

## Rule Extraction from a Mutagenicity Data Set Using Adaptively Grown Phylogenetic-like Trees

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

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

Received February 12, 2002

A public bacterial mutagenicity database was classified into 2-D structural families using a set of specific algorithms and clustering techniques that find overlapping classes of compounds based upon chemical substructures. Structure–activity relationships were learned from the biological activity of the compounds within each class and used to identify rules that define substructures potentially responsible for mutagenic activity. In addition, this method of analysis was used to compare the pharmacologically relevant substructure of test compounds with their potential toxic substructures making this a potentially valuable *in silico* profiling tool for lead selection and optimization.

### INTRODUCTION

Screening compounds for potential toxicity during the discovery phase of drug development can result in earlier and better decisions about dedication of resources. However, even *in vitro* testing of large compound libraries can often be costly and time-consuming. Alternatively, *in silico* screening can be useful in identifying a potential toxicity hazard before it becomes a stumbling block in the development process. Several *in silico* methods have been used to model a variety of toxicity endpoints with particular attention to mutagenicity and carcinogenicity.

Regulatory guidelines state that drug candidates must be assessed for genotoxicity in a battery of three tests prior to first human exposure.<sup>1</sup> The first test measures the ability of a compound to cause reverse mutations in a series of defined bacterial strains.<sup>2,3</sup> The assay is conducted in the absence and presence of a liver microsomal preparation. The latter is included to determine the impact of metabolic activation on the genotoxicity of the compound. The test takes several days to perform and analyze potentially requiring several grams of compound as the protocol requires testing a negative compound to the limit of its solubility or until it exhibits toxicity. A positive result for any strain whether with or without metabolic activation is sufficient for the test compound to be considered a bacterial mutagen.

Currently available computer based toxicity analysis systems are based upon two general methodologies: a mechanistic approach and a correlative approach. In the first case, rules, many of which are based upon the studies of Miller and Miller<sup>4</sup> as systematized by Ashby and collaborators,<sup>5–8</sup> are developed from expert knowledge of the relationship between chemical structure and plausible mechanism of toxicity and then used to predict the toxicity of test compounds (DEREK<sup>9</sup> and ONCOLOGIC<sup>10</sup>). In the second case, automated algorithms are used first to extract chemical substructures or descriptors from a database of compounds and then to compute a correlation between the substructure

or descriptor and toxicity (CASE,<sup>11</sup> MULTICASE<sup>12</sup> and TOPKAT<sup>13</sup>). This learned information is then used in the analysis of test compounds. Both approaches have the potential to incorporate structural elements that may modulate toxicity.

Since the underlying rationale of the mechanistic approach is often more transparent to toxicologists than the correlative, the rules-based method is frequently given more credibility. However, this approach is limited by the fact that it requires considerable effort by experts to generate new rules. In addition, systems that use a toxic alert approach do not provide either a quantitative assessment of the likelihood that a particular compound will be toxic or an evaluation of how well a compound is covered by the specific set of expert rules used. In contrast, the correlative approach is capable of defining novel toxic substructures but is dependent on the quality of the data used to derive the underlying models both with respect to biological endpoint and coverage of chemistry space in question. In general, systems using a correlative approach include information on both toxic and nontoxic compounds so that they can provide information on coverage for a particular test compound and, in addition, provide a qualitative indication of the potential for a test compound to be toxic.

We have taken a different approach to the general issue of extracting structure–activity relationships by analyzing relationships between and within classes learned with a phylogenetic-like tree algorithm.<sup>14</sup> The approach presented here is based upon the general principle of relating compounds to other similar compounds on the basis of shared substructures rather than focusing just on structure–activity correlations. However, by subsequently overlaying biological data on the learned structural classes, nonexpert derived structure–activity rules can be extracted.<sup>15</sup> We also show how the results of two analyses related to different biological endpoints can be compared on the basis of classes and their scaffolds. The methods are unbiased and complete in that inactive compounds that are related to the classes contribute to the rules.

\* Corresponding author phone: (505)995-8188 x208; fax: (505)995-8186; e-mail: bacha@bioreason.com.

This report describes the application of these techniques to the analysis of a bacterial mutagenicity database derived from public sources with the goal of finding substructures that are potentially related to toxicity. As a means of validating the biological relevance of this study, toxic substructures identified by this analysis were compared to the rules or structural alerts for mutagenicity described in the literature.<sup>5-8,13,16-20</sup> In most cases the structural alerts identified by the analysis described in this report were similar to those previously published based upon expert analysis. However, this investigation has suggested some modifications to the traditional set of toxicophores to define more specific mutagenic alerts.

## METHOD

**Data Source.** Bacterial mutagenicity data were primarily obtained from the Chemical Carcinogenesis Research Information System. This toxicology data file is maintained by the National Cancer Institute and made public through the National Library of Medicine's Toxicology Data Network (TOXNET; <http://toxnet.nlm.nih.gov>). It is a scientifically evaluated and fully referenced database with mutagenicity results for individual bacterial indicator strains (*S. typhimurium* TA97, TA1537, TA98, TA100, TA1535, and TA102 and *E. coli* WP2 *uvrA*) with and without addition of rat liver microsomal preparation to measure metabolic activation. These data were supplemented both with information from the Genetic Activity Profile database (<http://www.epa.gov/mdwgapdb>) maintained by the Environmental Protection Agency in association with the International Agency for Research on Cancer and with data from a series of literature references.<sup>21-27</sup>

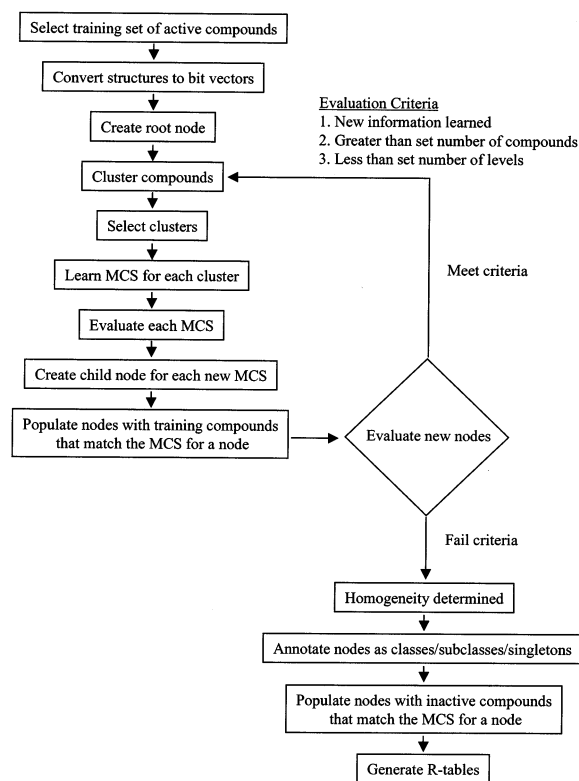
Two-dimensional structures in SMILES format were acquired for many of the compounds through TOXNET (ChemIDplus) and the National Cancer Institute's Developmental Therapeutics Program (<http://dtp.nci.nih.gov/>). The Merck Index and ChemFinder (<http://www.chemfinder.com>) were used to find additional structures.

Human tumor cell line screening data were downloaded from the National Cancer Institute's Developmental Therapeutics Program (<http://dtp.nci.nih.gov/>). Compound structural information in SD format was obtained from the same source.

**Data Processing and Compound Normalization.** Since multiple references for a particular compound and bacterial strain were possible, the biological data were screened for consistency. Only clearly mutagenic or nonmutagenic results were included in the final data set. Discrepant results were excluded. The final data set contained bacterial mutagenicity test results (scored as positive or negative) for a diverse set of chemicals, approximately 20% of which are pharmaceutical agents and physiological compounds.

The compound structure files were processed<sup>28</sup> to remove salts and stereochemical features from the SMILES or SD records. In addition, the structures for organometallics, polymers, and mixtures of compounds were eliminated from the file. The normalized compound structure file was then matched with the biological data file to create the final data set.

**Analytical Technique.** The algorithmic process used for generation of phylogenetic-like trees is described in detail in Nicolaou et al.<sup>14</sup> It is based upon a hybrid algorithm



**Figure 1.** Schematic of algorithmic process used to generate a phylogenetic-like tree.

employing a variety of techniques ranging from neural networks and genetic algorithms to expert rules and chemical substructure searching. It is repetitive in nature with several re-occurring steps in the main part of the algorithm linked to both a pre- and postprocessing step. The analysis is unbiased by activity or expert knowledge since activity is not used in the generation of the phylogenetic-like tree but only used to select the training set compounds.

The initial step is to construct a root node with all the training molecules in the data set converted to bit vectors based upon the particular chemical features present in each compound (Figure 1). In this case, structures for all of the compounds that were positive for mutagenicity in a particular bacterial strain were used as the training set.

Next the root node is subjected to the following series of steps to form the first level of clusters each defined by a distinctive maximum common substructure: 1) a clustering algorithm is used to group the molecules based upon feature vector similarity using a neural network based self-organizing map method, 2) the clustering results are processed and a subset selected based upon high compound similarity, 3) the maximum common substructure for each cluster is learned, 4) the common substructures are evaluated using a set of expert rules to eliminate any that either do not represent a significant gain in new knowledge or are identical to that of other clusters, 5) child nodes are created, one for each newly found common substructure, and 6) the structures of the compounds in the original node and each child common substructure are compared with those compounds with substructures that matched the common substructure of a child node are placed into a that node. For the purposes of this study, each such final group or node in the phylogenetic-like tree is called a structural class.

The clustering process is then repeated on each child node resulting in the formation of new classes with new substructures at each level until one of the following criteria is met: 1) no new information can be learned, 2) the number of compounds in the node is below a specified size, or 3) the tree has reached a preset depth level.

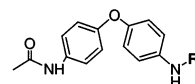
Again a hybrid system employing both expert rules and statistical methods is used to extract further information from the analysis during several postprocessing steps: 1) classes that have sufficient structure similarity based upon both fingerprint and graph-based representations of the compounds are marked as homogeneous, 2) homogeneous nodes are annotated as classes and subclasses based upon the composition of the nodes with compounds not covered in classes marked as singletons, 3) the structure of the nonmutagenic compounds are compared to the maximum common substructure of each class and compounds that match the substructure of a particular class are added to that class, and 4) R-tables are automatically generated for most classes.

**Analyses.** Altogether four different analyses are presented in this report.<sup>28</sup> The first analysis for finding structural alerts was conducted as described above with results from the *S. typhimurium* strain TA98 without metabolic activation (841 positive and 2972 negative compounds).

The next two analyses were designed to highlight potential differences in mechanism of action. Results from two different strains (TA98 with a base-pair deletion versus TA100 with a base-pair substitution) both in the absence of metabolic activation were combined to form the first comparison analysis (117 compounds positive in TA98 only, 181 positive in TA100 only, 463 positive in both test cases, and 2600 negative in both test cases). In addition, data from the TA98 strain with and without metabolic activation were used for the second comparison analysis (42 compounds positive without metabolic activation only, 295 positive with metabolic activation only, 409 positive in either test case, and 2341 negative in both test cases). For each analysis, a phylogenetic-like tree was created using compounds that were positive in either or both of the test cases being compared, e.g., TA98 without metabolic activation and TA100 without metabolic activation or TA98 with and without metabolic activation, and filtered with compounds that were negative in both test cases.

The fourth analysis was a cross-comparison of two different phylogenetic-like trees created to directly contrast the substructures potentially related to toxicological activity identified in one tree to those likely related to the pharmacological activity of drug-like compounds identified in another. Growth inhibition screening data for the A498 renal cancer cell line was used for analysis of substructures related to the pharmacological activity of the compounds. The ability of compounds to inhibit tumor cell growth was taken as a surrogate of pharmacological activity. Compounds that inhibited 50% of the cell growth at a concentration less than or equal to 0.1 micromolar ( $GI_{50} \leq 10^{-7}$  M) were considered active with the final data set containing 649 active compounds and 28319 inactive compounds. These compounds were tested *in silico* for mutagenic potential by filtering the active antitumor compounds through an analysis from a pooled bacterial mutagenicity test training set. The final mutagenicity data set consisted of 524 mutagenic and 1900 nonmutagenic compounds.

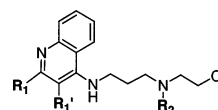
**Table 1.** R-Table for Diphenyl Ether Class<sup>a</sup>



compound no.	activity	R-group
172374-58-5	positive	OH, C2H3O
172374-60-6	positive	OC2H3O, C2H3O
172374-57-1	positive	OH
172374-59-3	positive	OH, CHO
135673-14-2	positive	O
2687-41-4	negative	
3070-86-8	negative	C2H3O

<sup>a</sup> Data taken from strain TA98 without metabolic activation.

**Table 2.** R-Table for Quinoline Class<sup>a</sup>



compound no.	activity	R <sub>1</sub> -R <sub>1'</sub>	R <sub>2</sub>
38914-96-4	positive	C4H4	
92280-00-7	positive	C8H8	
4251-89-2	positive	C8H8	C2H5
38915-14-9	positive	C4H4OC	C2H5
92280-01-8	negative	C4H8	
92279-99-6	negative	C4H8	C2H5

<sup>a</sup> Data taken from strain TA98 without metabolic activation.

## RESULTS

**Structural Alerts.** Three different techniques were used to extract rules for structure-activity relationships (SAR): R-table analysis within a class, multidomain comparison of compounds and their classes, and parent-child relationships between classes. All three of these techniques took advantage of the concept that classes with a higher percentage of positive compounds are more likely to be defined by common substructures that are related to mutagenic activity. However, it should be noted that while many such correlations can be found, not all substructures identified in this manner are necessarily directly responsible for induction of mutations.

The first approach used to extract SAR rules was to examine the differences between mutagenic and nonmutagenic compounds that fall within a single class. R-tables based upon the common scaffold of a class and individual compound substituents were used for induction of SAR information. The method was demonstrated with the TA98 (no metabolic activation) data set.

Twenty-two percent (841 compounds) of the 3813 compounds in this data set are mutagenic which is consistent with the findings of Zeiger and Margolin for a general survey of chemicals in commerce and industry.<sup>29</sup> Therefore, to find potentially significant structural alerts, we examined classes that contained at least 60% mutagenic compounds. The common scaffolds and R-tables for two such classes, a diphenyl ether class and a quinoline class, are presented in Tables 1 and 2, respectively. Both common scaffolds contain a nitrogen atom joined to an aromatic ring system, one of the primary literature mutagenicity alerts. Nonetheless, neither class consisted of only active compounds (71% and 67% mutagens, respectively). In the diphenyl ether class

(Table 1), all of compounds have a second aromatic amino-type alert. However, the five mutagenic compounds have nitroso, hydroxylamine, and aminoester groups attached to the core diphenyl ether structure, while the two nonmutagenic compounds have a less chemically reactive amino or acetamide group in this position. In the quinoline case (Table 2), the type of ring at the  $R_1$ – $R_1'$  position is critical for mutagenic activity. An aromatic ring in this position is associated with activity (four compounds), a saturated ring with inactivity (two compounds). In contrast, the addition of an ethyl group at  $R_2$  has no qualitative effect on mutagenicity.

These two examples show how the phylogenetic-like method of organizing a data set into classes and the automatically generated R-tables that organize compounds within a class can be used to extract rules. In these examples, the traditional alerts with respect to mutagenicity in the TA98 strain without metabolic activation were extended by the analysis to include the potential chemical reactivity of the "aromatic nitrogen" and the potential additive effect of having a polyaromatic system containing at least three rings.

In the phylogenetic-like method, a compound can appear in several classes that are defined by different common substructures or domains. This feature can be used as another approach for extracting the relationships between different domains of the molecule and activity. Specifically, the common substructure defining a class with a high percentage of mutagenic compounds is more likely to be related to mutagenicity than the common substructure that defines a class with a low percentage of mutagens.

Several compounds with an extensive multidomain classification were examined for potential structural alerts. Two such compounds, the doxorubicin derivative rudolfomycin and the mycotoxin alterperyleneol, are shown in Tables 3 and 4, respectively. Portions of the two compounds are highlighted in such a way as to show how the common substructure from different classes hit each molecule.

Starting with the rudolfomycin example (Table 3), the common substructure of class A defines a class primarily containing nonmutagenic compounds since only 13% of the compounds in this class tested positive in strain TA98 without metabolic activation. In contrast, 75% of the compounds in class D are mutagenic. Therefore, the anthraquinone structure with an additional hydroxyl group and sugar linker defining Classes D (9 mutagenic compounds and 3 nonmutagenic) appears to be more directly related to mutagenicity than any of the other common substructures shown here: the sugar chain of class A (13% actives; 126 mutagenic compounds and 417 nonmutagenic), the two-ring structure of class B (23% actives; 56 mutagenic compounds and 368 nonmutagenic), or the anthraquinone structure alone of class C (40% actives; 27 mutagenic compounds and 41 nonmutagenic).

The alterperyleneol compound (Table 4) was classified in a similar fashion by several different but overlapping domains. In this case, neither the partial ring structure defining class E with 9% actives (22 mutagenic compounds and 234 nonmutagenic), the phenolic structure of class F with 18% actives (143 mutagenic compounds and 640 nonmutagenic), nor the acetophenone structure in class G with 20% actives (86 mutagenic compounds and 352 nonmutagenic) would appear to be highly correlated with mutagenic activity. In contrast, both the partial and full

**Table 3.** Multidomain Classification of Rudolfomycin<sup>a</sup>

class	% positive	common substructure
A	13%	
	56 mutagenic 368 non-mutagenic	
B	23%	
	126 mutagenic 417 non-mutagenic	
C	40%	
	27 mutagenic 41 non-mutagenic	
D	75%	
	9 mutagenic 3 non-mutagenic	

<sup>a</sup> Data taken from strain TA98 without metabolic activation.

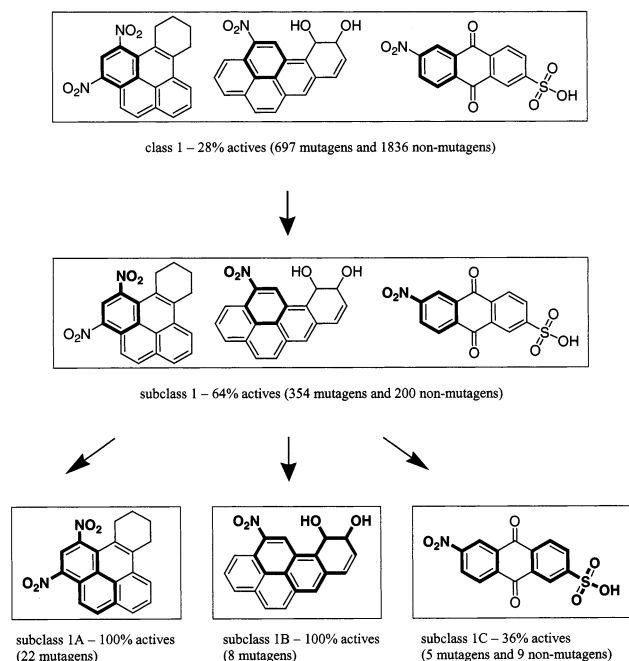
**Table 4.** Multidomain Classification of Alterperyleneol<sup>a</sup>

class	% positive	common substructure
E	9%	
	22 mutagenic 234 non-mutagenic	
F	18%	
	143 mutagenic 640 non-mutagenic	
G	20%	
	86 mutagenic 352 non-mutagenic	
H	76%	
	16 mutagenic 5 non-mutagenic	
I	89%	
	8 mutagenic 1 non-mutagenic	

<sup>a</sup> Data taken from strain TA98 without metabolic activation.

perylene core structure of the molecule with the two aliphatic alcohol groups (class H with 16 mutagenic compounds and 5 nonmutagenic and class I with 8 mutagenic compounds and one nonmutagenic) is more likely to be directly related





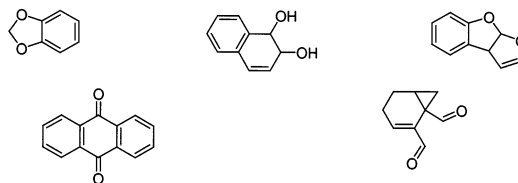
**Figure 2.** Aromatic nitro compounds: change in percent actives pathway for strain TA98 without metabolic activation.

to mutagenicity, as a much higher percentage of compounds in classes defined by these substructures are mutagenic (76% and 89%, respectively).

Rules generated from multidomain analyses define the mutagenicity alerts more specifically and with larger substructures than the general descriptions of small chemical features provided in the current literature.

The technique of finding differences in the percent of active compounds between a parent class and its subclasses was used to determine additional structure–activity relationships in this data set. For example, if the percent of compounds that are mutagenic in a particular subclass is greater than in the parent class, then the larger substructure defining the subclass is more likely to be correlated with genotoxicity. Alternatively, a decrease in the percent of active compounds as one moves along a path from a parent class to a subclass would indicate that the larger common substructure defining the subclass is potentially one that moderates mutagenic activity. However, since the number of compounds in each class decreases along the pathway, one needs to be careful in interpreting such results. For this reason, the smallest number of compounds in a class used for the examples presented here is eight.

Potential structural alerts and modifying elements were identified using this technique by examining substructures associated with substantial changes in the relative percent of actives along an analysis pathway. Figure 2 illustrates an example of such a pathway where there is both an increase and a decrease in the percent of mutagenic compounds as one moves from parent class to different subclasses. Three representative mutagenic compounds that appear in three different final subclasses are shown with common substructures highlighted. The common substructure for class 1 is an aromatic ring. Since the overall percentage of mutagenic compounds in the data set as a whole is 22% (841 mutagenic compounds and 2972 nonmutagenic) and there are 28% active compounds (697 mutagenic and 1836 nonmutagenic)

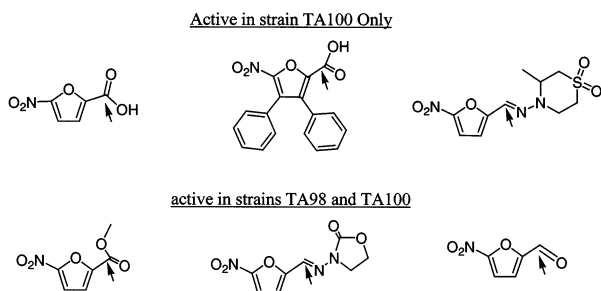


**Figure 3.** Oxygen-containing structural alerts for strain TA98 without metabolic activation.

in class 1, an aromatic ring by itself is not significant for mutagenicity. However, the presence of a nitro group in subclass 1 increases the percentage of active compounds in this class to 64% (354 mutagenic compounds and 200 nonmutagenic) and defines a common literature structural alert. The aromatic nitro alert is further enhanced in Subclasses 1A and 1B, both of which contain only mutagenic compounds with a polycyclic aromatic structure (22 and 8 mutagenic compounds, respectively). In addition, compounds in subclass 1A have a second nitro group while those in subclass 1B have a dihydroxy group. In contrast, only 36% (five mutagenic compounds and nine nonmutagenic) of subclass 1C are active compounds. The aromatic sulfonic acid group defining this subclass appears to be a modifying element that lowers the mutagenic potential for an aromatic nitro compound, confirming the earlier findings of Rosenkranz and Klopman.<sup>30</sup>

This example and others discovered using this parent-child analysis method show that a complete analysis of a data set is possible. Specifically, all parent classes containing the aromatic nitro alert were identified, and their subclasses examined for potential modifying elements. Typically aromatic nitro classes are comprised of approximately 60% mutagens. The substructure defining a subclass was considered to increase the potential of mutagenic activity if the subclass contained >80% mutagenic compounds. In contrast, subclasses with <40% active compounds were considered to be defined by substructures that decrease the likelihood of a compound being mutagenic. Overall, the presence of additional nitro groups, nitrogen containing side chains such as amines, acetamide, and *N*-methylol derivatives, or other substituents such as ethylene, butenal, or acetyl groups increases the probability of a compound being mutagenic. Likewise heteroaromatic nitro compounds containing oxygen, sulfur, or nitrogen, such as furans, thiophenes, or thiazoles are more likely to be mutagenic. Last, the presence of multiple aromatic or heteroaromatic rings (greater than or equal to three) increases the mutagenic potential of a compound. In contrast sulfonic and sulfamic acid groups decrease the likelihood that a structure is mutagenic.

Last, all three techniques to discover SAR were used to locate highly mutagenic classes that were not defined by traditional literature structural alerts. Since many nitrogen-containing substructures that we initially found could be considered an extension of the general toxic alert of a nitrogen atom joined to an aromatic ring system, we concentrated on oxygen-containing alerts. Our unbiased and complete analysis of classes enriched in mutagenic compounds for strain TA98 has highlighted the following alerts illustrated in Figure 3: benzodioxole, dihydro-naphthalenediol, dihydro-dioxo-cyclopenta[a]indene, terpene structures with unsaturated dialdehydes, and multiple aromatic ring systems with a quinone.



**Figure 4.** Difference in nitrofurans reactivity for strains TA98 and TA100.

**Table 5.** Examples of Compounds Requiring Metabolic Activation for Strain TA98

metabolized only	reaction	with or without metabolism
1A. 	epoxidation	1B. 
2A. 	n-hydroxylation	2B. 
3A. 	azo-reduction	3B. 

**Mechanistic Comparisons.** Phylogenic-like tree analyses can be used to compare the reactivity of compounds in two different assays by combining and coding the active compounds from both assays, generating phylogenetic-like classes, and examining classes that contain compounds that are active in only one or the other assay versus being active in both assay systems. In the first case, we explored the potential differences in mechanism of mutagenic activity in two different *Salmonella* test strains without metabolic activation: TA98 that detects frameshift mutagens and TA100 that is sensitive to missense mutagens. Derivatives of nitrofurans are universally mutagenic. However, not all derivatives are active in both test strains. As shown in Figure 4, substituents with a highly electrophilic  $\alpha$ -carbon (marked with arrow in figure) like carboxylic esters, aldehydes, and methylene-amino-oxazolidinone are mutagenic in both TA98 and TA100. In contrast, compounds with less electrophilic  $\alpha$ -carbon substitutions such as carboxylic acids and methylene-amino-thiomorpholinedioxide are positive in strain TA100 only.

We also examined the relationship between structure and the specific requirement for metabolic activation within the TA98 strain alone. As with the analysis described above, we examined classes that contained compounds that were active only with the addition of a metabolic activator or active regardless of whether a microsomal extract was added. This analysis was used to identify pairs of chemically similar compounds of interest for further investigation. Three such pairs of compounds are shown in Table 5. The compounds from each pair shown on the right-hand side of the table are positive in strain TA98 with or without metabolic activation. In contrast, the compounds from each pair shown on the left-

hand side of the table are only positive in the presence of a microsomal extract. As it turns out, known Phase I metabolic reactions can convert the compounds that require activation shown on the left in Table 5 (see arrow) to those that do not need to be metabolized as shown on the right.<sup>31</sup> First, the double bond in the dihydro-1,8-dioxo-cyclopenta[a]indene structure in Compound 1A can be oxidized to form the epoxide structure found in compound 1B (epoxidation). Next the acetamide side chain in compound 2A can be converted to the hydroxy-acetamide in compound 2B (N-hydroxylation). Last, the azo group joining the two aromatic rings in compound 3A can be reduced to release 2,4-dinitro-phenylamine (compound 3B; azo-reduction).

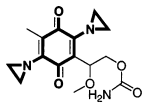
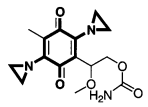
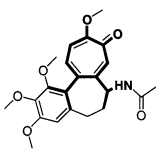
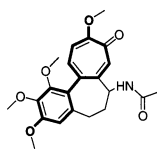
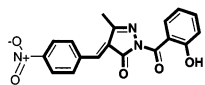
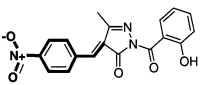
These findings suggest that although the actual metabolite responsible for mutagenic activity may not be identified, this analysis technique can be useful in the exploration of potential mechanistic pathways and is still relevant to the identification of the mutagenic potential of parent compounds regardless of their ultimate metabolic fate.

**Pharmacophore-Toxicophore.** Last, to test whether it is possible to use this type of analysis to compare the pharmacologically relevant substructure of a test compound with potential toxic substructures, information obtained from two different but related analyses were compared. Human tumor cell line screening data were evaluated based upon the pharmacological activity of the compounds tested, i.e., concentration that inhibited growth of a particular tumor cell line by 50%. Active compounds from the screen were used to create an analysis that was then filtered with the inactives. This analysis was used to find 2-D pharmacophore substructures for the potential antitumor agents based upon selection of high activity classes (classes with >50% compounds with antitumor cell activity). A second analysis was then conducted with the combined bacterial mutagenicity test data set using all of the compounds (mutagenic and nonmutagenic). Since this analysis was conducted using all compounds, classes generated by this analysis highlight both mutagenic (classes with a high percent of mutagens) and nonmutagenic (classes with a low percentage of mutagenic compounds) substructures. This second analysis was filtered with the active compounds from the tumor screen and used to determine potential toxicophore and toxicophobe substructures based upon the classes into which the tumor screen compounds filtered, i.e., classes with a high percentage of mutagenic compounds or ones with a low percentage of mutagens, respectively.

Three examples of comparisons between potential 2-D pharmacophore substructures and potential mutagenic/non-mutagenic substructures are shown in Table 6. The highlighting of substructures on the left-hand side of the figure represents the substructures that are potentially related to antitumor activity as derived from the cell line screen analysis. The classes defined by these substructures had an average growth inhibitory activity of  $<10^{-8}$  M. The structures highlighted on the right-hand side were taken from the bacterial mutagenicity test analysis and are indicative of substructures found by examining the class with the highest percent of mutagenic compounds in which a particular antitumor compound filtered.

The first comparison (compound A) shows a known anticancer agent, carbaziquone, which is also a known mutagen.<sup>32</sup> This compound filtered into a class in the bacterial

**Table 6.** Comparison of Relevant Substructures for Three Antitumor Compounds

potential 2-D pharmacophore	potential toxicophore/toxicophore	comments
A.		
		filtered into class with 100% mutagens 2 mutagens 0 non-mutagens known mutagen <sup>a</sup>
B.		
		filtered into class with only 8% mutagens 1 mutagens 13 non-mutagens known non-mutagen <sup>b</sup>
C.		
		filtered into class with 100% mutagens 8 mutagens 0 non-mutagens not tested for mutagenicity <sup>c</sup>

<sup>a</sup> Carbazilquinone.<sup>31</sup> <sup>b</sup> Colchicine.<sup>31</sup> <sup>c</sup> 2-(2-Hydroxybenzoyl-4-hydroxy(oxido)amino)benzylidene)-5-methyl-2,4-dihydro-2H-pyrazol-3-one.

mutagenicity analysis that contained two mutagenic compounds and no nonmutagenic compounds defined by an aziridiny-benzoquinone substructure. In this instance the potential pharmacophore structure and the toxicophore structure coincide. Likewise there is extensive overlap in the potential pharmacophore structure and the substructure identified by the bacterial mutagenicity analysis for the second example (compound B). However in this case, the test compound, colchicine, filtered into a class that was made up primarily of nonmutagenic compounds (only 8% mutagens with one mutagenic compound and 13 nonmutagenic). Therefore, the phenyl-cycloheptatriene substructure actually defines a nontoxic class which is consistent with the fact that colchicine is a known nonmutagen.<sup>32</sup>

The compound shown in the third example (compound C), 2-(2-hydroxybenzoyl-4-hydroxy(oxido)amino)benzylidene)-5-methyl-2,4-dihydro-2H-pyrazol-3-one, was found in a class in the bacterial mutagenicity test analysis that contained 100% mutagens (eight mutagenic compounds and no nonmutagenic) based upon a nitro-vinyl-benzene substructure. Though the actual bacterial mutagenicity activity of this compound is unknown, this analysis would suggest that it would be mutagenic. Interestingly, the potential pharmacophore substructure for this compound identified by the tumor cell line screening analysis overlaps with but does not entirely encompass the toxicophore substructure. Therefore, it is possible that this compound could be modified by substitution of the nitro group with another moiety, thus reducing its mutagenic potential and retaining its antitumor activity. In fact there were active compounds with hydroxyl or methoxy groups in this position in the same class in the antitumor analysis as this potentially mutagenic compound.

These three examples suggest ways in which this type of analysis can provide additional useful information on potential lead profiling and characterization. Such an evaluation can help distinguish situations in which potential pharma-

cophore and toxicophore structures coincide from those in which the potential toxic features of compounds can likely be removed during the optimization process without reducing their biological activity.

## CONCLUSION

Informed and rapid decision making is highly desirable in a world where getting new drugs to market more quickly and cost-effectively is critical. Screening compound libraries for potential undesirable attributes such as toxicity before they become stumbling blocks in the development process is key. The earlier such information is available the better, particularly when it is based upon specific structural characteristics. However, traditional *in vitro* and *in vivo* experimental methods have been pushed to the limit. In contrast, automated *in silico* approaches can potentially provide quick answers on a large number of compounds in an unbiased manner. Human experts can then spend more time exploring the knowledge extracted by these approaches and making informed decisions than on experimentally deriving new information or manually extracting knowledge and rules.

In this report we have described rule extraction methods for large data sets based on 2-D structural classes. These classes were then used to identify SAR derived within a structural class using automatically generated R-tables, from the multidomain nature of the compounds and classes and from the parent-child relationships between structural classes. Using the diverse public database of bacterial mutagenicity, the identified SAR, rules, and structural alerts were, in most cases, identical to those previously published based upon expert analysis of historical data.<sup>10,11</sup> However, given its unbiased and complete nature, this investigation has suggested modifications to the traditional set of toxicophores to define more specific mutagenic alerts with larger substructures than the small chemical features currently described in the literature.

By combining active compound sets and coding activity, this analysis has proven useful in examining both the relationship between structure and potential mechanism of mutagenic activity in different bacterial test strains and between structure and the specific requirement for metabolic activation. In fact, since the major limitation of toxicity analysis systems that use a mechanistic or expert rule-based approach is their inability to discovery new structure-toxicity relationships, the information gained by use of the correlative or algorithmic approach described in this report could easily be used to supplement mechanistic systems. In addition, extraction of new rules using common substructures from classes of compounds defined by phylogenic-like trees should prove to be easier than extraction of such rules from some of the other correlative analysis systems that define toxicity based upon small fragments divorced from the context of the entire molecule.

Most importantly, we have been able to use the methods described here to compare the pharmacologically relevant substructure of a compound, in this case directed at the ability to inhibit growth of a human tumor cell line, with its potential toxic substructures. The ability to easily do such comparisons should prove useful as an effective profiling tool in early lead selection, and is the key feature that separates this technique from the currently available toxicity analysis tools



that use a correlative approach. While these tools were designed to provide a direct assessment of the potential for a test compound to have a particular biological activity or toxicity, they do not currently directly provide the ability to integrate derived knowledge from different assays in the way the techniques described in this report can.

In conclusion, this report highlights the use of a phylogenetic-like tree algorithm to generate overlapping classes of compounds, the automatic building of R-tables from classes, and the extraction structure–activity relationships that can be used to define rules concerning mutagenic substructures. The use of this methodology can easily be expanded upon to address not only other types of toxicity data but also other general areas of drug design. The ability to augment human expertise with the data-driven discoveries of the type that can be made with this technology is a positive step forward in building automated systems for lead optimization analysis.

#### ACKNOWLEDGMENT

We thank Terence K. Brunck for his critical reading of the manuscript and insightful suggestions.

#### REFERENCES AND NOTES

- (1) International conference on Harmonization: Guidance on specific aspects of regulatory genotoxicity tests for pharmaceuticals. *Federal Register* **1996**, 61 (80), 18197–18202.
- (2) Ames, B. N.; McCann, H.; Yamasaki, E. Methods for detecting carcinogens and mutagens with the *Salmonella*/mammalian-microsome mutagenicity test. *Mutat. Res.* **1975**, 31, 347–364.
- (3) Maron, D. M.; Ames, B. N. Revised methods for the *Salmonella* mutagenicity test. *Mutat. Res.* **1983**, 113, 173–215.
- (4) Miller, J. A.; Miller, E. C. Ultimate chemical carcinogen as reactive mutagenic electrophiles. In *Origin of Human Cancers*; Hiatt, H. H., Watson, J. D., Winsten, J. A., Eds.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 1977; pp 605–627.
- (5) Ashby, J. Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity. *Environ. Mutagen.* **1985**, 7, 919–921.
- (6) Ashby, J.; Tennant, R. W.; Zeiger, E.; Stasiewicz, S. Classification according to chemical structure, mutagenicity to *Salmonella* and level of carcinogenicity of a further 42 chemicals tested for carcinogenicity by the U.S. National toxicology Program. *Mutat. Res.* **1989**, 223, 73–103.
- (7) Tennant, R. W.; Ashby, J. Classification according to chemical structure, mutagenicity to *Salmonella* and level of carcinogenicity of a further 39 chemicals tested for carcinogenicity by the U.S. National Toxicology Program. *Mutat. Res.* **1991**, 257, 209–227.
- (8) Ashby, J.; Tennant, R. W. Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U. S. NTP. *Mutat. Res.* **1991**, 257, 229–306.
- (9) Sanderson, D. M.; Earnshaw, C. G. Computer prediction of possible toxic action from chemical structure; the DEREK system. *Human Exp. Toxicol.* **1991**, 10, 261–273.
- (10) Woo, Y.-T.; Lai, D.; Argus, M.; Arcos, J. Development of structure–activity relationship rules for predicting carcinogenic potential of chemicals. *Toxicol. Lett.* **1995**, 79, 219–228.
- (11) Klopman, G. Computer automated structure evaluation of organic molecules. *J. Am. Chem. Soc.* **1984**, 106, 7315–7324.
- (12) Klopman, G. MULTI—CASE 1. A hierarchical computer automated structure evaluation program. *Quant. Struct. Act. Relat.* **1992**, 11, 176–184.
- (13) Enslein, K.; Gombar, V. K.; Blake, B. W. Use of SAR in computer-assisted prediction of carcinogenicity and mutagenicity of chemicals by the TOPKAT program. *Mutat. Res.* **1994**, 305, 47–61.
- (14) Nicolaou, C.; 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.* **2002**, 42, 1069–1079.
- (15) Bacha, P.; Nutt, R. *Development of Chemical-profiling software for early lead selection*. Predictive Toxicology; AdvanceTech Monitor: Woburn, MA, 2001; pp 238–250.
- (16) Rosenkranz, H.; Klopman, G. Structural alerts to genotoxicity: The interaction of human and artificial intelligence. *Mutagenesis* **1990**, 5, 333–361.
- (17) Cunningham, A. R.; Rosenkranz, H. S.; Zhang, Y. P.; Klopman, G. Identification of ‘genotoxic’ and ‘nongenotoxic’ alerts for cancer in mice: The carcinogenic potency database. *Mutat. Res.* **1998**, 398, 1–17.
- (18) Greene, N.; Judson, P. N.; Langowski, J. J.; Marchant, C. A. Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR, and METEOR. *SAR QSAR Environ. Res.* **1999**, 10, 299–314.
- (19) Mersch-Sundermann, V.; Rosenkranz, H. S.; Klopman, G. Structural basis of the genotoxicity of polycyclic aromatic hydrocarbons. *Mutagenesis* **1992**, 7, 211–218.
- (20) Levin, D. E.; Hollstein, M.; Christman, M. F.; Schwiers, E. A.; Ames, B. N. A new *Salmonella* tester strain (TA102) with A–T base pairs at the site of mutation detects oxidative mutagens. *Proc. Natl. Acad. Sci. U.S.A.* **1982**, 79, 7445–7449.
- (21) Haworth, S.; Lawlor, T.; Mortelmans, K.; Speck, W.; Zeiger, E. *Salmonella* Mutagenicity test results for 250 Chemicals. *Environ. Mutagen.* **1983**, 1(Suppl. 1), 3–142.
- (22) Mortelmans, K.; Haworth, S.; Lawlor, T.; Speck, W.; Tainer, B.; Zeiger, E. *Salmonella* mutagenicity tests: II. Results from the testing of 270 chemicals. *Environ. Mutagen.* **1986**, 8 (Suppl. 7), 1–119.
- (23) Zeiger, E.; Anderson, B.; Haworth, S.; Lawlor, T.; Mortelmans, K.; Speck, W. *Salmonella* mutagenicity tests: III. Results from the testing of 255 chemicals. *Environ. Mutagen.* **1987**, 9 (Suppl. 9), 1–110.
- (24) Zeiger, E.; Anderson, B.; Haworth, S.; Lawlor, T.; Mortelmans, K. *Salmonella* mutagenicity tests: IV. Results from the testing of 300 chemicals. *Environ. Mutagen.* **1988**, 11 (Suppl. 12), 1–158.
- (25) Zeiger, E.; Anderson, B.; Haworth, S.; Lawlor, T.; Mortelmans, K. *Salmonella* mutagenicity tests: V. Results from the testing of 311 chemicals. *Environ. Mutagen.* **1992**, 19 (Suppl. 2)1, 2–141.
- (26) Dunkel, V. C.; Zeiger, E.; Brusick, D.; McCoy, E.; McGregor, D.; Mortelmans, K.; Rosenkranz, H. S.; Simmon V. F. Reproducibility of Microbial Mutagenicity Assays: I. Tests with *Salmonella typhimurium* and *Escherichia coli* using a standardized protocol. *Environ. Mutagen.* **1984**, 6 (Suppl. 2), 1–254.
- (27) Dunkel, V. C.; Zeiger, E.; Brusick, D.; McCoy, E.; McGregor, D.; Mortelmans, K.; Rosenkranz, H. S.; Simmon V. F. Reproducibility of Microbial Mutagenicity Assays: II. Testing of carcinogens and noncarcinogens in *Salmonella typhimurium* and *Escherichia coli* using a standardized protocol. *Environ. Mutagen.* **1985**, 7 (Suppl. 5), 1–248.
- (28) LeadPharmer, DrugPharmer, and TreeViewer software from Bioreason, Inc., Santa Fe, NM was used for the analysis. For classification of compounds, the standard multidomain, 3-level method was selected.
- (29) Zeiger, E.; Margolin, B. H. The proportion of mutagens among chemicals in commerce. *Reg. Toxicol. Pharmacol.* **2000**, 32, 219–225.
- (30) Rosenkranz, H. S.; Klopman, G. Structural basis of the mutagenicity of phenylazoaniline dyes. *Mutation Res.* **1989**, 221, 217–239.
- (31) Gibson, G. G.; Skett, P. *Introduction to Drug Metabolism*, 2nd ed.; Stanley Thornes Publishers Ltd.: Cheltenham, UK, 1999; pp 3, 8, and 10.
- (32) Chemical Carcinogenesis Research Information System available through TOXNET at <http://toxnet.nlm.nih.gov>.

CI020366Q