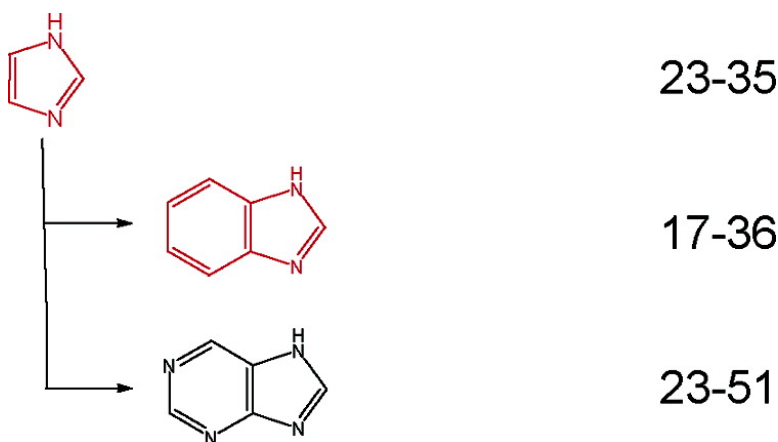


## Ring Systems in Mutagenicity Databases

Richard Kho, Jason A. Hodges, Mark R. Hansen, and Hugo O. Villar

*J. Med. Chem.*, **2005**, 48 (21), 6671-6678 • DOI: 10.1021/jm050564j • Publication Date (Web): 21 September 2005

Downloaded from <http://pubs.acs.org> on March 29, 2009



### More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 2 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

[View the Full Text HTML](#)

## Ring Systems in Mutagenicity Databases

Richard Kho,\* Jason A. Hodges, Mark R. Hansen, and Hugo O. Villar

Altoris, Inc., 11575 Sorrento Valley Road, Suite 214, San Diego, California 92121

Received June 15, 2005

The distribution of ring systems in public mutagenicity databases is analyzed. An automated enumeration of substructures permits determination of the occurrence of different scaffolds in data sets. The counts are used to perform population analysis via proportions and odds ratios of mutagenic compounds. Pairwise calculations of odds ratios between scaffolds allow comparison of ring systems for isostere replacement studies. These findings are presented in tables that readily show which scaffold is likely to occur in mutagenic compounds. Also, rings identified in public domain mutagenicity data sets are compared to rings in drugs data sets; unfortunately, public mutagenicity data sets do not reflect the types of scaffolds in drugs and those typically used in medicinal chemistry. The findings bring into question the utility of predictive models that were derived from public domain data sets. The automated ring identification and statistical approaches used here can be applied to other pharmacological properties to yield information about chemical scaffolds.

### Introduction

As a public health concern, the early detection of mutagens in the environment and the food supply chain is critical because industrial pollutants, pesticides, and certain toxicants can have detrimental effects.<sup>1</sup> Regulatory agencies routinely monitor for mutagens and toxicants to protect the public from potentially harmful chemicals.<sup>2</sup> In drug discovery, the identification of potential mutagens early in the discovery process is also critical because mutagenicity would constitute an undesirable toxicological profile leading to potentially harmful effects. Positive outcomes in mutagenicity tests would likely raise concerns in regulatory authorities and cause the discontinuation of further work in the chemical class.

The Ames test<sup>3–5</sup> has been a favored short-term in vitro assay aimed at detecting the genetic damage caused by chemicals and as a predictor of carcinogenicity. Despite all its limitations, the positive predictive power of the Ames test for carcinogenicity in rodents ranges from 77% to 90%.<sup>6</sup> In practice, the Ames test is not the final end-point for determining carcinogenicity because it is typically followed by a battery of in vitro and in vivo genotoxicity assays.<sup>7,8</sup> Nevertheless, the Ames test has become an important tool for weighing whether a chemical series should be advanced in the preclinical track. Its relative simplicity compared to other mutagenicity or carcinogenicity determinations makes it a good fit in current drug discovery paradigms, which are mostly concerned with the evaluation of a large number of chemicals in the early stages of the project.<sup>9</sup> Quick and simple assays have gained relevance in a drug discovery workflow characterized by a myriad of potential chemicals from which few are selected for further development.

The Ames protocol is based on inducing growth in genetically altered strains of the bacterium *Salmonella*

*typhimurium* that lack the capability of synthesizing histidine and need external sources of this amino acid for growth.<sup>3</sup> When the bacteria are exposed to certain mutagens, the *Salmonella* can undergo mutations that restore their ability to produce histidine, therefore growing without histidine in the medium. Because the mutant bacteria revert to their original character, these bacteria are sometimes referred to as revertants. Several strains of *Salmonella* are used when evaluating the mutagenicity of compounds because they reflect different mechanisms by which revertants become mutated. The TA-98 and TA-100 strains are the most commonly used, but this panel of strains is not considered to be fully satisfactory, and current guidelines suggest including the TA-1535 and TA-1537 strains as well.<sup>10,11</sup> The use of multiple strains to check the mutagenic character of chemicals is done to increase the sensitivity of the assays as well as its range of applicability.

Many chemicals are not inherently mutagenic but are transformed into mutagens by metabolic activation. The practitioners of the Ames test utilize a metabolic activation mixture to mimic this aspect of the mutagenic potential of a chemical due to in vivo metabolism. Thus, each chemical is tested in two formats per strain, with and without metabolic activation. On the basis of the four aforementioned strains, a total of eight determinations are recommended to test the mutagenic character of a chemical.

As with every other aspect of toxicology, in silico prediction<sup>12</sup> or derivation of simple rules that may bias compound selection against mutagenic chemicals has traditionally been and continue to be of great practical interest.<sup>12–17</sup> In drug discovery, such rules could help to better define the likelihood of success of a chemical. The methods for predicting mutagenicity and observations made on the properties of mutagenic compounds rely on the modeling of existing data sets. Because those data sets were compiled over long periods of time for disparate reasons, assay conditions are variable and no motif exists for how the number and variety of chemicals

\* To whom correspondence should be addressed. Phone: 858-259-8161. Fax: 858-259-8162. E-mail: rkho@altoris.com.

were selected for study. The lack of homogeneous data and the wide-ranging nature of the chemicals could be problematic when attempting to derive general conclusions or model the data. For that reason, we have two goals in this study. First, we want to summarize the information available in Ames test databases by analyzing the types of ring systems that are present in the available data sets and studying their frequency in Ames positive and negative compounds. The study of fragments prevalent in mutagenic compounds has attracted considerable attention over the years and has been at the center of the development of techniques to predict potential mutagens. Historically, methods such as CASE and MULTICASE have pioneered the knowledge-based approach for predictive purposes.<sup>15,18</sup> Our emphasis is not on predictive algorithms<sup>19,20</sup> but instead on organizing the currently available data sets in terms that can be readily useful to chemists. We identified simple scaffolds present in the data sets and organized them in a hierarchy according to their complexity. The parent and child relationships that exist in the set were analyzed, and changes in mutagenic character according to scaffold complexity were observed.

Our second goal is to evaluate the relevance of the currently available mutagenicity databases for work in drug design. The available literature suggests that the most common algorithms to predict mutagenicity do not perform particularly well in the case of drugs.<sup>17</sup> We decided to investigate whether the problems are due to the types of structures that are being used to train the available algorithms. To that end, we compared the frequency and distribution of ring systems in a small data set of marketed drugs to the chemicals in the largest toxicological database. In other words, we wanted to determine if the chemicals evaluated for mutagenicity are representative of the types of chemical substructures that are found in drugs. As before, we employed fragmentation of chemicals according to the ring systems they contain and organized the results by complexity. We then compared the scaffolds present in the mutagenicity databases and the drugs databases. Our results strongly suggest that the chemical diversity of compounds evaluated in mutagenicity assays is significantly less than the diversity of marketed drugs. For that reason, methods based on public domain data may be insufficient to derive conclusions that would be valuable for drug design. Since larger private data sets that better represent druglike compounds likely exist, we describe the methods and statistical approaches that can be used to analyze them.

## Methods

**Data Sets.** The CCRIS (Chemical Carcinogenesis Research Information System) database was used for this study following methods detailed in prior studies.<sup>19,20</sup> All results for the four main test strains were retrieved (TA-98, TA-100, TA-1535, and TA-1537), including those with or without metabolic activation using the rat liver S9 mix protocol. The classification of compounds as Ames positive or negative is not straightforward because there are ambiguities in the data reported through time by different laboratories. To improve accuracy, only molecules that had a consistent outcome in at least 80% of the studies were taken into

consideration. The CCRIS database lacks structural data for the compounds; we obtained structural information by merging the CCRIS data with the structures contained in other databases, using the CAS number as a common reference.<sup>19</sup> This step reduced the number of compounds available for study, since we could not identify structures or CAS numbers for all of the compounds in the CCRIS data set. The final data set used for this study contained 6039 compounds with at least one Ames test result and a corresponding chemical structure. The data set can be downloaded from our Web site (<http://www.altoris.com>).

For comparison purposes, we compiled in-house a set of 3882 commercial drugs. Among others, it contains compounds found in the "To Market, To Market" chapters of the *Annual Reports in Medicinal Chemistry*.<sup>21</sup> The data set contains information on a wide assortment of different pharmacological classes, therapeutic activities, and routes of delivery.

**Computational Method for Scaffold Identification.** The data sets were analyzed using our program, SARvisionPlus 1.5 (<http://www.chemapps.com>, ChemApps, San Diego, CA). The program carries out an enumeration of molecular fragments frequently found in the database. The scaffold perception algorithm does not resort to the use of predefined lists; instead, it carries out an exhaustive comparison of all molecules.<sup>22</sup> SARvision applies a series of knowledge-based rules to reduce the total number of scaffolds considered to those that are chemically meaningful, which in turn increases the efficiency of the algorithm. Among the rules, only entire ring structures are counted as scaffolds rather than ring or functional group fragments, and differences in simple alkyl substitutions are insufficient to define a new scaffold. These two simple restrictions greatly reduce the total number of scaffolds and consequently speed up the analysis but without affecting the results. The rules have the advantage of defining substructures that are consistent with chemical intuition and simplifying the subsequent analysis of the data.

One of the major challenges when substructural analysis is carried out is the organization and presentation of the different scaffolds or fragments found. Specifically, the representation of relationships among them can be cumbersome. SARvision utilizes a hierarchical organization of the molecular fragments and presents the hierarchy as a tree, allowing facile identification of the relationships between scaffolds. In this respect, the program is similar to another recently described application,<sup>23</sup> as it organizes molecular fragments in a hierarchy according to complexity. The simplest structures are at the highest level of the tree, and other, more complex superstructures are placed as leaves in the tree, resulting in easy-to-navigate parent and child relationships. Since the program requires nontrivial additions to differentiate among scaffolds, branching points in the tree are only found when nontrivial variants of the parent scaffold exist. Moreover, many scaffolds appear only in a small number of molecules. Throughout this study, we report on scaffolds that occur in a minimum of five molecules. The program provides a count of the number of occurrences of a particular scaffold at a tree node or leaf.

**Statistical Analysis of the Proportions.** For the analysis of the distribution of scaffolds, we resorted to some simple tools of population and categorical data analysis: analysis of proportions and odds ratio. The *proportion* of Ames negative compounds is defined as

$$p = \frac{n_{\text{neg}}}{N}$$

where  $n_{\text{neg}}$  is the number of compounds found to be Ames negative in all of the assays for all strains and  $N$  is the total number of observations made. Note that the proportion is a value between 0 and 1, inclusive. The standard error of the proportion at 5% significance level is calculated as

$$\text{SE} = \pm 1.96 \sqrt{\frac{(1-p)p}{N}}$$

The 95% confidence interval can then be determined for the proportion using the SE. If a proportion of 0.5 is contained in the confidence interval, then the same proportion of positive or negative compounds can be expected for that scaffold.

The odds ratio (OR), commonly used in epidemiological and biomedical research, is a statistical measure useful for describing the relationship between two populations.<sup>24,25</sup> As its name indicates, the odds ratio is simply the ratio of the odds of two events. In this analysis, the odds of observing a negative outcome refers to the number of negative outcomes ( $n_{\text{neg}}$ ) divided by number of non-negative outcomes. The odds ratio between two cohorts of compounds ( $\text{OR}_{1,2}$ ) can be calculated from

$$\text{OR}_{1,2} = \frac{\left[ \frac{n_{\text{neg}}}{N - n_{\text{neg}}} \right]_1}{\left[ \frac{n_{\text{neg}}}{N - n_{\text{neg}}} \right]_2}$$

where  $N$  is the total number of observations. The 95% confidence interval for the OR can be calculated at the 5% significance level using

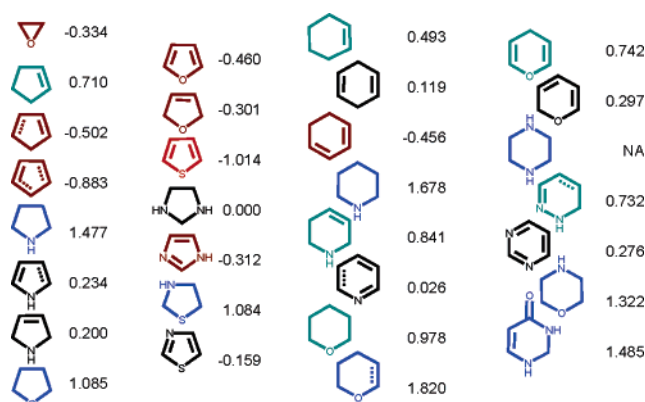
$$\ln \text{OR} \pm \sqrt{[1/n_{\text{neg}} + 1/(N - n_{\text{neg}})]_1 + [1/n_{\text{neg}} + 1/(N - n_{\text{neg}})]_2}$$

The antilogarithms provide the confidence intervals.

An OR of unity would indicate equal odds of negative or positive outcomes. However, the OR is not a symmetric function around  $\text{OR} = 1$ , which makes the assimilation of results less immediate. For this reason, we adopted the use of the decimal logarithm (log to the base 10) of the odds ratio. In this way, the scale becomes symmetric with respect to the equal odds ( $\log \text{OR} = 0$ ); a positive log OR value indicates greater odds of having an Ames negative outcome, and a negative log OR value indicates greater odds of having an Ames positive outcome (see results; for example, use of log OR). For the statistical analysis, compounds that showed a positive outcome in any of the eight Ames tests were counted as positive and compounds that had no positive results reported were counted as negative.

## Results and Discussion

**Population Analysis for Simple Scaffolds.** The program used to identify scaffolds organizes the results



**Figure 1.** Simple scaffolds identified in mutagenicity databases. Scaffolds are colored according to a log OR gradient, where the log OR ( $\log_{10}$  odds ratio) was calculated for Ames negative versus Ames positive counts. Blue scaffolds have most Ames negative character; black scaffolds contain equal odds of being in Ames negative or Ames positive compounds; and red scaffolds have higher odds of being in Ames positive compounds. The actual log OR values are shown for each scaffold. A large positive number (blue scaffold) indicates the most Ames negative character. Scaffolds are arranged by increasing complexity (from upper left).

hierarchically according to parent and child relationships. The highest level of the tree structure is composed mostly of simple heterocycles. Figure 1 shows some of these substructures and the log OR values for Ames negative versus Ames positive results.

If the value of the log OR is positive, the odds of the scaffolds being in compounds reported to be Ames negative is larger than that of the scaffold being part of Ames positive compounds. That is, positive values are associated with scaffolds that are less likely to be found in mutagenic compounds. For example, thiophene has a negative log OR ( $-1.014$ ); therefore, the scaffold is more likely to be found in compounds that fail the Ames test, and the value reveals that the odds of this scaffold being found in an Ames positive compound are greater than 10 to 1. At the other end of the spectrum, no compound containing piperazine as a scaffold was found to be Ames positive. This is the only case where a parent scaffold shows no Ames positive result. Compounds containing piperidine or morpholine are also unlikely to be mutagenic, having log OR values of 1.678 or 1.322, respectively. This is equivalent to saying that the odds of finding these scaffolds in Ames negative compounds are 48 to 1 or 21 to 1, respectively. Some heterocycles such as imidazolidine and pyrrole are found with comparable odds in active and inactive compounds. A point to keep in mind is that these results are not predictive, but simply summarize the observations in the data set analyzed, the CCRIS database. If the data set contains biases, then the analysis will also reflect those biases.

A common problem in medicinal chemistry is that of substitutions and replacements using bioisosteres. The replacement of a given functional group or scaffold is a common practice to improve the activity or to change the toxicological, physicochemical, or pharmacokinetic properties of a compound. Pairwise comparisons of scaffolds can consequently have more practical use than comparing the odds of positive versus negative compounds in a population as in Figure 1. For that purpose,




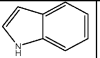




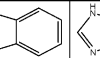





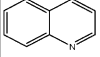
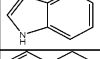
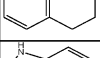
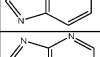
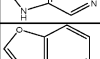
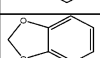
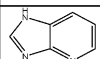
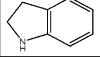
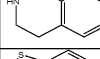
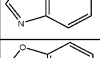
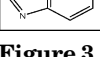
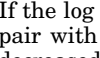
	0.0000	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
	1.9682	0.0000	↑	↑	↑	0.1119	↑	←	←	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
	2.3822	0.4139	0.0000	↑	-0.1027	←	↑	←	←	←	←	↑	0.0802	↑	↑	↑	↑	↑	↑	↑
	4.0143	2.0461	1.6321	0.0000	←	←	0.3662	←	←	←	0.3544	←	-0.2246	←	←	←	←	←	←	↑
	2.4849	0.5167	0.1027	↑	0.0000	←	↑	←	←	←	↑	0.183	↑	↑	↑	↑	-0.383	↑	-0.1382	↑
	1.8563	-0.1119	↑	↑	↑	0.0000	↑	0.383	-0.0344	-0.1278	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
	3.6481	1.6798	1.2659	-0.3662	←	←	0.0000	←	←	←	-0.0118	←	↑	←	0.3064	←	←	←	←	↑
	1.4733	-0.4949	-0.9089	-2.541	-1.0116	-0.383	-2.1748	0.0000	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑
	1.8907	-0.0775	↑	↑	↑	0.0344	↑	←	←	←	0.0000	-0.0934	↑	↑	↑	↑	↑	↑	↑	↑
	1.9841	0.0159	↑	↑	↑	0.1278	↑	←	0.0934	0.0000	↑	-0.3178	↑	↑	↑	↑	↑	↑	↑	↑
	3.6599	1.6917	1.2777	-0.3544	←	←	0.0118	←	←	←	0.0000	←	-0.579	←	0.3182	←	←	←	←	↑
	2.3019	0.3337	-0.0802	↑	-0.183	←	↑	←	←	←	0.3178	↑	0.0000	↑	↑	↑	↑	↑	↑	↑
	4.2389	2.2707	1.8568	0.2246	←	←	←	←	←	←	0.579	←	0.0000	←	←	←	←	←	←	↑
	3.0239	1.0557	0.6417	-0.9904	0.539	1.1676	-0.6242	1.5506	1.1332	1.0398	-0.636	0.722	-1.215	0.0000	-0.3178	0.156	0.0177	0.4008	↑	-0.5108
	3.3417	1.3735	0.9595	-0.6726	0.8568	1.4854	-0.3064	1.8684	1.451	1.3576	-0.3182	1.0398	-0.8972	0.3178	0.0000	←	0.3355	←	↑	-0.193
	2.8679	0.8997	0.4857	-1.1464	0.383	1.0116	-0.7802	1.3946	0.9772	0.8838	-0.792	0.566	-1.371	-0.156	↑	-0.4738	0.0000	-0.1383	0.2448	↑
	3.0062	1.038	0.624	-1.0081	0.5213	1.1499	-0.6419	1.5329	1.1155	1.0221	-0.6537	0.7043	-1.2327	-0.0177	-0.3355	0.1383	0.0000	0.3831	↑	-0.5285
	2.6231	0.6548	0.2409	↑	0.1382	←	↑	←	←	←	↑	←	←	-0.4008	↑	-0.2448	-0.3831	0.0000	↑	↑
	5.3116	3.3433	2.9294	1.2973	2.8267	3.4553	1.6635	3.8383	3.4208	3.3274	1.6517	3.0096	1.0726	2.2877	1.9699	2.4437	2.3054	2.6885	0.0000	←
	3.5347	1.5665	1.1526	-0.4796	1.0498	1.6784	-0.1133	2.0614	1.644	1.5506	-0.1252	1.2328	-0.7042	0.5108	0.193	←	←	←	↑	0.0000

**Figure 2.** Pairwise comparisons of individual monocyclic ring scaffolds for log odds ratio of Ames negative to Ames positive compounds. If the log OR confidence interval does not include the equal odds event (log OR = 0), then an arrow points to the scaffold of the pair with higher odds of being found in Ames negative compounds; that is, the scaffold that should be preferred in terms of decreased mutagenic potential.

an interesting way to analyze the data is to look at the odds ratio of Ames outcomes for pairs of scaffolds. The odds that a compound containing a given scaffold shows no Ames positive result for any strain compared to another scaffold is directly applicable to the problem of identifying more suitable bioisosteres.

A two-way entry table (Figures 2 and 3) can be built using such pairwise comparisons of scaffolds. The log OR for any two given scaffolds can be calculated, and a preference in terms of mutagenic potential can be

determined for each scaffold relative to the other. If the confidence interval does not contain the equal odds value (log OR = 0), then the log OR is statistically significant ( $p \leq 0.05$ ) and one of the scaffolds is preferable over the other, with all other considerations being equal. The preference is indicated in Figures 2 and 3 by an arrow that points to the scaffolds that have favorable odds. Figure 2 shows pairs of simple, one-ring scaffolds, while Figure 3 compares pairs of bicyclic scaffolds. If the confidence interval for the log OR

												
	0.0000	↑ -0.3896	↑ -0.351	← 0.2847	-0.1965	-0.1840	-0.1840	← 0.7358	↑ -1.8291	↑ -0.597	↑ -1.0182	↑ -3.0331
	← 0.3896	0.0000	0.0386	← 0.6743	0.1932	0.2056	0.2056	← 1.1254	↑ -1.4395	-0.2074	↑ -0.6286	↑ -2.6435
	← 0.351	-0.0386	0.0000	← 0.6357	0.1546	0.1671	0.1671	← 1.0868	↑ -1.4781	-0.2460	↑ -0.6672	↑ -2.6821
	↑ -0.2847	↑ -0.8743	↑ -0.6357	0.0000	↑ -0.4811	↑ -0.4687	↑ -0.4687	0.4511	↑ -2.1138	↑ -0.8817	↑ -1.3029	↑ -3.3178
	0.1965	-0.1932	-0.1546	← 0.4811	0.0000	0.0125	0.0125	← 0.9323	↑ -1.6327	-0.4006	↑ -0.8218	↑ -2.8367
	0.1840	-0.2056	-0.1671	← 0.4687	-0.0125	0.0000	0.0000	← 0.9198	↑ -1.6452	-0.4130	↑ -0.8342	↑ -2.8491
	0.1840	-0.2056	-0.1671	← 0.4687	-0.0125	0.0000	0.0000	← 0.9198	↑ -1.6452	-0.4130	↑ -0.8342	↑ -2.8491
	↑ -0.7358	↑ -1.1254	↑ -1.0868	-0.4511	↑ -0.9323	↑ -0.9198	↑ -0.9198	0.0000	↑ -2.5649	↑ -1.3328	↑ -1.754	↑ -3.7689
	← 1.8291	← 1.4395	← 1.4781	← 2.1138	← 1.6327	← 1.6452	← 1.6452	← 2.5649	0.0000	← 1.2321	← 0.8109	↑ -1.204
	← 0.597	0.2074	0.2460	← 0.8817	0.4006	0.4130	0.4130	← 1.3328	↑ -1.2321	0.0000	-0.4212	↑ -2.4361
	← 1.0182	← 0.6286	← 0.6672	← 1.3029	← 0.8218	← 0.8342	← 0.8342	← 1.754	↑ -0.8109	← 0.4212	0.0000	↑ -2.0149
	← 3.0331	← 2.6435	← 2.6821	← 3.3178	← 2.8367	← 2.8491	← 2.8491	← 3.7689	← 1.204	← 2.4361	← 2.0149	0.0000

**Figure 3.** Pairwise comparisons of individual bicyclic ring scaffolds for log odds ratio of Ames negative to Ames positive compounds. If the log OR confidence interval does not include the equal odds event (log OR = 0), then an arrow points to the scaffold of the pair with higher odds of being found in Ames negative compounds; that is, the scaffold that should be preferred in terms of decreased mutagenic potential.

contains zero, then there is no reason to prefer one of the scaffolds over the other. It should be pointed out that if larger numbers of observations were available for statistical analysis, the confidence intervals would be narrower and fewer comparisons would include log OR = 0 unless there was truly an equal odds of the scaffold occurring in Ames positive and Ames negative compounds.

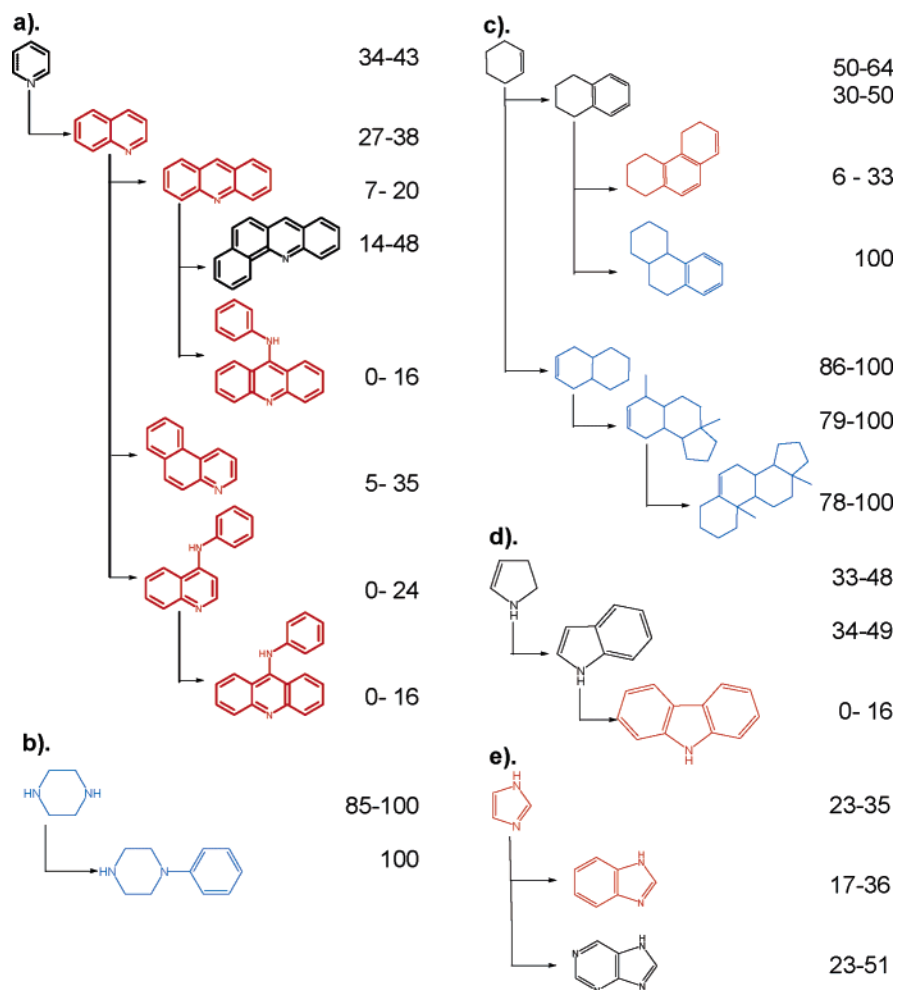
Some results can be quickly visualized using Figures 2 and 3. For example, any other ring should be preferred over aziridine (Figure 2, first datum row), since the log OR between it and all other scaffolds are negative values. Epoxides (Figure 2, second data row) are also quite commonly found in mutagenic compounds, but the odds of finding compounds that are mutagenic that contain it are less than the odds for compounds containing either thiophene (Figure 2, eighth data row) or aziridine. Therefore, in the absence of other discriminating features, epoxides should be preferred over aziridinyl or thiophenyl substituents. In the case of the bicyclic scaffolds (Figure 3), the data sets are sparser and, consequently, the confidence intervals are wider. There is less chance that the data will show preference for one scaffold over another. However, the benzoxazole (Figure 3, last data row) and benzothiazole (Figure 3, second to last data row) rings are less likely to be part of mutagenic compounds.

A comparison of the properties of pairs of scaffolds as shown in Figures 2 and 3 has clear applications for the replacement of functional groups or scaffolds in medicinal chemistry research. They are simple to interpret, and the knowledge can be readily applied. The results are not necessarily predictive, but they sum-

marize the properties of compounds in biologically meaningful ways. While the property studied here is the Ames mutagenicity test, similar tables can be constructed for other properties, providing an alternative way to look at chemical and biological data simultaneously.

**Increasing Complexity of Scaffolds.** As the complexity of the scaffolds increases, some correlations can be observed between the size of the scaffolds and the Ames outcome. Figure 4 shows a few examples of parent and child relationships where expanding upon the parent scaffold changes the proportions of Ames positive compounds. The analysis is performed using proportions for simplicity. The proportions indicate the ratio of Ames negative compounds relative to the entire data set and are presented as a percentage: a proportion of 50 indicates that the scaffold is found equally in Ames negative or positive compounds, while a proportion greater than 50 indicates that the scaffold occurs in more Ames negative compounds and vice versa. The interval of the confidence will depend on the number of observations accumulated.

The proportion of Ames negative compounds for each scaffold could change significantly depending on the substituents and the resulting topology of the ring system. For example, going from a pyridine (Figure 4a) to a quinoline (first child) changes the confidence interval of the proportion of Ames negative compounds from 34% to 43% to a range of 27–38%, indicating that the quinoline scaffold is more likely to be found in Ames positive compounds than the parent pyridine. At one lower level on the tree is acradine, with a proportion in the range of 7–20%, meaning that the addition of



**Figure 4.** Examples of parent and child relationships in mutagenicity databases. Compounds where the proportion of Ames negative to Ames positive counts is high are colored blue. Black scaffolds represent equal proportions, and red scaffolds represent higher proportions of Ames positive outcome. Numbers indicate the confidence interval of proportions of each scaffold: (a) pyridine, (b) piperazine, (c) hexanaphthylene, (d) pyrroline, and (e) imidazole branches of the tree.

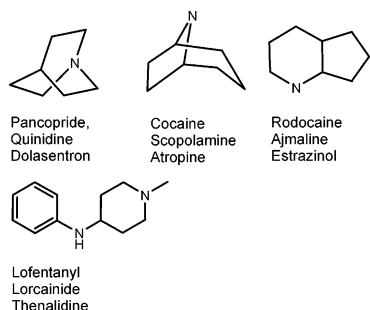
another benzene substituent seems to increase the proportion of Ames positive compounds in the set.

In other instances, the extension of a parent scaffold can lead to more complex scaffolds that do not meaningfully change the proportion of Ames negative compounds. The piperazine ring (Figure 4b), for example, has an *N*-phenylpiperazine as a child leaf that does not significantly change the Ames outcome. More interesting is the tetrahydronaphthalene branch (Figure 4c, first child), where depending on the substituents, a more Ames positive (6–33%) or more Ames negative (100%) scaffold results. As a general rule, increasing aromaticity or extension of the conjugation of chemicals increases the odds that the compounds containing those scaffolds will become mutagenic. Conversely, an increase in the aliphatic character of the ring results in a decrease in the mutagenic potential of the compounds. Overall, except for those simple trends, the relationships between larger scaffolds and the substructures they contain do not appear to have strong correlations to Ames test results.

**Relevance of the Substructure Set in Drug Design.** The above results summarize the observations in the CCRIS data sets. Because this database was not compiled for drug discovery purposes, its relevance to that end is unclear. Indeed, most studies in the past

have relied on the available data sets including the CCRIS without addressing the question of the relevance of the data set, particularly when the results will be applied to guide drug discovery. The applicability of the CCRIS database can be evaluated by comparing the substructures present in it to the types of substructures used in drug discovery work. For that purpose, we compiled a set of 3882 commercial drugs and compounds that underwent clinical trials and applied the same algorithms used in the fragmentation of the CCRIS data set.

The drugs compilation resulted in 750 ring systems represented by at least five molecules. The Ames data set contained only 427 ring systems. This result in itself is revealing because it shows that a smaller set of molecules (drug data set) contains a much larger number of ring systems. The two sets only have 199 ring systems in common, which is only a fraction of the rings identified in drugs. Figure 5 shows examples of scaffolds found in the drugs database but not in the Ames set. The reciprocal is less significant in our case given the small size of drug collection used. An interesting point to note is that natural products such as opiates and taxanes provide a large number of ring substructures. The presence of complex natural products in drugs accounts for the large number of rings found in the



**Figure 5.** Examples of scaffolds found in the analysis of the drugs database that are not found with sufficient frequency in the publicly available mutagenicity data sets. The names of a few drugs that contain the scaffold are provided.

drugs compilation. The CCRIS data set also has natural products, including some opiates, but in smaller numbers. Since we are restricting the number of rings analyzed by requiring that they be present in at least five molecules in the set, the scaffolds from natural products are less prominent in the CCRIS data set.

## Conclusions

Our aim was to summarize the information available in the CCRIS database, focusing on the ring systems present. Other prior studies have focused on the global properties of the molecules and on the influence of simple functional groups. Analysis of the distribution of ring systems in the CCRIS database shows that some scaffolds have lower odds of being part of mutagenic compounds. In particular, aliphatic rings are less likely to be part of mutagenic compounds, which agrees with previous studies.

The study of scaffold distributions using SARvision software provides a direct method for summarizing results in large data sets, which have become the norm in the early stages of drug discovery. In the past, the study of the distribution of substructures was limited to analysis of proportions and frequencies. Population analysis provides the tools necessary to more thoroughly exploit the information in fragment data sets. The study was enabled by progress made in scaffold detection algorithms, where exhaustive enumeration of scaffolds was combined with some simple knowledge-based rules to work only with complete scaffolds that are sufficiently different from others in the set. Previously, the large number of scaffolds would have made the analysis and the presentation of results impractical. The use of scaffold rules makes the process feasible because it reduces the total number of scaffolds to be analyzed, making the computational requirements manageable.

In addition, statistical approaches to analyze categorical data that have not been exploited in cheminformatics work were illustrated. The use of odds ratios and proportions for population analysis of chemical substructures is promising. Population-based studies are not as widely used in cheminformatics or for derivation of structure–activity relationships, but we show that these statistics can provide powerful alternative means to analyze chemical–biological data. Odds ratios allowed the pairwise comparison of rings in the data set, as well as the odds that the molecules containing the scaffold are mutagenic. In particular, the pairwise odds

ratios can be valuable when searching for replacement groups and their use can be extended to other sets of data beyond mutagenicity. Population analysis goes beyond the conventional frequency studies for functional groups and can be a very powerful way to summarize the data. The approach outlined to summarize large data sets and create simple tables is immediately applicable to medicinal chemistry work.

Our analysis also reveals that the distribution of scaffolds in a small set of drugs is different from the fragments identified in the CCRIS data set because there is only a limited overlap in the rings present in both data sets. These results may account for the lack of sensitivity shown by currently available methods in the identification of mutagenic pharmaceuticals. The result brings into question the practice of using public domain data sets to derive predictive models without examining their relevance for drug design. It may be necessary to construct such data sets more carefully in such a way to represent the greater diversity of types of scaffolds found in drug data sets. Ideally, the frequency and distribution of the different scaffolds should also mimic what is found in drug and druglike molecules used in medicinal chemistry research so that improved predictive models can be developed.

## References

- (1) Luch, A. Nature and nurture. Lessons from chemical carcinogenesis. *Nat. Rev. Cancer* **2005**, *5*, 113–125.
- (2) MacGregor, J. T.; Casciano, D.; Muller, L. Strategies and testing methods for identifying mutagenic risks. *Mutat. Res.* **2000**, *455*, 3–20.
- (3) Josephy, P. D.; Cruz, P.; Nohmi, T. Recent advances in the construction of bacterial genotoxicity assays. *Mutat. Res.* **1997**, *386*, 1–23.
- (4) Maron, D. M.; Ames, B. N. Revised methods for the Salmonella mutagenicity test. *Mutat. Res.* **1983**, *113*, 173–215.
- (5) McCann, J.; Ames, B. N. Detection of carcinogens as mutagens in the Salmonella/microsome test: assay of 300 chemicals: discussion. *Proc. Natl. Acad. Sci. U.S.A.* **1976**, *73*, 950–954.
- (6) Mortelmans, K.; Zeiger, E. The Ames Salmonella/microsome mutagenicity assay. *Mutat. Res.* **2000**, *455*, 29–60.
- (7) Fenech, M. The in vitro micronucleus technique. *Mutat. Res.* **2000**, *455*, 81–95.
- (8) Marzin, D. New approaches to estimating the mutagenic potential of chemicals. *Cell Biol. Toxicol.* **1999**, *15*, 359–365.
- (9) Atterwill, C. K.; Wing, M. G. In vitro preclinical lead optimisation technologies (PLOTs) in pharmaceutical development. *Toxicol. Lett.* **2002**, *127*, 143–151.
- (10) Kirkland, D. J.; Gatehouse, D. G.; Scott, D.; Cole, J.; Richold, M., Eds.; *Basic Mutagenicity Tests: UKEMS Recommended Procedures*; Cambridge University Press: Cambridge, U.K., 1990.
- (11) Gatehouse, D.; Haworth, S.; Cebula, T.; Gocke, E.; Kier, L.; Matsumura, T.; Melcion, C.; Nohmi, T.; Ohta, T.; Venitt, S.; Zieger, E. Recommendations for the performance of bacterial mutation assays. *Mutat. Res.* **1994**, *312*, 217–233.
- (12) Dearden, J. C. In silico prediction of drug toxicity. *J. Comput.-Aided. Mol. Des.* **2003**, *17*, 119–127.
- (13) Kazius, J.; McGuire, R.; Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.* **2005**, *48*, 312–320.
- (14) Patlewicz, G.; Rodford, R.; Walker, J. D. Quantitative structure–activity relationships for predicting mutagenicity and carcinogenicity. *Environ. Toxicol. Chem.* **2003**, *22*, 1885–1893.
- (15) Klopman, G.; Zhu, H.; Fuller, M. A.; Saiakhov, R. D. Searching for an enhanced predictive tool for mutagenicity. *SAR QSAR Environ. Res.* **2004**, *15*, 251–263.
- (16) Cariello, N. F.; Wilson, J. D.; Britt, B. H.; Wedd, D. J.; Burlinson, B.; Gombur, V. Comparison of the computer programs DEREK and TOPKAT to predict bacterial mutagenicity. Deductive estimate of risk from existing knowledge. Toxicity prediction by computer assisted technology. *Mutagenesis* **2002**, *17*, 321–329.
- (17) Snyder, R. D.; Pearl, G. S.; Mandakas, G.; Choy, W. N.; Goodsaid, F.; Rosenblum, I. Y. Assessment of the sensitivity of the computational programs DEREK, TOPKAT, and MCASE in the



- prediction of the genotoxicity of pharmaceutical molecules. *Environ. Mol. Mutagen.* **2004**, *43*, 143–158.
- (18) Rosenkranz, H. S.; Klopman, G. New structural concepts for predicting carcinogenicity in rodents: an artificial intelligence approach. *Teratog., Carcinog., Mutagen.* **1990**, *10*, 73–88.
- (19) Llorens, O.; Perez, J. J.; Villar, H. O. Toward the design of chemical libraries for mass screening biased against mutagenic compounds. *J. Med. Chem.* **2001**, *44*, 2793–2804.
- (20) Llorens, O.; Perez, J. J.; Villar, H. O. Investigation of structural and electronic biases in mutagenic compounds. *Int. J. Quantum Chem.* **2002**, *88*, 107–117.
- (21) See, for example, the following. Hegde, S.; Carter, J. To Market, To Market—2003. In *Annual Reports in Medicinal Chemistry*; Doherty, A. M., Ed.; Elsevier Academic Press: Amsterdam, 2004; Vol. 39, pp 337–368.
- (22) Bone, R. G. A.; Villar, H. O. Exhaustive enumeration of molecular substructures. *J. Comput. Chem.* **1997**, *18*, 86–107.
- (23) Wilkens, S. J.; Janes, J.; Su, A. I. HierS: Hierarchical scaffold clustering using topological chemical graphs. *J. Med. Chem.* **2005**, *48*, 3182–3193.
- (24) *Foundations of Epidemiology*, 3rd ed.; Lilienfeld, D. E., Stolley, P. D., Eds.; Oxford University Press: New York, 1994.
- (25) Wolter, K. M. *Introduction to Variance Estimation*; Springer-Verlag: New York, 1985.

JM050564J