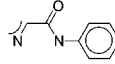


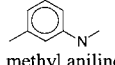
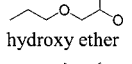
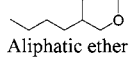
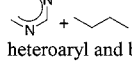
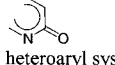

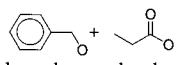
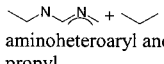
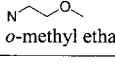
descriptor	no. actives	avg. activity	s.d.	avg. MW	no. inactives
 amide/urea heteroaryl system	16	1.75	0.45	1160	2
 alkyl urea	19	1.89	0.32	394	63
 aniline/thiocarbonyl	52	1.81	0.40	397	455
 hydroxyether/amide	35	1.80	0.41	595	182
 hydroxyether and isobutyl	75	1.67	0.47	585	673

interest (Table 8). These classes are discovered by using the parameter “percent actives”, which is defined as the number of active compounds divided by the total number of compounds in a class. A high percent actives attribute of a class indicates that the defining substructure(s) differentiated the active compounds relative to the inactive compounds in a particular HTS screen. These classes may be targeted for 2D substructure virtual screening for the preparation of test lists from an unscreened compound library. The purpose of this procedure is to prepare new lists of compounds to be tested that should be enriched in active compounds relative to the total unscreened library. All of the active compounds within the unscreened library will not be found; however, the subset of compounds that are screened should be enriched in active compounds relative to inactive compounds. The new active compounds discovered will have substructure similarities with regard to the original active set.

Class prioritization also yields classes of lower rank. Two types of lower priority classes may be discerned using DrugPharmer: threshold classes and classes that contain compounds that have a nonspecific interaction


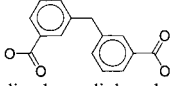
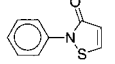
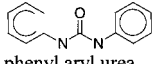
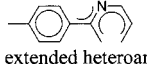
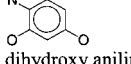

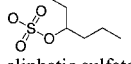

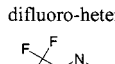


**Table 7.** Selected Classes with High Median Activity and High Range of Activity

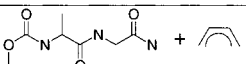
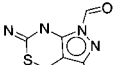
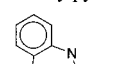
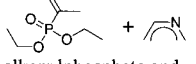
descriptor	no. actives	avg. activity	s.d.	avg. MW	no. inactives
 phosphate/ether	49	1.73	0.45	727	53
 diphenyl diazo	114	1.53	0.50	790	202
 phenylamide/heteroaryl	22	1.64	0.49	996	99
 heteroaryl	110	1.66	0.47	528	640
 tetrahydrofuran	183	1.62	0.49	530	1345
 methyl aniline	73	1062	0.49	575	646
 hydroxy ether	86	1.62	0.49	711	785
 Aliphatic ether	33	1.58	0.50	669	359
 heteroaryl and butyl	161	1.52	0.50	505	1771
 heteroaryl system	131	1.58	0.50	503	1941
 imide	39	1.51	0.51	369	1176
 benzyloxy and carboxy	40	1.53	0.51	733	619
 aminoheteroaryl and propyl	31	1.52	0.51	767	543
 o-methyl ethanolamine	116	1.56	0.50	561	2665

with the assay. Threshold classes may be difficult to optimize and therefore are of lower priority. Two conditions define threshold classes. Threshold classes contain active compounds whose activities are close to the threshold, as well as similar inactive compounds that make SAR difficult to determine (Table 9). Drug-Pharmer extracted the classes that met some of the requirements of threshold classes, but the chemist must determine whether the actives are structurally comparable to the inactive compounds and whether SAR can be discovered or not. If there is no comprehensible SAR, the class may be considered at the threshold. If there is SAR, then the class may be a low activity class, which would have higher priority. One active and several inactive compounds in one threshold class are illustrated in Figure 3. Compounds **12** and **13** were moderately active, but there were no useful SAR relationships that could be discovered when comparing compounds **12** and **13** to compounds **14**–**16**. This suggests that this class would be difficult to optimize

**Table 8.** Unique Active Substructures Defined by the Classification: Classes with a High Percent Active Value

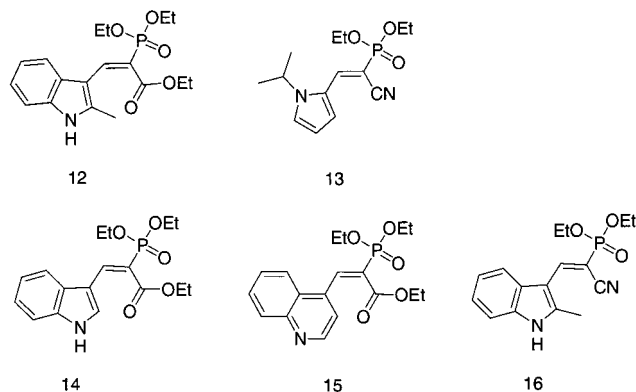
descriptor	no. actives	avg. activity	s.d.	% actives
 o-methyl-thiophenol	105	1.42	0.50	24
 dicarboxy diphenylmethane	69	1.25	0.43	34
 phenyl-isothiazolone	18	1.5	0.51	23
 phenyl aryl urea	16	1.06	0.25	70
 extended heteroaryl system	15	1.13	0.35	29
 dihydroxy aniline	13	1.23	0.44	26
 benzo-fused heteroaryl system	11	1.36	0.51	38
 aliphatic sulfate	7	1.29	0.49	37
 difluoro-heteroaryl system	7	1.43	0.54	23
 difluoromethyl sulfide	6	1.5	0.55	46

**Table 9.** Threshold Classes

descriptor	no. actives	avg. activity	s.d.	% actives
 peptide and aryl	7	1.00	0.00	5.3
 carbonylpyrazole	16	1.00	0.00	32
 benzodiazepinone	5	1.00	0.00	26
 alkenylphosphate and heteroaryl	7	1.00	0.00	32


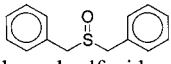
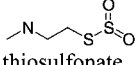
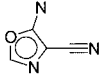
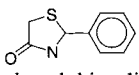
and would not be a desirable choice for follow-up chemistry studies.

The classes that contain compounds that interact with the assay in a nonspecific manner have similar measured activities and contain few, if any, inactive structures. Classes of this type will have a signature of a very low range of activity and a very high percent active value and, thus, are easily discovered within the data set. No nonspecific inhibitors were discovered in the NCI



**Figure 3.** Threshold class: compounds **12** and **13** were moderately active. Compounds **14–16** were inactive.

**Table 10.** Potential False Positives

descriptor	no. actives	% actives
 imidazoline thione	2	7.1
 benzyl sulfoxide	2	6.3
 thiosulfonate	3	7.3
 amino-cyano oxadiazole	4	6.6
 phenyl thiazolidinone	4	3.1

HIV data set: this may be due in part to the inclusion of toxic compounds in the inactive set.

Classes that contain potential false positives are characterized by having a small number of active compounds and a large number of similar inactive compounds (Table 10). Members of these classes should be retested to determine their activity. DrugPharmer extracted the classes that met some of the requirements of false positive classes. However, because many classes contain inactive structures that are not of great similarity to the active compounds, the chemist must determine whether these classes truly need retesting. False positives are exemplified in a class containing a 4-cyano-5-amino-oxazole substructure (Figure 4). The active compound (compound **17**) may be a uniquely active compound but should be retested considering all of the inactives of similar structure.

**SAR Analysis.** DrugPharmer extracts the classes with a higher range of activity because these classes may possess SAR. These classes are targeted for SAR analysis in DrugPharmer and TreeViewer.

SAR can be discovered within the classification in three ways. SAR is extracted within a class directly from R-tables or from two methods using class to subclass relationships that are derived from the defining substructures and corresponding average activity changes. The second and third methods have enabled the dis-

covery of SAR by observing the increased substructure definition of the child node or the subclass with concomitant changes in activity. Although statistically significant differences between two compound sets are the most useful in determining SAR, a qualitative understanding of early SAR derived from noisy HTS data sets can be valuable. In addition, the SAR learned from HTS data sets may be qualitatively distinct from the SAR learned from a combinatorial or focused library.

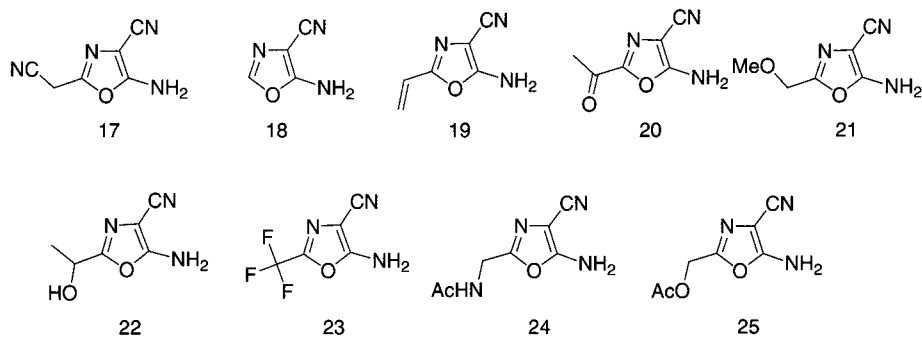
Using the first method of SAR extraction, relationships within a class can be observed using automatically generated R-tables. Visualization and sorting of R-tables in DrugPharmer allow the user to extract SAR from a class. This type of early SAR extraction may be observed later in biological secondary screens. Several cases of this type of SAR extraction are exemplified in three distinct classes (see examples in Figures 1–3). The compounds in the nucleoside class (Figure 5) were distributed into several categories based on the R<sub>1</sub> and R<sub>4</sub>, yielding potency-enhancing and -lowering features. The greater the number of representative compounds within a category, the greater the statistical significance of the SAR discovered. The 3'-azido-thymidine and 3'-fluoro-thymidine analogues are of greater potency, while the 3'-hydroxy-N-alkyl-cytosine or 3'-hydroxy-thymidine analogues were of lower potency. The N-palmitoyl-cytosine and cytosine analogues were of greater potency.

The 4,4'-diamino-styrene 2,2'-bis-sulfonate class was also studied in order to discover SAR from R-tables (Figure 6). In examining this class, several potency-enhancing and -lowering features were found. Two of the three azo-containing categories of compounds were potency-enhancing. The categories that contained phenyl or quinoline R-groups were potency-lowering.

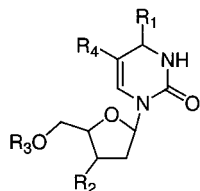
The third class that was examined to discover SAR was a diverse phenyl sulfonamide class (Figure 7). Within this class were two similar categories containing a chlorothiophenol. Of these two, the benzamide in the position meta to the sulfonamide was potency-enhancing, while the methyl substituent in the position meta to the sulfonamide was potency-lowering. A third category containing 13 compounds contained an acridine system, but this substitution had no net effect on activity.

The second method of automated SAR extraction is defined by tree growth. A parent node to child node relationship that is determined by the PGLT algorithm can reveal SAR; however, the compounds within a parent node must have high structural similarity in order to ensure the extraction of relevant information. Several examples of SAR discovered using this method are shown in Table 11 (classes 3–23). In this table, six parent classes (classes 3, 6, 9, 13, 16, and 20) were compared to those two or three child classes with larger common substructures that had the potential to reveal the most SAR information based upon the difference in average activity between the parent and the child.

Figure 8 shows the relationship between the parent class 3 and two of its more highly defined children, classes 4 and 5. Class 4 contained compounds of higher average potency and was described by an azidothymidine phosphate ester. Class 5, described by a thymidine phosphonamidate, contained compounds of lower aver-

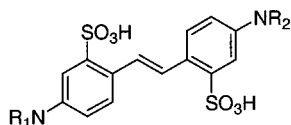


**Figure 4.** Potential false positives: compound **17** is active, but compounds **18–25** are inactive.



no. actives	R <sub>1</sub>	R <sub>2</sub>	avg. activity	activity change	effect on potency
113 (total)			1.68		
54	O	N <sub>3</sub>	1.85	0.17	enhancing
3	NH-n-C <sub>18</sub> H <sub>37</sub>	O or OCHO	1.00	-0.68	lowering
6	NHCO-n-C <sub>15</sub> H <sub>31</sub>	H or phosphate ester with nucleoside	1.83	0.15	enhancing
5	O	H	2.00	0.32	enhancing
3	O	OH	1.00	-0.68	lowering
10	NH	H	1.90	0.22	enhancing
10	O	F	2.00	0.32	enhancing

**Figure 5.** SAR derived from an R-table based on the active compounds of a nucleoside class. R<sub>4</sub> = CH<sub>3</sub>, H, or F; R<sub>5</sub> = H, ester, phosphodiester, or phosphotriester.

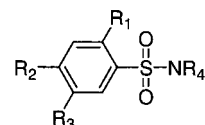


no. actives	R <sub>1</sub>	R <sub>2</sub>	avg. activity	activity change	effect on potency
45 (total)			1.49		
4	substituted phenyl or quinoline	H	1.00	-0.49	lowering
4	azo naphthol bis-sulfonic acid	same as R <sub>1</sub>	2.00	0.51	enhancing
4	azo naphthol sulfonic acid	same as R <sub>1</sub>	2.00	0.51	enhancing
5	azo naphthyl mono-sulfonic acid	same as R <sub>1</sub>	1.40	-0.09	lowering
14	substituted phenyl	same as R <sub>1</sub>	1.36	-0.13	lowering

**Figure 6.** Qualitative SAR derived from an R-table of a 4,4'-diamino-styrene 2,2'-bis-sulfonate class.

age potency, suggesting that the phosphonamidate is a potency-lowering feature.

The polyoxygenated ether parent class 6 has two more highly defined child classes (Figure 9). The more potent class 7 was described by a large multicyclic diester functionality, while the less potent class 8 was described by a hexose triacetate substructure.



no. cmpds	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	avg. activity	activity change	effect on potency
88 (total)					1.13		
5	SH	Cl	benzamide	hetero-cycle	1.40	0.27	enhancing
31	SH	Cl	methyl	hetero-cycle	1.03	-0.10	lowering
3	H	H	2'(alkyl-thioester)-benzamide	H, COCH <sub>3</sub>	1.50	0.37	enhancing
3	H	hetero-aromatic system	H	C(NH)(NH <sub>2</sub> )	1.00	-0.13	lowering
13	H	extended acridine system	H	H	1.18	0.05	minimal

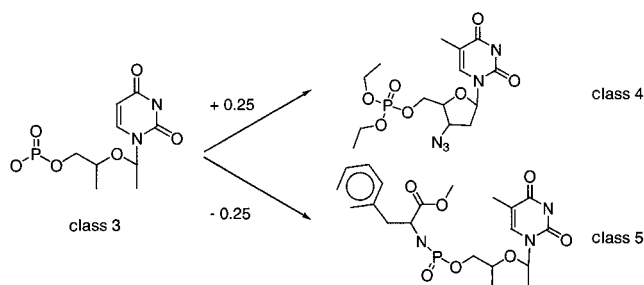
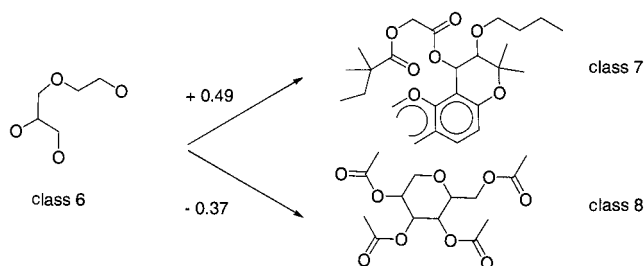
**Figure 7.** Qualitative SAR derived from an R-table of a phenyl sulfonamide class.

Class 9 was specified by two noncontiguous substructures, phenyl and nitro (Figure 10). Three child classes of varying average activity were subsequently defined. While the biphenyl sulfone (class 11) was an activating feature, an *o*-nitrophenyl-phenyl sulfone (class 10) gave rise to a highly active set of compounds. Class 12, defined by a phenyl sulfonate moiety, which was a substructure noncontiguous to the phenyl and nitro functionalities, contained compounds of lower activity.

From the automated classification of a subset of the total active set, consisting of compounds with a molecular weight of 500 or less, a new categorization rich in SAR information was produced. The parent to child relationships exemplified in classes 13–23 (shown Table 11) revealed an expanded understanding of the SAR within the data set. Class 13 included the pyrimidine nucleosides (Figure 11). The 3-azidothymidine members of this class (class 14) created a new class of higher overall potency. The compounds containing a 3'-hydroxylamine functionality (class 15) were of lower potency. Class 16, defined by a smaller substructure than class 13 and containing a broader sampling of nucleosides, was subdivided into three classes of varying activity (Figure 12). Class 17 contained a highly active set of 3'-fluoropyrimidine nucleosides. Class 18 contained methylcytosine derivatives of higher average potency than the parent class. Class 19, defined by a purine nucleoside substructure, contained compounds of lower potency.

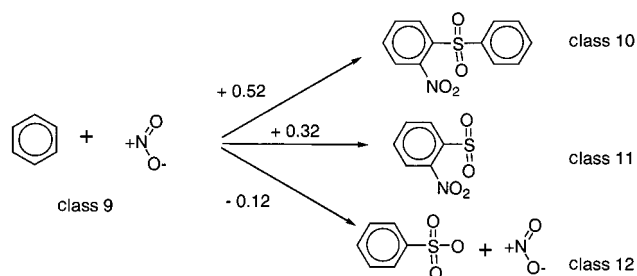
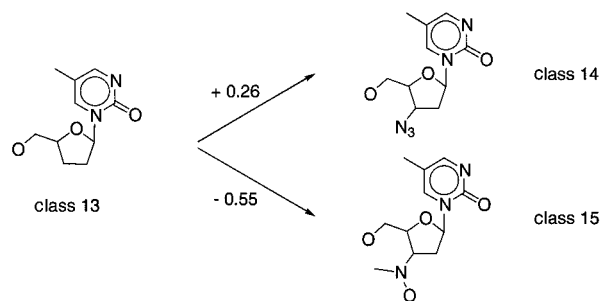
**Table 11.** SAR Derived from the Automated Classification: Parent to Child Relationships, Revealing Potency-Enhancing and -Lowering Substructures

	no. actives	avg. activity	% actives	activity change	description of class	effect on potency
class 3	42	1.74	68		pyrimidine ether phosphorus ester	
class 4	7	2.00	100	+0.26	AZT phosphate ester	enhancing
class 5	8	1.50	100	-0.24	thymine phosphonamidate	lowering
class 6	51	1.37	5.0		polyoxygenated ether	
class 7	14	1.86	93	+0.49	multicyclic diester	enhancing
class 8	6	1.00	3.5	-0.37	hexose triacetate	lowering
class 9	77	1.22	2.9		phenyl; nitro	
class 10	10	1.75	47	+0.52	<i>o</i> -nitrophenyl-phenyl sulfone	enhancing
class 11	55	1.55	35	+0.32	<i>o</i> -nitrophenyl sulfone	enhancing
class 12	10	1.10	24	-0.12	sulfonate	lowering
class 13	29	1.69	45		pyrimidine nucleosides	
class 14	19	1.95	66	+0.26	AZT analogues	enhancing
class 15	7	1.14	33	-0.55	hydroxylamine pyrimidine nucleosides	lowering
class 16	68	1.54	20		nucleosides	
class 17	7	2.00	78	+0.46	fluoro thymidine nucleosides	enhancing
class 18	8	1.75	35	+0.21	methylcytosine nucleosides	enhancing
class 19	11	1.27	12	-0.27	purine nucleosides	lowering
class 20	31	1.29	3.9		heteroaromatic with 2 nitrogen atoms	
class 21	4	2.00	80	+0.71	thionobenzimidazoles	enhancing
class 22	18	1.22	25	-0.07	phenyl dihydrothiophenes	none
class 23	5	1.10	1.1	-0.29	benzo-fused heteroaromatic	lowering

**Figure 8.** Parent and child node relationships. Phosphorus nucleosides. The numbers indicate the difference in average activity of the compounds within the parent class and the child class.**Figure 9.** Parent and child node relationships. Polyoxygenated ethers. The numbers indicate the difference in average activity of the compounds within the parent class and the child class.

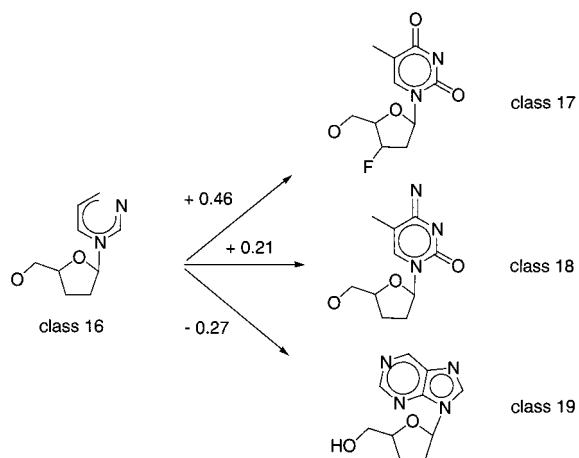
Class 20 was a broadly defined class containing a variety of nitrogen-containing heteroaromatics. This class was subsequently refined in the classification to yield three smaller classes of varying activities (Figure 13). The most active subset of class 20, described by a thionobenzimidazole substructure, afforded a highly potent set of compounds (class 21). The majority of members of class 22 contained a phenyl dihydrothiophene substructure. Class 22 was of similar average activity as the parent class. The benzo-fused heteroaromatic substructure defined class 23, which was of lower potency.

The third method of automated SAR extraction is defined by class/subclass relationships. SAR from class

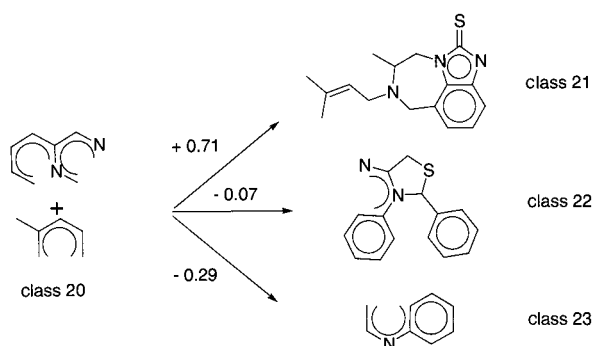
**Figure 10.** Parent and child node relationships. Phenyl- and nitro-containing class. The numbers indicate the difference in average activity of the compounds within the parent class and the child class.**Figure 11.** Parent and child node relationships. Pyrimidine nucleosides. The numbers indicate the difference in average activity of the compounds within the parent class and the child class.

to subclass or across subclasses, a property independent of tree growth, can be discovered. Subclasses, in general, have greater structural definition than the corresponding classes. An example of the investigation of SAR from subclass relationships within a broadly defined class is summarized in Figure 14 and Table 12. The nucleoside-containing class 24, defined by two noncontiguous heteroaromatic and alkane substructures, was further refined to the slightly more potent class 25, defined by a substructure that included a nitrogen-containing aromatic attached to a sugar moiety. Class 25 was split into three classes of varying average activities. Two classes contained potency-enhancing features: the 3'-





**Figure 12.** Parent and child relationships. Nucleosides. The numbers indicate the difference in average activity of the compounds within the parent class and the child class.



**Figure 13.** Parent and child relationships. Heteroaromatic compounds. The numbers indicate the difference in average activity of the compounds within the parent class and the child class.

fluoro-pyrimidine nucleosides (class 26) and the thymidine nucleosides (class 27). The purine nucleosides (class 28), also a subset of class 25, revealed a potency-lowering substructure. The thymidine nucleoside class 27 was increasingly defined by the classification to yield more SAR information. Class 29, described by a 3'-azido functionality, contained compounds of greater average activity, while class 30, described by a 3'-hydroxy moiety, contained compounds of less average activity.

The three methods of SAR extraction are complementary. SAR extraction can differ in nature depending upon the type of data set. While using the automated classification to derive SAR, diverse data sets exhibit coarse SAR, whereas focused data sets such as combinatorial libraries display refined SAR. For smaller, well-defined classes, the R-tables are advantageous for the discovery of SAR. In contrast, for large classes, or for classes that do not have R-tables, the second and third methods are superior.

**Pharmacophore Point Identification.** The relevance of particular pharmacophore points can be determined by the association of structural motifs with the activity of a class. Furthermore, the percent actives characteristic is an important indicator for pharmacophore point identification. Substructures that define the active set relative to the inactive set are more likely to interact with the enzyme or receptor binding site. Key observations must be made in order to test the effect of

learned structural elements on activity. The SAR associated with the defining substructures may be a function of the active compounds alone (average activity) and/or a function of the total data set (percent actives). Each classification can be explored for relevance of the substructure keys to biological activity observed in a specific assay.

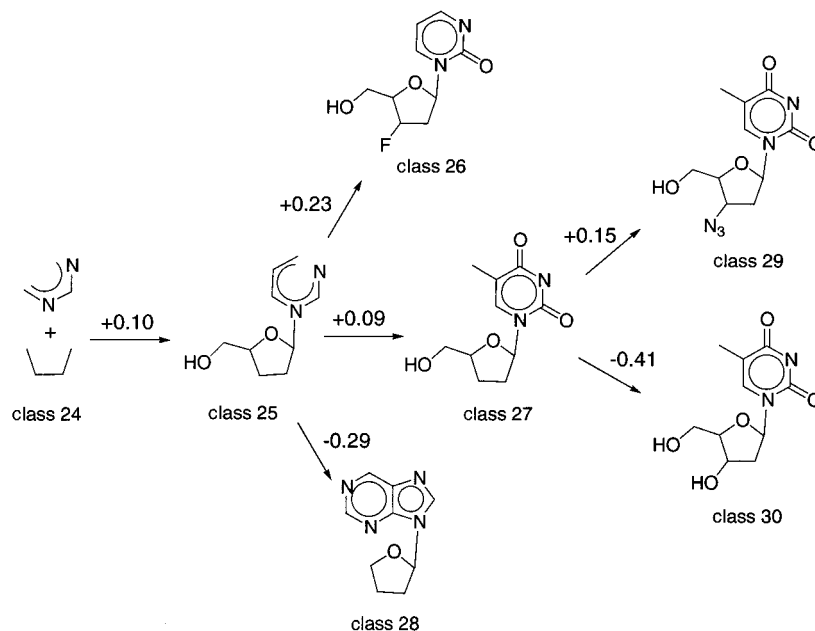
Several examples of the information that may be derived from multidomain classification are described. The multidomain nature of compound **26** is illustrated in three classes (Figure 15). Class 31 contains compounds described by overlapping substructures: an aromatic sulfonic acid with a carbonyl in the meta position and an aromatic hydroxyl. Class 32 contains compounds with a 3-hydroxy-biphenyl substructure. Class 33 contains a tricyclic system, but this descriptor defined a threshold class and may not be a key pharmacophore point. Classes 31 and 32 exhibit a range of activity and are potential SAR classes. When compared to classes 32 and 33, class 31 has greater potency and better definition with respect to the inactive compounds, as revealed by the percent actives value. From this information, it is possible to hypothesize that the *m*-carbonyl-phenyl sulfonic acid with the aryloxy functionality is the key pharmacophore point. Compound **26** yields the most useful information when clustered with compounds in class 31.

The multidomain nature of compound **27** is illustrated in three classes (Figure 16). Class 34, which is defined by two noncontiguous substructures, an aryl ether and a propane, contains compounds of greater average activity and a percent actives value of 11.69. Class 35, with phenyl propene and ethanol descriptors, is comprised of compounds of lower potency and a lower percent actives value. Class 36, described by an acridone system, contains only compounds of lower activity and is considered a potential threshold class. Compound **27** is contained in a higher priority potential SAR class, class 34. Class 34 may be a more valuable classification based on optimization potential. Furthermore, class 34 is described by two substructures that may be more valuable for describing potential pharmacophore points: the aryl ether and the hydrophobic propane.

Compound **28** was characterized in two classes, as illustrated in Figure 17. Class 37 was defined by two noncontiguous substructures: an ethylamine and a phenyl sulfonamide. Class 38 was defined by a large substructure containing a triaminotriazine attached to a phenyl sulfonamide. Because it had a range of activity, class 37 contained more SAR information. If the classification was based solely on the triamino-triazine class 38, compound **28** would be considered a compound in a threshold class of lower priority. The large substructure describing class 38 was not an important pharmacophore point. However, when placed in class 37, compound **28** provided valuable SAR information in a class that has optimization potential.

## Discussion

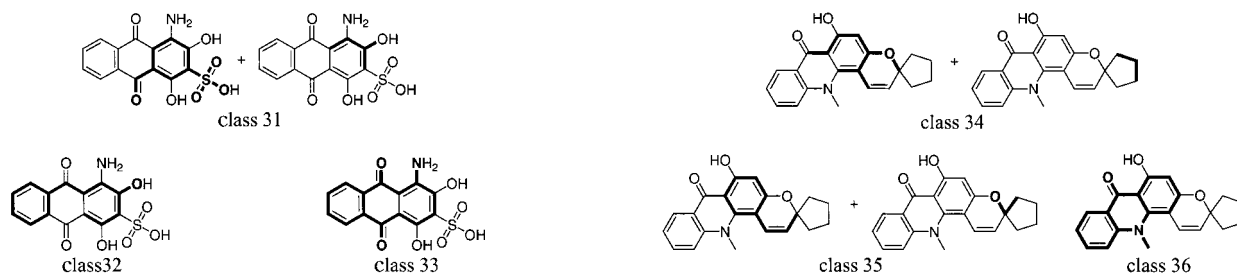
The desire to speed up the drug discovery process while increasing throughput has created a need for a new set of cost-effective computational tools for rapid and in-depth data analysis. Visual data-mining tools, while allowing experts to explore and identify trends



**Figure 14.** Discovering SAR from subclass relationships. The numbers indicate the change in the average activity of the compounds in a class from a larger, diverse class to a smaller class defined by a larger substructure.

**Table 12.** Discovering SAR from Subclass Relationships

	no. actives	avg. activity	% actives	description of class	effect on potency
class 24	42	1.52	8.3	heteroaromatic containing 2 nitrogen atoms; butane	
class 25	7	1.62	26	nucleosides	enhancing
class 26	8	1.85	72	3'-fluoro nucleosides	enhancing
class 27	51	1.71	47	thymidine nucleosides	enhancing
class 28	14	1.33	10	purine nucleosides	lowering
class 29	6	1.86	86	3'-azido-thymidine nucleosides	enhancing
class 30	77	1.30	19	3'-hydroxy-thymidine	lowering



	no. actives	avg. activity	s.d.	% actives	learned Substructures
class 31	7	1.43	0.5	30.44	m-carbonyl-phenyl sulfonic acid; aryloxy
class 32	77	1.21	0.4	12.16	3-hydroxy-diphenylmethane
class 33	4	1	0	9.09	amino anthraquinone

**Figure 15.** Multidomain compound **26**, moderately active. Highlighting indicates the learned substructures. The compound in class 31 is described by two overlapping substructures depicted separately in the figure.

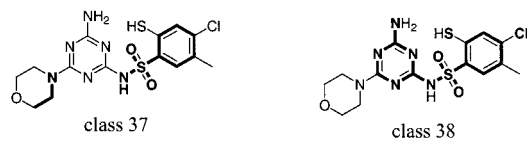
	no. actives	avg. activity	s.d.	% actives	learned Substructure
class 34	38	1.47	0.51	11.69	aryl ether, propyl
class 35	37	1.24	0.43	3.11	phenylpropene; ethanol
class 36	3	1	0	17.65	acridinone

**Figure 16.** Compound **27**, moderately active. Highlighting indicates the learned substructures. The compound in class 34 is described by two overlapping substructures depicted separately in the figure.

and relations in small- to medium-sized data sets, have fallen short in their ability to analyze large, complex data sets. There is an increasing demand for automated approaches that can rapidly discover key information contained in screening data in an unbiased manner. Human experts can explore knowledge extracted by these approaches, making informed decisions rather than manually extracting knowledge and rules.

In this paper, we have analyzed the data from the full NCI AIDS antiviral screen using a set of algorithms

to grow and postprocess PGLTs. This general approach compared favorably with the techniques previously reported for analysis of a similar data set using MCASE<sup>10,11</sup> or LeadScope.<sup>12</sup> The advantage of the MCASE system is its ability to learn and provide a set of structural fragments statistically linked to activity. In addition, it is unique in its ability to predict the potential biological activity of an unknown compound based upon these fragments. However, fragments in a system like MCASE are limited in size



	no. actives	avg. activity	s.d.	% actives	learned substructures
class 37	32	1.22	0.42	4.6	ethyl amine; phenyl sulfonamide
class 38	8	1	0	57	triamino-triazine-phenyl-sulfonamide

**Figure 17.** Multidomain compound **28**, moderately active. Highlighting indicates the learned substructures.

(2–10 atoms) and are linear and, thus, do not contain closed rings. Moreover, the number of compounds that can be used in the analysis is limited to a few thousand, forcing the user to examine only the structural properties of the active compounds while ignoring all of the inactive compounds that may contain valuable SAR information.

In contrast, LeadScope can easily reduce the information obtained from an HTS screening set to a manageable level by classifying compounds into a set of hierarchically structured chemical families using 2D molecular fragments. To a large extent, the fragments in LeadScope contain substructures based on functional groups as well as cyclic and acyclic fragments. LeadScope applies the same dictionary of structural fragments for classification to each data set; the usefulness of the resulting analysis is highly dependent on how well the data set is classified by its fixed dictionary. In addition, the structural variations that may describe SAR may not be identified by LeadScope techniques because the variations in structure may be either too small or too complex to be discovered with predefined fragments.

PGLT, like LeadScope, has been used to efficiently analyze HTS screening data sets by using a hierarchy of substructures as descriptors for structural classes. In contrast to both MCASE and LeadScope, PGLT fragments are based on cyclic or acyclic substructures that are learned from small clusters of compounds within the data set. As compared to other methods, these fragments are more varied in size and complexity. The fragments may be smaller, based on as little as two atoms; larger, based on unlimited atom count or predefined fragment size; or more complex, based on more than one type of contiguous functionality or substructure. Moreover, the extraction of SAR from a large data set is simplified when using the common substructures defining PGLT classes and the R-tables derived from those classes. The multidomain clustering method described in this paper has enabled the automatic identification of classes, their prioritization based on class-based reasoning, and the efficient extraction of SAR information contained in the data.

## Conclusion

The PGLT classification results exhibited 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. The data analysis has also shown that the method allows for the formation of structural families from a given data set and the extraction of detailed SAR information in an efficient manner. In

addition, partially because of the multidomain nature of the PGLT classification, the method has proven to be robust to the highly desirable property of discovering under- and overrepresented structural families of compounds within the data set. Key to this outcome are 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.

## Abbreviations Used

HTS, high-throughput screening; SAR, structure–activity relationship; QSAR, quantitative structure–activity relationship; NCI, National Cancer Institute; HIV, human immunodeficiency virus; PGLT, phylogenetic-like tree.

**Acknowledgment.** We thank Bobi K. Den Hartog, Brian P. Kelley, and Christos A. Nicolaou for the application of their programming expertise to computational and medicinal chemistry issues and Terence K. Brunck for critical reading of the manuscript.

## References

- 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.
- 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.
- Frank, R. B. Quantitative Structure–Activity Relationship Studies Using Gaussian Processes. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 830–835.
- 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.
- 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.
- Willet, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *39*, 983–996.
- 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.
- Matter, H.; Potter, T. Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211–1225.
- 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.
- Klopman, G. MULTICASE I. A Hierarchical Computer Automated Structure Evaluation Program. *Quant. Struct.-Act. Relat.* **1992**, *11*, 176–184.
- Klopman, G. Artificial Intelligence Approach to Structure–activity Studies. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315–7321.
- 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.
- Latest articles include (a) Jalali-Herari, M.; Parastar, F. Use of Artificial Networks in a QSAR Study of Anti-HIV Activity for a Large Group of HEPT Derivatives. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 147–154. (b) Knaggs, M. H.; McGuigan, C.; Harris, S. A.; Heshmati, P.; Cahard, D.; Gilbert, I. H.; Balzarini, J. A QSAR Study Investigating the Effect of L-alanine Ester Variation on the Anti-HIV Activity of Some Phosphoramidate Derivatives of d4T. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 2027–2078. (c) Huuskonen, J. QSAR Modeling with the Electropotential State: Tibo derivatives. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 425–429. (d) Filippini, E.; Cruciani, G.; Tabarrini, O.; Cecchetti, V.; Fravolini, A. QSAR Study and VolSurf Characterization of Anti-HIV Quinolone Library. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 203–217.

- (14) 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.
- (15) The normalization process is a module within the LeadPharmer software package.
- (16) Nicolaou, C. A.; Tamura, S. Y.; Kelley, B. P.; Bassett, S. I.; Nutt, R. F. Analysis of Large Screening Datasets Via Adaptively Grown Phylogenetic-like Trees. *J. Chem. Inf. Comput. Sci.*, in press.
- (17) McGregor, J. J. Backtrack Search Algorithms and the Maximal Common Subgraph Problem. *Software-Practice and Experience*, **1982**, *12*, 23–34.
- (18) Bayada, D. M.; Simpson, R. W.; Johnson, A. P.; Laurencio, C. An Algorithm for the Multiple Common Subgraph Problem. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 680–685.
- (19) This calculation is based on the classes, while omitting the subclasses and singletons.

JM010535I