Substructure Mining Using Elaborate Chemical Representation

Jeroen Kazius,*[,†] Siegfried Nijssen,[‡,§] Joost Kok,[‡] Thomas Bäck,[‡,#] and Adriaan P. IJzerman[†]

Division of Medicinal Chemistry, Leiden-Amsterdam Center for Drug Research, Leiden University, P.O. Box 9502, Einsteinweg 55, 2300 RA Leiden, The Netherlands, and Algorithms Cluster, Leiden Institute of Advanced Computer Science, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands

Received September 2, 2005

Substructure mining algorithms are important drug discovery tools since they can find substructures that affect physicochemical and biological properties. Current methods, however, only consider a part of all chemical information that is present within a data set of compounds. Therefore, the overall aim of our study was to enable more exhaustive data mining by designing methods that detect all substructures of any size, shape, and level of chemical detail. A means of chemical representation was developed that uses atomic hierarchies, thus enabling substructure mining to consider general and/or highly specific features. As a proof-of-concept, the efficient, multipurpose graph mining system Gaston learned substructures of any size and shape from a mutagenicity data set that was represented in this manner. From these substructures, we extracted a set of only six nonredundant, discriminative substructures that represent relevant biochemical knowledge. Our results demonstrate the individual and synergistic importance of elaborate chemical representation and mining for nonlinear substructures. We conclude that the combination of elaborate chemical representation and Gaston provides an excellent method for 2D substructure mining as this recipe systematically explores all substructures in different levels of chemical detail.

INTRODUCTION

A key element in the rational design of compounds is knowledge of substructures with biological or physicochemical relevance. In pharmaceutical research, for instance, structure−activity/property relationships (SARs/SPRs) can aid synthesis decision, high-throughput screening library design, hit prioritization, lead optimization, and prioritization of pharmacological or toxicological assays. Traditionally, human experts have used experience to identify structural features that are responsible for biological effects. Structural alerts, for example, have been derived for the ability of compounds to cause mutations in DNA, i.e., mutagenicity.[1,2] As descriptors, substructures (or fragments) are both easy to interpret by toxicologists and chemists and readily applicable in the design of new compounds. Furthermore, several properties of compounds, such as mutagenicity, can be better explained with substructures than with other descriptors.[3,4]

Substructure mining algorithms systematically detect, or learn, substructures from a data set. Importantly, these algorithms can be combined with data mining methods to extract substructures that are statistically associated with biological or physicochemical properties. Though substructure mining can also be interpreted as a form of data mining, in this paper we handle the term data mining only for the extraction of those substructures that are most discriminative for mutagenicity. For mutagenicity in particular, substructure mining algorithms[5−8] and other methods[3,9−21] exist that classify and predict the mutagenicity of chemically diverse training and test sets with acceptable to satisfactory precision.[4,22−29] Current methods for substructure mining, however, consider limited chemical information with respect to substructure shape and atom type as they examine only small, linear fragments with default atom and bond types.[4−8] Furthermore, most methods are limited to relatively small data sets of hundreds of compounds.

There are at least two directions to further exploit the chemical information that exists in a set of molecules. First, next to using default atom and bond types to represent compounds, additional chemical information can be considered. Data can be attached to certain heteroatoms regarding, for instance, their formal charge or the number of connected hydrogens. Also, aromatic atoms or hydrogen donor atoms can be grouped. Such additional information enables subsequent substructure mining methods to consider both more general and more specific features. Second, graphs are more comprehensive than linear atom chains in terms of representing substructures of any two-dimensional shape because graphs can also contain branches and cycles. Efficiency is critical in the detection of substructures of any shape because, overall, there are many more nonlinear than linear fragments present in a compound. Recent developments in graph mining research have enabled the analysis of the complete repertoire of substructures that are present in a database of thousands of compounds.[30−34]

For these reasons, the overall aim of our study was to enable more exhaustive data mining by designing methods that detect all substructures of any size, shape, and level of

* Corresponding author phone: +31 (0)71 527 4513; e-mail: j.kazius@lacdr.leidenuniv.nl.

† Leiden-Amsterdam Center for Drug Research, Leiden University.

‡ Leiden Institute of Advanced Computer Science, Leiden University.

§ Current address: Institute for Computer Science, Machine Learning Lab, Freiburg University, Georges Köhler Allee 79, D-79110 Freiburg/Br., Germany.

NuTech Solutions, Martin-Schmeisser-Weg 15, 44227 Dortmund, Germany.

10.1021/ci0503715 CCC: $33.50
Published on Web 01/19/2006

chemical detail. The specific aims of the current study were as follows:

(1) to develop an elaborate method of graph-based chemical representation that increases the level of chemical detail considered in substructure mining analyses;

(2) to test the usefulness of an efficient graph-based substructure mining algorithm, named Gaston,[34,35] in learning substructures from a reasonably large mutagenicity data set;

(3) to automatically extract SARs, in the form of a compact set of nonredundant, information-rich substructures that are significantly related to mutagenicity, by subjecting the output of Gaston to a data mining method;

(4) to determine the benefits of using elaborate chemical representation and of considering nonlinear substructures in SAR extraction.

In the first section of this paper, we discuss how a mutagenicity data set was pruned prior to analysis. Subsequently, we examine the means of representing the compounds and the process of mining this data set for substructures. In the section thereafter, the method of data mining is described. Finally, the results of applying these methods to this data set are discussed, and conclusions are drawn about the significance of combining chemical representation, substructure mining, and data mining.

METHODS

Mutagenicity Data Set. To test the validity of the results of the current study we used a previously characterized data set. In our earlier study,[20] a large mutagenicity data set (4337 entries) was constructed from the Chemical Carcinogenesis Research Information System.[36] In this data set, each compound was categorized as a nonmutagen if only negative Ames test results were reported for it, while it was categorized as a mutagen if one or more positive Ames test results were available for it.[36] This data set predominantly contains organic compounds as well as several salts and mixtures. Prior to the present analysis, this data set was subjected to stricter criteria. First, mixtures were discarded entirely, and counterions were removed from salts. To exclude atypical entries such as proteins, we also removed those molecules with a molecular weight of over 500. Subsequently, the remaining compounds were converted into a canonical SMILES[37] format.[38−41] Stereoisomers were considered to be duplicate compounds since substructure mining algorithms search for two-dimensional fragments. Compounds with different formal charges, protonation states, isotopes, or resonance structures were also considered to be duplicates. Compounds that described duplicate canonical SMILES representations were removed. Application of these filters to the original mutagenicity data set resulted in a final subset of 4069 distinct compounds, of which 2294 were categorized as mutagens and 1775 as nonmutagens. For the purpose of comparison, the settings handled in the present study (settings of chemical representation, frequency constraints and statistics) were based on the corresponding settings used to determine general toxicophores in the previous analysis.[20,42]

Chemical Representation. In the present analysis, compounds of the mutagenicity data set were represented in two ways: as standard chemical graphs and, by using an extension to standard chemical notation, as elaborate graphs.

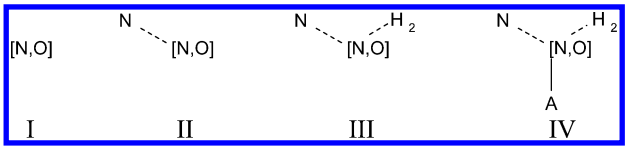

Figure 1. Construction of an atomic hierarchy. The first three substructures are descriptions of a single atom in different levels of chemical detail. The first substructure (**I**) describes a single general label ([N,O]) that detects either a nitrogen or an oxygen atom. The second substructure (**II**) contains this general label as well as a specific label (N) which indicates that the described atom is a nitrogen. The third substructure (**III**) shows a complete atomic hierarchy with an additional specific label (H_2) that indicates that this atom is a primary amine. The fourth substructure (**IV**) is nonlinear and describes two atoms: a single bond connects an aromatic atom, symbolized as a general label for aromatic atoms (A), to a primary amine, symbolized as the atomic hierarchy also shown in substructure **III**.

Graph-based representations of compounds commonly consist of atoms, which are symbolized as nodes labeled with default atom types (C, N, O, etc.), that are connected through bonds. These bonds, in turn, are symbolized as edges labeled as single, double, or triple (or sometimes aromatic). To compare the impact of elaborate chemical representation, which will be introduced in the following paragraphs, this standard chemical notation was also used to represent compounds for substructure mining.

Though an atom is commonly labeled with its atom type, a wildcard label can also be used to indicate the presence of any atom, irrespective of its atom type.[43,44] Here, however, atoms were also represented with atomic hierarchies, rather than with single labels for wildcards or atom types. Atomic hierarchies are small tree-shaped structures that consist of one central atom label, the root, to which further atom labels are attached. Figure 1 shows how an atomic hierarchy was constructed and used in a substructure of two atoms. We employed the OpenBabel C++ libraries[45] to design atom labels with user-defined SMARTS expressions.[46] In theory, any atom label can serve as either the root of a tree or as additional information. In this study, we chose to label every root of an atomic hierarchy with a general label. A general label describes a property that is shared by multiple atom types, such as aromatic atoms, halogens, acidic groups, or hydrogen donors or acceptors. Additional labels, called specifiers, describe more atom-specific chemical information, such as the atom type, its formal charge and the number of connected hydrogens. Specifiers are connected through virtual edges to the roots of atomic hierarchies. Finally, the root (including the extra information) replaces the original atom in the molecule. For instance, substructure **III** from Figure 1 describes an atomic hierarchy of a primary amine that consists of a general label and two specifiers. The most important advantage of using atomic hierarchies instead of standard atom types is the ability to include more general and/or very specific features in the substructure mining process. Like SMARTS patterns,[46] the resulting substructures describe different degrees of chemical detail.

The next paragraphs explain how an elaborate method of chemical representation was used to describe the data set of 4069 compounds. More specifically, during elaborate chemical representation two settings were handled because each setting enables the detection of substructures that could not be detected with the other setting. Figure 2 shows a

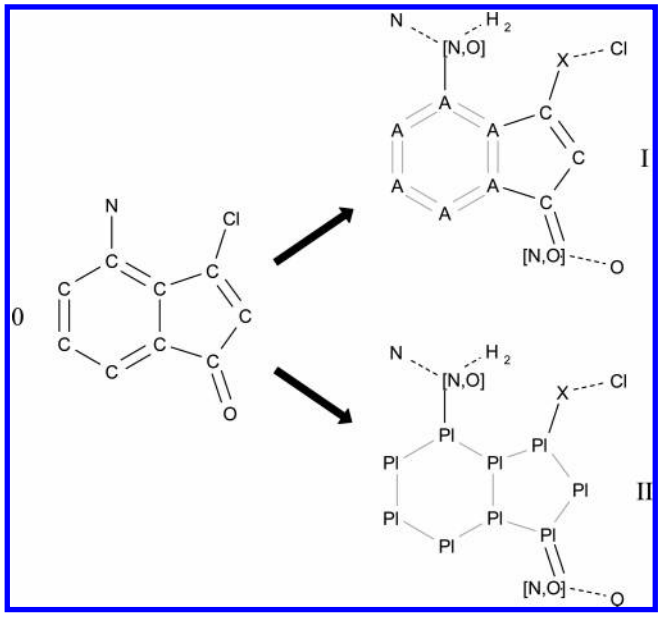

Figure 2. Elaborate chemical representation. A hypothetical compound in standard chemical notation (**0**) is represented according to two settings of elaborate chemical representation: aromatic (**I**) and planar (**II**). Different bond labels correspond to single, double, aromatic (grey double bonds in **I**), planar (grey single bonds in **II**), and virtually added bonds (dashed bonds in **I** and **II**), which can attach additional information. Various atom labels correspond to carbon (C), nitrogen (N), oxygen (O), small heteroatom ([N,O]), halogen (X), chlorine (Cl), aromatic atom (A), planar atom (Pl), and the number of implicit hydrogens (H_2).

hypothetical compound in standard chemical notation and in the corresponding chemical representations of each setting.

For both settings that were handled during elaborate chemical representation, **I** and **II**, the default rule was that atoms that did not satisfy any of the criteria mentioned below were represented by their atom type. Both settings also had in common that:

- aliphatic oxygen and nitrogen atoms were labeled as small heteroatoms with specifiers for their atom type and the number of connected hydrogens, as shown in Figures 1 and 2;
- aliphatic sulfur and phosphorus were labeled as large heteroatoms with atom type specifiers;
- chlorine, bromine, and iodine atoms were labeled as halogens with atom type specifiers, as shown in Figure 2.

In the first setting (**I**), named the aromatic setting, four bond types were handled: aromatic bonds and aliphatic single, double, and triple bonds. All aromatic atoms were labeled as aromatic, and specifiers for atom types were attached to aromatic heteroatoms. Examples of aromatic bonds and atoms are shown in chemical representation **I** in Figure 2.

In the second setting (**II**), named the planar setting, all aromatic atoms were labeled as planar atoms. Additionally, aliphatic atoms that were part of a planar five- or six-membered ring (see Supporting Information for the exact definition of planarity) were also labeled as planar atoms with specifiers for their atom type. All aliphatic bonds within such planar rings and all aromatic bonds were labeled as planar bonds. The planar setting thus introduces a new bond type for planar bonds. Examples of planar bonds and atoms are shown in chemical representation **II** in Figure 2. One effect of this setting is that the primary amine in Figure 2, which is connected to an aromatic atom in the aromatic setting (**I**), is connected to a planar atom in the planar setting (**II**). Planar atoms also describe several cyclic sp_2-hybridized aliphatic atoms. As a result, the latter substructure and the aromatic atom-bound amine will show different statistics. By using both settings during elaborate chemical representation such differences were also considered during subsequent data mining.

Substructure Mining. Substructure mining algorithms are designed to extract all substructures (or fragments) that satisfy a predefined constraint from a data set of compounds. As noted before, we handle the term data mining to describe a different step than substructure mining. For two-dimensional fragments, the following classes of structural complexity exist:

- paths, which denote simple, linear graphs (unbranched and noncyclic);
- trees, which represent branched graphs that do not contain a cycle;
- graphs, which can have any shape (including cycles and branches).

The computational work required to search for paths, trees, and graphs increases with their structural complexity. Focused search algorithms exist to efficiently search for fragments of each of these classes.[34] A novel graph-based substructure mining algorithm, named Gaston,[34,35] uses this knowledge to split up the substructure mining process into several phases by mining for all paths, trees, and graphs, respectively. To limit the output to those substructures that are of potential interest, constraints can be set by applying, for example:

- a minimum number of molecules that each substructure needs to detect (minimum support or frequency constraint) [This frequency constraint can also be expressed as the percentage of compounds of the entire data set.];
- a maximum number of atoms or bonds, set for either the number of atoms that belong to a predefined class or the total number of atoms (size constraint);
- a maximum structural complexity level that can limit the search to paths only, paths and/or trees, or all graphs, including paths, trees, and cyclic graphs (shape constraint or structural complexity constraint).

Substructure mining algorithms iteratively perform a step that consists of both substructure generation and the corresponding substructure search, thus determining which molecules this substructure detects. Typically, substructure mining starts with small substructures (single atoms) that are extended atom by atom until the extended substructure does not satisfy any of the predefined constraints. Figure 1 depicts four example substructures that a substructure mining method would detect if it analyzed compound **I** of Figure 2. In this process, many substructure mining algorithms[30−34,47] make use of the Apriori property[48] to detect all fragments that satisfy a frequency constraint. This property states that if a given substructure does not satisfy a frequency constraint, then no larger substructure that contains this substructure can satisfy this constraint (as it will detect an equal or lower number of compounds). The Apriori property has proved critical in substructure mining as it efficiently disregards infrequent substructures (and their larger derivatives). For instance, if according to the frequency constraint there are

too few compounds that contain oxygen, then peroxide-containing substructures are not explored.

Consecutive steps of substructure searches can be performed in a breadth-first or depth-first order. A breadth-first search first considers all substructures of the same size one by one, then enlarges all these, and considers all new substructures of one size bigger. In contrast, a depth-first search recursively considers and extends one small substructure before extending the next small substructure. A depth-first search is more memory efficient as, during each step of the analysis, only the data of a single substructure needs to be memorized, while a breadth-first search needs to store data of all substructures of the same size.

At the start of this process, the initial substructure is mapped everywhere it fits into every molecule in the data set. For each such mapping, the atoms at the neighboring positions of this substructure are stored. The corresponding atom and bond type information is then used to prevent computer-intensive searches for larger substructures that are infrequent or even nonexistent. Most methods perform duplicate substructure searches since a single substructure can be generated in many ways.[34,49] Gaston,[34,35] however, most efficiently prevents such duplicate searches.[49]

Four different substructure learning scenarios were considered in total: standard or elaborate chemical representation and with or without also considering nonlinear substructures. More specifically, after elaborate chemical representation of all compounds, Gaston was run individually for each setting, aromatic and planar, and the resulting substructures were fused into a single set of substructures. In the previous study,[20] a frequency constraint of 70 or more mutagens was handled during extraction of general toxicophores. Accordingly, for each scenario, Gaston was used to find all substructures that occurred in 70 or more mutagens[42] with no constraint on size.

Data Mining. In our analysis, the purpose of data mining was to extract a small set of substructures from the considerable set of information-rich substructures that resulted from substructure mining. Moreover, the extracted collection should contain nonredundant substructures that are discriminative for mutagenicity.

In the current study, we chose a data mining method similar to an approach for creating decision lists,[50] which are linear decision trees.[51,52] Each scenario produced a collection of substructures that was analyzed separately. For each substructure in such a collection, its statistical association with mutagenicity, expressed as the p-value, was determined from the amounts of mutagens and nonmutagens it detected in the complete chemical data set. It was then determined which substructure was most strongly associated with mutagenicity, that is, which substructure possessed the lowest p-value. This substructure was then selected to split the chemical data set into two subsets. Each split generated one subset of compounds that contain this substructure and another subset of compounds that lack this substructure. This latter subset was used to recompute the p-values of all substructures. From these p-values, the next most mutagenic substructure was determined and then used to split this chemical subset in two, and so on. After six splits for instance, all compounds from the original database are divided over seven subsets. In cases where multiple substructures have the lowest p-value, the largest substructure was selected to split the data set . For example, synonymous substructures can describe an aromatic primary amine because it can be connected to a single aromatic atom or a chain of two, three, four, or five aromatic atoms. In the present analysis, splits were made with the constraint that each selected substructure was required to detect a set of compounds of which over 60% were mutagenic.[42] If the p-value of the best selected substructure with respect to the analyzed chemical subset exceeded 10^{-20}, then it was not used to split the data set, and no further splits were made. As a result, the size of a trained decision list depends on the number of substructures that satisfy these two criteria.

Cross-validation is a standard statistical technique to test whether data mining methods can extract relevant patterns (in our case, substructures that predict mutagenicity). The complete data set was split into several parts of equal size, in this case 10% of the compounds. One by one, each part was excluded from the training set and put in a corresponding test set. Each training set was used to construct a classification model, here a decision list, which was then used to make predictions for the corresponding test set. This way, the mutagenicity of each compound of the complete data set was predicted once. In the current study, 10-fold cross-validation produced training sets that contained 90% of the complete data set. Accordingly, thresholds of 63 (90% of 70[42]) or more mutagens were handled for training sets during cross-validation. For each of the four scenarios, cross-validation was performed with identical training and test sets.

RESULTS

In total, Gaston learned substructures that occurred in 70 or more mutagens for four different scenarios: standard or elaborate chemical representation and with or without also considering nonlinear substructures. The scenario of elaborate chemical representation that also included nonlinear fragments was considered in more detail. Data mining of the corresponding substructure set resulted in a set of six substructures. This set was visualized as a decision list in Figure 3. The substructure that was extracted to split the complete data set describes a planar group (**1**) that is shown at the start of the decision list of Figure 3 (top left corner). The second substructure, a nitroso/azo-type group (**2**), was extracted from the subset of compounds that lack substructure **1**, etc.

Table 1 describes the statistics of the individual substructures of Figure 3 with respect to the entire data set (right section) and the subset that was analyzed to extract each substructure (left section). As an example, the substructure of a three-membered heterocycle (**3** in Figure 3 and Table 1) detected 191 mutagens and 35 nonmutagens in the entire data set. Note that the subset that was analyzed did not include compounds that also contained a polycyclic planar system or a nitroso/azo-type group. The analyzed subset therefore contained 153 mutagens and 32 nonmutagens. In other words, substructure **3** detected 185 compounds from the analyzed subset, of which 83% were mutagenic.

During classification, the decision list in Figure 3 first tested whether a compound contained substructure **1.** If so, the arrow indicated by **Y** was followed, and the compound was classified as a mutagen. If not, the arrow indicated by **N** was followed, and the compound was tested for the

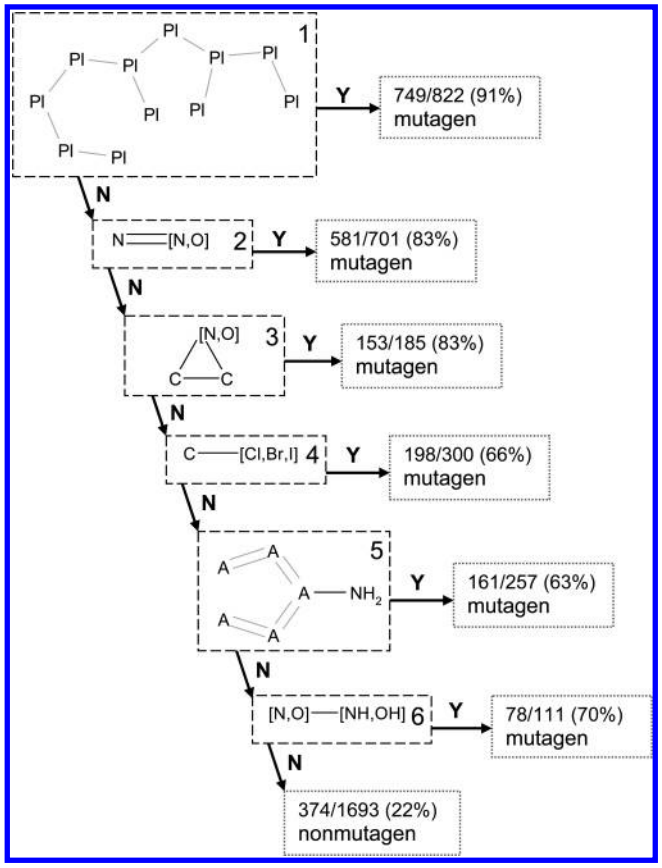

Figure 3. Decision list extracted from the complete data set. Substructures are numbered and shown in black dashed boxes. Grey dotted boxes describe statistics of individual subsets of the data set: the number of mutagens, the total number of compounds, the percentage of mutagens, and the corresponding classification. Arrows indicated by **Y** and **N**, respectively, show the direction to follow if a substructure is (**Y**) or is not (**N**) present in a compound. Where possible, atoms were depicted in the standard chemical representation.

presence of substructure **2**. Each compound was tested for substructures until it was put in a subset, where it was classified. Figure 3 shows that compounds which contained any of the six substructures were classified as mutagens. For instance, because 91% of the compounds that contained substructure **1** were mutagenic, all were classified as mutagens. Compounds that did not contain any of these substructures were classified as nonmutagens because only 22% of these were mutagenic. This classification method resulted in an overall classification error of 20%. The last row of Table 2 corresponds to the scenario of elaborate chemical representation that included nonlinear substructures. This row contains a more detailed statistical overview of the classification performance made by the decision list from Figure 3. These results demonstrate that 1920 mutagens (47% of the entire data set) were classified correctly, while 374 mutagens (9%) were classified incorrectly. In other words, mutagens were detected with a sensitivity of 84%. The observed specificity for detecting nonmutagens was 74%.

For each substructure learning scenario, 10 decision lists were constructed during cross-validation. Each list was used to predict the mutagenicity of compounds that were not in the training set. Table 3 shows the overall statistics of mutagenicity prediction of the complete data set as based on 10-fold cross-validation. The last row of Table 3 shows that 1904 mutagens (47% of the entire data set) of all 2294 mutagens were predicted correctly. The sensitivity and specificity that were observed in cross-validation, 83% and 74% respectively, were very close to their corresponding values in the entire analysis, 84% and 74%, of Table 2.

DISCUSSION

Extracted Substructures. In this section, we examine the substructures that are shown in the decision list of Figure 3. Only chemical differences between the current substructures and substructures found earlier are discussed below, because

Table 1. Statistics of Individual Substructures[a]

	analyzed subset				entire data set			
substructure	mutagens	nonmutagens	fraction of mutagens	p-value	mutagens	nonmutagens	fraction of mutagens	p-value
polycyclic planar system (**1**)	749	73	91%	$\ll 10^{-20}$	749	73	91%	$\ll 10^{-20}$
nitroso/azo-type group (**2**)	581	120	83%	$\ll 10^{-20}$	763	122	86%	$\ll 10^{-20}$
three-membered heterocyle (**3**)	153	32	83%	$\ll 10^{-20}$	191	35	84%	$\ll 10^{-20}$
aliphatic halide (**4**)	198	102	66%	$\ll 10^{-20}$	277	104	73%	$\ll 10^{-20}$
aromatic primary amine (**5**)	161	96	63%	$\ll 10^{-20}$	357	109	77%	$\ll 10^{-20}$
heteroatom-bound heteroatom (**6**)	78	33	69%	$\ll 10^{-20}$	117	38	85%	$\ll 10^{-20}$

[a] Statistics are shown with respect to both the subset that was analyzed to derive each substructure (left section) and the entire data set (right section). The number of mutagens, nonmutagens, the fraction of mutagens, and the p-value are shown for each data set. The corresponding substructures are shown in Figure 3. The order of the substructures in this table reflects the order in which they were extracted, that is, the order in the decision list.

Table 2. Overall Statistics of Mutagenicity Classification of the Entire Data Set According to Each Substructure Learning Scenario[a]

substructure learning scenario	error	sensitivity	true positives	false negatives	specificity	true negatives	false positives
standard and linear only	**29%**	**59%**	33% (1359)	23% (935)	**87%**	38% (1549)	6% (226)
standard and nonlinear also	**26%**	**65%**	37% (1492)	20% (802)	**85%**	37% (1505)	7% (270)
elaborate and linear only	**24%**	**74%**	42% (1696)	15% (598)	**78%**	34% (1382)	10% (393)
elaborate and nonlinear also	**20%**	**84%**	47% (1920)	9% (374)	**74%**	32% (1319)	11% (456)

[a] Each row describes the statistics for one scenario. Columns describe the percentages of the overall error, sensitivity, and specificity in bold or the fractions of the data set with the corresponding number of compounds in brackets.

Table 3. Overall Statistics of Mutagenicity Prediction of the Entire Data Set According to Each Substructure Learning Scenario as Based on 10-Fold Cross-Validation[a]

substructure learning scenario	error	sensitivity	true positives	false negatives	specificity	true negatives	false positives
standard and linear only	**30%**	**58%**	32% (1322)	24% (972)	**87%**	38% (1544)	6% (231)
standard and nonlinear also	**29%**	**61%**	34% (1393)	22% (901)	**85%**	37% (1516)	6% (259)
elaborate and linear only	**25%**	**73%**	41% (1665)	15% (629)	**78%**	34% (1390)	9% (385)
elaborate and nonlinear also	**21%**	**83%**	47% (1904)	10% (390)	**74%**	32% (1322)	11% (453)

[a] Each row describes the statistics for one scenario. Columns describe the percentages of the overall error, sensitivity and specificity in bold or the fractions of the data set with the corresponding number of compounds in brackets.

their mutagenic modes of action were discussed before[20] and because minor statistical differences were due to the stricter elimination of entries from the original data set.

The first, highly branched substructure (**1**) in the decision list of Figure 3 contains 11 planar atoms that are connected with planar bonds. Of all possible substructures, it is selected first, which means that this substructure is most discriminative for mutagenicity. Due to its branched shape and the amount of atoms, it describes a polycyclic planar system consisting of at least three rings. This substructure is a typical example of a substructure that could not have been determined with other substructure mining methods[4−8,30−33] since it is a large, nonlinear fragment and since it uses the planarity definition. It is related to the general toxicophores for a polycyclic aromatic system and a polycyclic planar system, which were derived in the previous study.[20] In comparison, the description of planarity that was handled here (see Supporting Information for the exact definition) is more exact than the one used in the previous analysis as it excludes more nonplanar rings.

The second substructure (**2**) in the decision list contains a nitrogen atom that is connected through a double bond to a nitrogen or oxygen atom. In a biochemical sense, this nitroso/azo-type substructure is very general as it encompasses three previously characterized general toxicophores: the aromatic nitro, the nitroso, and the azo-type group.[1,2,20] Since the data mining method for selecting substructures is statistics-based, it does not necessarily distinguish between chemically distinct groups. As a result, this nitroso/azo-type substructure also detects compounds that were previously not classified or predicted as mutagens.

The three-membered heterocycle (**3**) detects only aliphatic epoxides and aziridines and so differs slightly from the matching general toxicophore detected earlier.[1,2,20]

Aliphatic halogen (which detects chlorine, bromine, and iodine) (**4**) and aromatic primary amine (**5**) are identical to the corresponding general toxicophores detected earlier.[1,2,20]

The heteroatom-bonded heteroatom substructure (**6**) is very similar to the previously detected general toxicophore for an unsubstituted, heteroatom-bonded heteroatom.[20] Substructure **6** differs only slightly from the general toxicophore as it detects heteroatom-bound secondary amines, while it does not detect all heteroatom-bound primary amines. Both versions detect hydroxylamines, primary peroxides, and primary amines that are connected to secondary amines.[20]

In all, the current analysis provides further statistical support to the general toxicophores derived in our previous study.[20] For predictive purposes, we favor the usage of more specific toxicophores for which individual mechanistic hypotheses and confidence estimates exist.

Chemical Representation. Substructure mining methods, such as CASE,[6] MultiCASE/MCASE,[7] MOLFEA,[5] and others,[8,15] consider only combinations of default atom (C, N, O, etc.) and bond types (single, double, triple, and aromatic). King et al.[14] made an extension to this by describing atoms with extra data. Helma et al.[4] described atoms with single, nonwildcard labels that are standard in the SMARTS language (C for aliphatic carbon, c for aromatic carbon, etc.).[45,46] The usage of wildcards in substructure mining was introduced by Hofer et al.,[44] who used a very general wildcard (any atom). The concept of describing chemical information in more subtle levels of detail is not new, as exemplified by the existence of the SMARTS language.[45,46] However, one contribution of the current study lies in exploring different levels of chemical detail during substructure mining. This was made possible by a graph-based representation method that uses atomic hierarchies to describe single atoms with general and specific labels.

Even from the small set of six detected substructures from Figure 3, the value of such an elaborate chemical representation is evident: all six detected substructures contain general wildcards, and two substructures, the aromatic primary amine and the heteroatom-bound heteroatom, contain additional specifiers. The only limitation of the use of wildcards can be the statistical extraction of oversimplified fragments, as was the case for the nitroso/azo-type substructure. However, this can be overcome by a subsequent analysis of the subset of compounds that contain the general substructure. Namely, a set of more specific substructures will result when wildcards are excluded in the chemical representation step of this analysis. The final section of the discussion more specifically describes the effect of using elaborate chemical representation rather than standard chemical notation on the overall predictive performance.

Substructure Mining. Substructure mining methods have been very useful for computational SAR extraction from chemically diverse data sets, including mutagenicity data sets.[4,22−29] In principle, CASE,[6] MultiCASE,[7] and MOLFEA[5] consider only linear fragments. To partially overcome this limitation, CASE and MultiCASE were extended to detect several predefined branches and three- or four-membered rings.[6,7] Another limitation lies in their ability to detect fragments with up to 10 atoms. In comparison, an independently developed CASE-analog[8] considers fragments with very short branches, that is, directly-neighboring bonds and atoms, but their approach only detects fragments with up to eight atoms. MOLFEA,[5] on the other hand, considers only linear fragments, with the advantage of analyzing fragments of any size. None of these methods consider the majority of existing graphs, that is, branched and cyclic graphs. The scenario of elaborate chemical representation that also

considers nonlinear substructures resulted in six substructures, of which two are branched and one is cyclic. This already suggests that searching for all graphs provides added value over searching for paths only. The effect of also considering nonlinear substructures on the overall predictive performance is explained in the final section of the discussion.

In addition to the above substructure mining approaches, algorithms exist that mine for fragments of every two-dimensional shape,[30−34,47] but so far they have received relatively little attention from the chemoinformatics community. For mutagenicity in particular, one search for graphs has been performed with a frequency constraint of 20% in a data set of 230 aromatic nitro compounds.[53] Such a high constraint (20% compared to <2% in the present analysis[42]) was handled to this small data set of limited chemical diversity due to efficiency limitations.[53]

As discussed in the methods section, Gaston splits up the substructure mining process into several phases by mining for all paths, trees, and graphs, respectively. Because Gaston also prevents duplicate substructure searches, it is currently the most efficient algorithm for learning all (linear and nonlinear) substructures from a set of chemicals.[34,49] As a consequence, Gaston can process chemically diverse databases of thousands of compounds with low-frequency constraints. When elaborate chemical representation is used prior to analysis, substructure mining can explore additional, information-rich substructures.

Classification and Prediction of Mutagenicity. The separation of the data mining step from the steps of chemical representation and substructure mining enables the independent use of various machine learning algorithms and validation methods. This was skillfully exemplified by Helma et al.,[4] who tested different data mining algorithms for their ability to derive SARs.

The last row of Table 2 shows that the small decision list of Figure 3 produces an overall error of 20% in classifying the mutagenicity of this data set. Importantly, Table 3 shows that this classification error is comparable to the average prediction error of 21% that resulted from the corresponding 10-fold cross-validation. In comparison, our earlier study[20] obtained errors of 18% and 15% in classification and prediction, which lie closer to 15%, the average error of interlaboratory reproducibility of Ames tests.[54,55] As expected, the current results did not outperform the previous, more comprehensive collection of toxicophores for mutagenicity prediction.[20] The suggested, minor loss in performance may be attributed to the strictness of the settings that were handled in the fully automated steps of substructure mining and decision list creation. As a result, the scenario with elaborate chemical representation that included nonlinear substructures yielded decision lists of only six substructures from both the complete analysis and cross-validation. However, this strictness may have also been responsible for the negligible difference between the errors of classification and prediction that are reported in Tables 2 and 3. This small difference indicates that the current study provided a robust method for extracting substructures that were most discriminative for mutagenicity.

Table 4 shows the error percentages of mutagenicity prediction that were reported for a variety of substructure mining methods. For one, Table 4 shows that we have analyzed the largest mutagenicity data set currently available. Because studies by other authors were based on different data sets, we warn that the extent to which predictive performance can be compared is limited. For comparison, we note that one study[28] reported error percentages that were computed differently. These percentages were based on only those parts of the analyzed data sets for which 'conclusive predictions'[29] could be made, thereby omitting 19% of the compounds and their corresponding predictions. Considering this, the results from Table 4 suggest that our overall error in prediction was comparable and often even favorable with respect to error percentages achieved in earlier studies. Surprisingly, other substructure mining methods need significantly more substructures to yield acceptable error percentages, whereas our results are based on only six substructures. The combination of elaborate chemical representation and substructure mining without shape restrictions enables a single substructure to cover different moieties that share general and/or specific chemical features. We therefore hypothesize that fewer substructures are needed to describe the chemical information that is most discriminative for mutagenicity.

Table 4. Overall Statistics of Mutagenicity Prediction Errors Reported for Various Substructure Mining Methods and Data Sets[a]

method (name or acronym)	no. of compds in data set	fraction of mutagens in data set	overall error in prediction
CASE[23]	93	45%	24%
CASE[23]	89	40%	29%
CASE[21]	114	50%	28%
MultiCASE[21]	114	51%	20%
MultiCASE[24 b]	123	39%	28%[b]
MultiCASE[24 b]	516	55%	19%[b]
MultiCASE[26]	70	40%	24%
MultiCASE[28]	2513	43%	12%
MultiCASE[28]	52	40%	15%
CASE analog[22]	551	52%	26%
MOLFEA[4]	684	50%	22%
current study (learning scenario: elaborate and nonlinear also)	4069	56%	21%

[a] Data sets are only listed if they contain roughly as many mutagens as nonmutagens. The first column shows the name or acronym of the method. The second and third columns show the number of compounds and the fraction of mutagens in the analyzed data set. The fourth column shows the corresponding error percentage of mutagenicity prediction of this data set with this method. [b] Compounds used for training were not excluded prior to prediction, which may have resulted in an overlap between the training and test set.

When the overall classification performance is compared for different substructure learning scenarios, Tables 2 and 3 show that any gain in sensitivity comes at a cost of specificity. This corresponds with the notion that when more mutagens are recognized some nonmutagens are also falsely considered as mutagens. Tables 2 and 3 also show that the overall classification and prediction error of a substructure learning scenario decreases considerably in two cases: (1) when nonlinear substructures are also considered during mining and (2) when elaborate representation is used instead of standard chemical notation. The individual positive effect of handling elaborate chemical representation seems fairly larger than the effect of also considering nonlinear substructures. Furthermore, the observed decreases of overall error percentages in classification and prediction suggest that both beneficial effects are additive.

CONCLUSION

We developed an elaborate, graph-based method of chemical representation that fully exploits the chemical information present in a large mutagenicity data set. This method uses atomic hierarchies to describe single atoms with general and specific labels. Subsequent substructure mining with Gaston enabled the efficient detection of substructures of any size, shape, and level of chemical detail. Decision lists containing six discriminative, nonredundant substructures were extracted that classified and predicted the mutagenicity of this data set with overall errors of 20% and 21%, respectively. Data mining of all these substructures yielded structure−activity relationships that are robust, information-rich, and directly applicable to chemical design. The extracted substructures represent relevant biochemical knowledge because they are highly similar to toxicophores recognized in previous studies. Our results demonstrate the individual and synergistic importance of elaborate chemical representation and mining for nonlinear substructures. We conclude that the combination of elaborate chemical representation and Gaston provides an excellent method for 2D substructure mining as this recipe systematically explores all substructures in different levels of chemical detail.

Supporting Information Available: Individual SMILES data sets of the mutagens and nomutagens and SMARTS strings of aromatic and planar settings of chemical representation. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Ashby, J. Fundamental Structural Alerts to Potential Carcinogenicity or Noncarcinogenicity. *Environ. Mutagen.* **1985**, *7*, 919−921.

(2) 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.

(3) 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.

(4) Helma, C.; Cramer, T.; Kramer, S.; De Raedt, L. Data Mining and Machine Learning Techniques for the Identification of Mutagenicity Inducing Substructures and Structure Activity Relationships of Noncongeneric Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1402−1411.

(5) Helma, C.; Kramer, S.; DeRaedt, L. The Molecular Feature Miner MOLFEA. In *Proceedings of the Beilstein Workshop 2002: Molecular Informatics: Confronting Complexity*; Beilstein Institut: Frankfurt, 2002.

(6) Klopman, G. Artificial Intelligence Approach to Structure−Activity Studies: Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315−7321.

(7) Klopman, G. Multi-CASE: a Hierarchical Computer Automated Structure Evaluation Program. *QSAR* **1992**, *11*, 172−184.

(8) Malacarne, D.; Pesenti, R.; Paolucci, M.; Parodi, S. Relationship Between Molecular Connectivity and Carcinogenic Activity: A Confirmation with a New Software Program Based on Graph Theory. *Environ. Health Perspect.* **1993**, *101*, 332−42.

(9) Sanderson, D. M.; Earnshaw, C. G. Computer Prediction of Possible Toxic Action from Chemical Structure; the DEREK System. *Hum. Exp. Toxicol.* **1991**, *10*, 261−273.

(10) Smithing, M. P.; Darvas, F. Hazardexpert: an Expert System for Predicting Chemical Toxicity. In *Food Safety Assessment*; American Chemical Society: Washington, DC, 1992; pp 192−200.

(11) Enslein, K.; Gombar, V. K.; Blake, B. W. International Commission for Protection Against Environmental Mutagens and Carcinogens. Use of SAR in Computer-Assisted Prediction of Carcinogenicity and Mutagenicity of Chemicals by the TOPKAT Program. *Mutat. Res.* **1994**, *305*, 47−61.

(12) Woo, Y. T.; Lai, D. Y.; Argus, M. F.; Arcos, J. C. Development of Structure−Activity Relationship Rules for Predicting Carcinogenic Potential of Chemicals. *Toxicol. Lett.* **1995**, *79*, 219−228.

(13) Ridings, J. E.; Barratt, M. D.; Cary, R.; Earnshaw, C. G.; Eggington, C. E. et al. Computer Prediction of Possible Toxic Action from Chemical Structure: an Update on the DEREK System. *Toxicology* **1996**, *106*, 267−279.

(14) King, R. D.; Muggleton, S. H.; Srinivasan, A.; Sternberg, M. J. Structure−Activity Relationships Derived by Machine Learning: the Use of Atoms and Their Bond Connectivities to Predict Mutagenicity by Inductive Logic Programming. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 438−442.

(15) Young, S. S.; Gombar, V. K.; Emptage, M. R.; Cariello, N. F.; Lambert, C. Mixture-Deconvolution and Analysis of Ames Mutagenicity Data. *Chemom. Intell. Lab. Syst.* **2002**, *60*, 5−11.

(16) Bacha, P. A.; Gruver, H. S.; Den Hartog, B. K.; Tamura, S. Y.; Nutt, R. F. Rule Extraction from a Mutagenicity Data Set Using Adaptively Grown Phylogenetic-Like Trees. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1104−1111.

(17) Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Oloff, S. et al. Three New Consensus QSAR Models for the Prediction of Ames Genotoxicity. *Mutagenesis* **2004**, *19*, 365−377.

(18) Li, H.; Ung, C. Y.; Yap, C. W.; Xue, Y.; Li, Z. et al. Prediction of Genotoxicity of Chemical Compounds by Statistical Learning Methods. *Chem. Res. Toxicol.* **2005**, *18*, 1071−1080.

(19) Mahe, P.; Ueda, N.; Akutsu, T.; Perret, J. L.; Vert, J. P. Graph Kernels for Molecular Structure−Activity Relationship Analysis with Support Vector Machines. *J. Chem. Inf. Model.* **2005**, *45*, 939−951.

(20) Kazius, J.; McGuire, R.; Bursi, R. Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J. Med. Chem.* **2005**, *48*, 312−320.

(21) Contrera, J. F.; Matthews, E. J.; Kruhlak, N. L.; Benz, R. D. In Silico Screening of Chemicals for Bacterial Mutagenicity Using Electrotopological E-State Indices and MDL QSAR Software. *Regul. Toxicol. Pharmacol.* **2005**, *43*, 313−323.

(22) Klopman, G.; Rosenkranz, H. S. Testing by Artificial Intelligence: Computational Alternatives to the Determination of Mutagenicity. *Mutat. Res.* **1992**, *272*, 59−71.

(23) Perrotta, A.; Malacarne, D.; Taningher, M.; Pesenti, R.; Paolucci, M.; Parodi, S. A Computerized Connectivity Approach for Analyzing the Structural Basis of Mutagenicity in Salmonella and its Relationship with Rodent Carcinogenicity. *Environ. Mol. Mutagen.* **1996**, *28*, 31−50.

(24) Zeiger, E.; Ashby, J.; Bakale, G.; Enslein, K.; Klopman, G. et al. Prediction of Salmonella Mutagenicity. *Mutagenesis* **1996**, *11*, 471−484.

(25) Pearl, G. M.; Livingston-Carr, S.; Durham, S. K. Integration of Computational Analysis as a Sentinel Tool in Toxicological Assessments. *Curr. Top. Med. Chem.* **2001**, *1*, 247−255.

(26) Dearden, J. C.; Barratt, M. D.; Benigni, R.; Bristol, D. W.; Combes, R. D.; et al. The Development and Validation of Expert Systems for Predicting Toxicity: Report and Recommendations of an ECVAM/ECB Workshop (ECVAM Workshop 24). *ATLA* **1997**, *25*, 223−252.

(27) White, A. C.; Mueller, R. A.; Gallavan, R. H.; Aaron, S.; Wilson, A. G. A Multiple In Silico Program Approach for the Prediction of Mutagenicity from Chemical Structure. *Mutat. Res.* **2003**, *539*, 77−89.

(28) Snyder, R. D.; Pearl, G. S.; Mandakas, G.; Choy, W. N.; Goodsaid, F.; et al. 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.

(29) 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.

(30) Inokuchi, A.; Washio, T.; Motoda, H. An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data. In *Proceedings of the 4th European Conference on Principles of Knowledge Discovery and Data Mining (PKDD)*; 2000; pp 13−23.

(31) Kuramochi, M.; Karypis G. Frequent Subgraph Discovery. In *Proceedings of the International Conference on Data Mining (ICDM)*; 2001; pp 313−320.

(32) Borgelt, C.; Berthold, M. R. Mining Molecular Fragments: Finding Relevant Substructures of Molecules. In *Proceedings of the International Conference on Data Mining (ICDM)*; 2002; pp 51−58.

(33) Yan, X.; Han, J. gSpan: Graph-Based Substructure Pattern Mining. In *Proceedings of the International Conference on Data Mining (ICDM)*; 2002; pp 721−724.

(34) Nijssen, S.; Kok, J. N. A Quickstart in Frequent Structure Mining can make a Difference. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*; 2004; pp 647−652.

(35) Nijssen, S. Gaston at http://www.liacs.nl/∼snijssen/gaston/.

(36) Chemical Carcinogenesis Research Information System is available through TOXNET at http://toxnet.nlm.nih.gov.

(37) Weininger, D. SMILES, a Chemical Language and Information System 1. Introduction and Encoding Rules. *J. Chem. Inf. Comput. Sci*. **1988**, *28*, 31−36.
(38) CACTVS, at http://www.xemistry.com/.
(39) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, S.; Sasaki, S. Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach Toward Modularity and Flexibility. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 109−116.
(40) The used CACTVS-Tcl script was kindly provided by Dr. Marc Nicklaus.
(41) Frowns, at http://frowns.sourceforge.net/.
(42) For the purpose of comparison, the settings handled in the present study (settings of chemical representation (choice of atom and bond groupings), frequency constraints (≥70 mutagens), and data mining (≥60% mutagenic)) were based on the corresponding settings used to extract general toxicophores in the previous analysis (Kazius et al., 2005).
(43) Judson, P. N. Rule Induction for Systems Predicting Biological Activity. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 148−153.
(44) Hofer, H.; Borgelt, C.; Berthold, M. R. Large Scale Mining of Molecular Fragments with Wildcards. In *Advances in Intelligent Data Analysis V (IDA)*; 2003; pp 380−389.
(45) OpenBabel, at http://openbabel.sourceforge.net/.
(46) Daylight Chemical Information, Inc., Santa Fe, at www.daylight.com/dayhhtml/doc/theory/theory.smarts.html.
(47) An overview of subgraph mining methods is available at http://hms.liacs.nl/.
(48) Agrawal, R.; Srikant, R. Fast Algorithms for Mining Association Rules. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)*; 1994; pp 487−499.
(49) Wörlein, M.; Meinl, T.; Fischer, I.; Philippsen, M. A Quantitative Comparison of the Subgraph Miners MoFa, gSpan, FFSM, and Gaston. In *Proceedings of the 4th European Conference on Principles of Knowledge Discovery and Data Mining (PKDD)*; 2005.
(50) Clark, P.; Niblett, T. The CN2 Induction Algorithm. *Machine Learn.* **1989**, *3*, 261−284.
(51) C4.5, release 8, at http://rulequest.com/.
(52) Quinlan, J. R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann: San Mateo, 1993.
(53) Inokuchi, A.; Washio, T.; Okada, T.; Motoda, H. Applying the Apriori-Based Graph Mining Method to Mutagenesis Data Analysis. *J. Comput.-Aided Chem.* **2001**, *2*, 87−92.
(54) Benigni, R.; Giuliani, A. Computer Assisted Analysis of Interlaboratory Ames Test Variability. *J. Toxicol. Environ. Health* **1988**, *25*, 135−146.
(55) Piegorsch, W. W.; Zeiger, E. Measuring Intra-Assay Agreement for the Ames Salmonella Assay. In *Lecture notes in Medical Informatics*; Springer-Verlag: Heidelberg, 1991; pp 35−41.

CI0503715