Data Mining and Machine Learning Techniques for the Identification of Mutagenicity Inducing Substructures and Structure Activity Relationships of Noncongeneric Compounds

Christoph Helma,*,† Tobias Cramer,† Stefan Kramer,†,‡ and Luc De Raedt†

Institute for Computer Science, Machine Learning Lab, University Freiburg, Georges Köhler Allee 79, D-79110 Freiburg/Br., Germany, and Institute for Computer Science, Technical University Munich, Boltzmannstrasse 3, D-85748 Garching, Germany

Received November 7, 2003

This paper explores the utility of data mining and machine learning algorithms for the induction of mutagenicity structure−activity relationships (SARs) from noncongeneric data sets. We compare (i) a newly developed algorithm (MOLFEA) for the generation of descriptors (molecular fragments) for noncongeneric compounds with traditional SAR approaches (molecular properties) and (ii) different machine learning algorithms for the induction of SARs from these descriptors. In addition we investigate the optimal parameter settings for these programs and give an exemplary interpretation of the derived models. The predictive accuracies of models using MOLFEA derived descriptors is ∼10−15%age points higher than those using molecular properties alone. Using both types of descriptors together does not improve the derived models. From the applied machine learning techniques the rule learner PART and support vector machines gave the best results, although the differences between the learning algorithms are only marginal. We were able to achieve predictive accuracies up to 78% for 10-fold cross-validation. The resulting models are relatively easy to interpret and usable for predictive as well as for explanatory purposes.

1. INTRODUCTION

The development of models for *structure−activity relationships (SARs)* has a long and successful history in medicinal chemistry, but applications of SAR techniques for toxicological effects are still rather sparse. Although the objectives, the prediction of biological activities from chemical structures, are closely related, there are fundamental differences between both application areas.

SAR studies in pharmaceutical research rely typically on a few (several tens to hundreds) compounds, containing a basic structure responsible for activity (i.e. they are congeneric). Variations in activity are caused by secondary features (e.g. presence, length or composition of certain side-chains). In many cases the cellular target (e.g. receptor, active site of an enzyme) is known, and some information about biological mechanisms is also available. Based on this knowledge it is possible to select a limited set of descriptors for the chemical structures, which might be relevant for the activities of the investigated compounds.

The situation is much more complicated for the majority of toxic effects: Many different molecular mechanisms and cellular targets may be involved in a single toxic effect. As a consequence, chemicals with very different structures (*noncongenerics*) may cause the same toxicological effect. With such a limited amount of information, the selection of appropriate descriptors for SAR studies is more or less a trial and error process.[1] In addition many toxicity assays are rather time-consuming and expensive, especially those which are considered to have the highest relevance for human health (e.g. in vivo experiments). Under these circumstances it is in most cases impossible to perform experiments especially for SAR studies. Therefore SAR models have to use existing databases, where the composition and distribution of structures and activities is far from optimal, because they have been selected according to different criteria (e.g. production volumes, results from short-term tests).

Under these conditions the identification of a toxicological structure−activity relationship is essentially a data mining problem: The researcher has to identify regularities within the chemical structures and their toxic activities that allow the construction of a model that predicts the activity of untested compounds. More specifically, the data mining task can be decomposed into two steps:

1. Generation of descriptors for the chemical structures
2. Induction of the SAR model

In the present paper, the use of various data mining techniques[2] for both steps is systematically explored on a database of mutagenic activities. The goal is to obtain insight into the utility of such data mining techniques for building SARs from toxicological databases. Among the data mining techniques selected, special attention was given to symbolic machine learning techniques, because they are able to derive understandable and interpretable models. It is therefore possible to use the induced models for predictive as well as for explanatory purposes (i.e. to identify structural features that may cause a certain toxic effect).

In particular we have investigated the following:

- the utility of a newly developed system for the generation of molecular substructures (MOLFEA[3,4]) as compared to molecular properties calculated by various computational chemistry programs,

* Corresponding author phone: ++49-761-203-8013; e-mail: helma@informatik.uni-freiburg.de.
† University Freiburg.
‡ Technical University Munich.

10.1021/ci034254q CCC: $27.50
Published on Web 06/10/2004

- the suitability of various machine learning programs for the generation of mutagenicity SARs from the descriptors mentioned before and
- the optimization of parameters for these programs, to identify the optimal conditions for the detection of mutagenicity SARs within data sets with noncongeneric compounds.

2. SYSTEM AND METHODS

2.1. Database. The mutagenicity data set was extracted from the carcinogenic potency database (CPDB) (http://potency.berkeley.edu/cpdb.html[5]). The CPDB provides mutagenicity classifications (mutagens and nonmutagens) as determined by the *Salmonella*/microsome assay (Ames test[6]). Additional information (mutagenic potencies, results in individual strains, necessity of metabolic activation) is not available.

The data set contains a very diverse set of chemicals of predominatly industrial and pharmaceutical origins. A visual representation of all structures and their mutagencity classifications can be downloaded from http://www.predictive-toxicology.org/data/cpdb_mutagens/mutagens.pdf. Chemical structures came from various sources on the Internet (see http://www.predictive-toxicology.org/db_links/). They were converted to SMILES strings,[7] checked for validity, and corrected according to the procedures described in a previous publication.[8] After the elimination of mixtures and undefined structures, 684 compounds remained for the experiments.

The whole data set with chemical structures, fragments for various thresholds (see below), molecular properties, and mutagenic activities is available from http://www.predictive-toxicology.org/data/cpdb_mutagens (also see Supporting Information).

2.2. Fragment Generation. To predict the mutagenic activity of the compounds in our database, we first describe the chemical structures in terms of substructural fragments. The fragments are generated automatically from the data set using the molecular feature miner MOLFEA. MOLFEA is an inductive database system[9] tailored toward discovering substructures within sets of small molecules. Inductive databases are databases that can be queried not only for data but also for patterns and regularities (in our case: substructures) that occur within the data and fulfill certain user defined criteria. An example query in our investigation would be to ask for all substructures that occur with a high frequency in mutagenic compounds and a low frequency in nonmutagenic compounds.

More precisely, the substructures considered in the current version of MOLFEA are linear molecular fragments, i.e., sequences of atoms and connecting bonds. An example fragment would be C−C-c:c:c−O, which stands for two nonaromatic carbons connected by a single bond, followed by three aromatic carbons, connected by a single bond to an oxygen. The language of fragments used in MOLFEA is a subset of the SMARTS language (http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html). Please note that the restriction to linear substructures is not an integral part of MOLFEA but only represents a starting point for our investigations.

A query in MOLFEA can be composed of several conditions, each of which has to be fulfilled in order to make it a solution fragment. These conditions are expressed in terms of primitive constraints considering e.g. the syntax of the fragments or their frequency in different data sets. A syntactic constraint could e.g. state that the fragments of interest should be a substructure (or a superstructure) of a given structure. A frequency constraint requires that the fragment occurs in at least (respectively at most) x% of the molecules belonging to a given data set. In the present investigation we have used only frequency constraints. Hence we will focus on them while explaining the algorithm of MOLFEA. More detailed information about MOLFEA can be found in previous publications.[3,4,10]

Most general fragment g**:** `c:c-c:c`

Most specific fragment s**:** `c:c:c:c:c-c:c:c:c:c:c`

All solutions: `c:c-c:c`, `c:c-c:c:c`, `c:c-c:c:c:c`, `c:c-c:c:c:c:c`, `c:c-c:c:c:c:c:c`, `c:c:c-c:c:c:c`, `c:c:c-c:c:c:c:c`, `c:c:c-c:c:c:c:c:c`, `c:c:c:c-c:c:c:c:c`, `c:c:c:c-c:c:c:c:c:c`, `c:c:c:c:c-c:c:c:c:c`, `c:c:c:c:c-c:c:c:c:c:c`

Figure 1. Example for the version space representation of a MOLFEA query result.

If $freq(f, D)$ denotes the frequency of a fragment f on a set of molecules D, the frequency of a fragment f in a database D is defined as the fraction of molecules in D in which f is occurring. In MOLFEA we may pose queries concerning the minimum and the maximum frequency on (possibly different) data sets. We may write such queries as $(freq(f, A) < t_A) \wedge freq(f, B) > t_B)$ where t_A and t_B are relative frequencies and A and B are different sets of molecules. This constraint denotes that the frequency of the fragment f in the data set A (e.g. nonmutagens) should be less than t_A and the frequency in B (e.g. mutagens) should be greater than t_B.

To efficiently compute the fragments that satisfy a given query the generality relation on fragments is exploited. This is important because it will allow us to prune the space of possible solutions.

More formally, for linear fragments of molecules, we say that a fragment g is more general than a fragment s, if and only if g is a subsequence of s or g is a subsequence of the reversal of s. The fragment c−C (aromatic carbon connected to an aliphatic carbon) for example is more general than fragment C−c:c (aliphatic carbon connected to two aromatic carbons). Whenever a specific fragment s occurs in a molecule it must be the case that all subfragments of s (i.e. all fragments that are more general than s, the so-called generalization) will also occur in the same molecule.

This knowledge is important in the light of the frequency constraints. Indeed, whenever a fragment s satisfies a constraint of the form $freq(s, D) > t$, then all generalizations g of s will also satisfy the constraint.

This property actually imposes a lower border (called the S-set) on the space of possible solutions (cf. Figures 1 and 2). S contains all maximally specific fragments that satisfy the constraint. It is called a border because all fragments more general than an element of S will also satisfy the constraint and all fragments that are not more general than at least one fragment in S will not satisfy the constraint. Dually, whenever a fragment g satisfies a constraint of the form $freq(g, D) < t$, then all of its specializations (superfragments) will satisfy the constraint as well. This property−

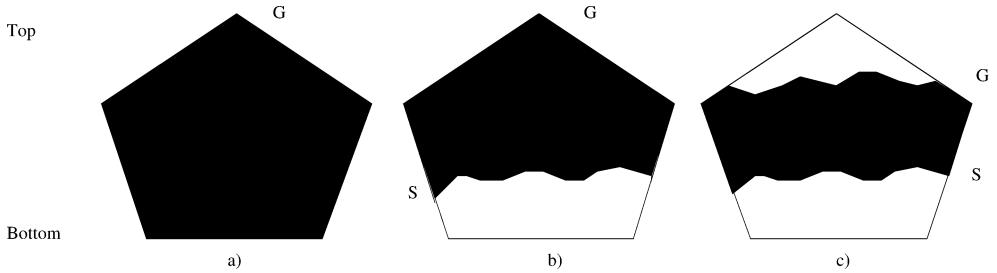


Figure 2. Evolution of the version space for the conjunction of a minimum frequency query and a maximum frequency query. The solutions are indicated by the black area. (a) depicts the initial situation, (b) is the solution for the minimum frequency query, and (c) is the final result for both constraints. *G* indicates the set of the most general solutions, *S* is the set of the most specific solutions, the top of the version space (T) is the "empty" fragment, and the bottom of the version space (⊥) is a hypothetical fragment, that is more specific than all other fragments.

in turn—imposes an upper border (the *G*-set) on the space of possible solutions.

Now, because of these properties the set of solutions to a query of the form ($freq(f, A) > t_A \wedge freq(f, I) < t_I$, *A* is the set of active compounds and *I* is the set of inactive compounds) can be completely characterized by its two borders *S* and *G*. Indeed, the first constraint induces the lower border *S*, and the second one induces the upper border *G* (Figure 2). Therefore, all fragments being more general than a fragment in *S* and more specific than an element in *G* will be solutions to the query. In machine learning terminology, the solution space is called a version space. Figure 1 illustrates the concept of a version space, where *G* = ={c: c−c:c} and *S* = {c:c:c:c:c:c−c:c:c:c:c:c}.

Because the space of all possible solutions to a conjunctive query of the form $c_1 \wedge ... c_n$ (c_i are individual constraints) is a version space, and version spaces are completely characterized by their border sets, it suffices to compute the border sets with respect to (wrt) such queries in order to have a complete characterization of the solution set. This is also the underlying idea in the MOLFEA system. To compute the borders wrt a given query $c_1 \wedge c_2 \wedge ... c_n$, MOLFEA will incrementally process the constraints from left to right and update the border sets for each c_i correspondingly.

Figure 2 depicts the evolution of the version space for the conjunction of a minimum frequency and a maximum frequency constraint ($freq(f, A) > t_A \wedge freq(f, I) < t_I$). The initial version space (before processing the first primitive constraint c_1, see Figure 2a) contains all possible fragments. We write that $G_0 = \{T\}$ and $S_0 = \{\perp\}$, where T is the most general linear fragment (that is, the "empty" fragment), and ⊥, by definition, is the most specific linear fragment. [Please note that the ⊥ in the case of linear molecular fragment does not really exist. It is merely a theoretical construct denoting a fragment that is more specific than any other fragment.] Then we process the constraints c_i sequentially by looking for fragments that fulfill the given criteria (the exact procedure is described below). After each step c_i, we obtain an updated *G* and *S* set, either the *G* set becomes more specific, or the *S* set more general and the set of all solutions becomes smaller. On the example query ($freq(f, A) > t_A \wedge freq(f, I) < t_I$) in Figure 2, the minimum frequency constraint updates *S* (Figure 2b), and the maximum frequency constraint updates *G* (Figure 2c).

Let us now explain the details of the algorithm. To update the borders a minimum frequency threshold of the form $freq(f, A) > t_A$, MOLFEA applies algorithm 1 (Figure 4). It essentially combines the famous levelwise algorithm from

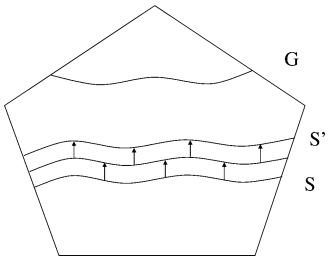


Figure 3. Level-wise search upward for a minimum frequency constraint to update *S* to *S*′.

input:
the current borders S and G
a constraint of the form $freq(f, D) > t$
output:
the updated borders S and G w.r.t. the constraint

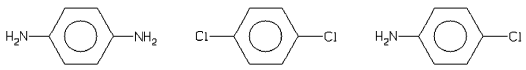
$i := 0$; $C_i := \{\top\}$; $F_i := \{\top\}$; $G := \{g \in G \mid freq(g, D) > t\}$
while $C_i \neq \emptyset$ **do**
$i := i + 1$;
$C_i := \{f \odot a \mid f \in F_{i-1}, a$ is an atom, and $\odot$ a bond $\}$
$C_i := \{a \odot f \mid (a \odot f) \in C_i$ and $f \in F_i\}$
$C_i := C_i - \{c \in C_i \mid c_i$ is not more general than any element of $S\}$
$F_i := \{c \in C_i \mid freq(c, D) > t\}$
endwhile
$S := \{c \mid c$ is maximally specific in $\bigcup_i F_i$ and c is more specific than an element in $G\}$

Figure 4. Algorithm 1 for solving minimum frequency queries (see Figure 5 for an illustrative example).

data mining[2] with principles of version spaces. The effect of the algorithm is illustrated in Figure 3. Given a minimum frequency threshold, the *S* border will be generalized.

Algorithm 1 works as follows (An example run on a small data set is depicted in Figure 5): First, those fragments that do not satisfy the minimum constraint are deleted (as they cannot contribute to a solution). Second, the elements of *S* are updated using a levelwise search algorithm. This algorithm keeps track of a list of candidates C_i and a list L_i of solutions to the ($freq(f, D) > t$) constraint. Both lists are initialized with the maximally general element T and iteratively updated. During each iteration, the candidates $f \odot a$ at level *i* are computed by refining existing fragments *f* at level $i - 1$. Here, *a* is an atom and ⊙ a bondtype. Because fragments of the form $a \odot f$ will not satisfy $freq(f, A) > t$ when *f* does not satisfy this constraint, the algorithm prunes away the corresponding candidates. Finally, those candidates that are not more general than the already specified *S* border cannot be a solution to the query $c_1 \wedge ... c_i$ and are therefore pruned away as well. All remaining candidates are then evaluated on the data in order to obtain their frequencies. The frequencies of these candidates are evaluated on the given data set with the Daylight SMARTS libraries

A :



```
LEVEL 1
Candidates:     [He] [Li] [Be] B C c N n O o F [Ne] [Na] [Mg] [Al] [Si] P p
S s Cl [Ar] [K] [Ca] [Sc] [Ti] [V] [Cr] [Mn] [Fe] [Co] [Ni] [Cu] [Zn] [Ga]
[Ge] [As] [Se] Br [Kr] [Rb] [Sr] [Y] [Zr] [Nb] [Mo] [Tc] [Ru] [Rh] [Pd]
[Ag] [Cd] [In] [Sn] [Sb] [Te] I [Xe] [Cs] [Ba] [Hf] [Ta] [W] [Re] [Os] [Ir]
[Pt] [Au] [Hg] [Tl] [Pb] [Bi] [Po] [At] [Rn] [Fr] [Ra] [Pt] [Ac] [La] [U]
[Sm] [Ce] [Nd] [Eu] [Gd] [Dy] [Er] [Rh]
Frequent:       c N Cl

LEVEL 2
Candidates:     c-c c=c c#c c:c c-N c=N c#N c:N c-Cl c=Cl c#Cl c:Cl N-N N=N N#N
N:N N-Cl N=Cl N#Cl N:Cl Cl-Cl Cl=Cl Cl#Cl Cl:Cl
Frequent:       c:c c-N c-Cl

LEVEL 3
Candidates:     c:c:c c:c-N c:c-Cl c-N-c N-c-N N-c-Cl c-Cl-c Cl-c-Cl
Frequent:       c:c:c c:c-N c:c-Cl

LEVEL 4
Candidates:     c:c:c:c c:c:c-N c:c:c-Cl N-c:c-N N-c:c-Cl Cl-c:c-Cl
Frequent:       c:c:c:c c:c:c-N c:c:c-Cl

LEVEL 5
Candidates:     c:c:c:c:c c:c:c:c-N c:c:c:c-Cl N-c:c:c-N N-c:c:c-Cl Cl-c:c:c-Cl
Frequent:       c:c:c:c:c c:c:c:c-N c:c:c:c-Cl

LEVEL 6
Candidates:     c:c:c:c:c:c c:c:c:c:c-N c:c:c:c:c-Cl N-c:c:c:c-N N-c:c:c:c-Cl
Cl-c:c:c:c-Cl
Frequent:       c:c:c:c:c:c c:c:c:c:c-N c:c:c:c:c-Cl N-c:c:c:c-N N-c:c:c:c-Cl
Cl-c:c:c:c-Cl

LEVEL 7
Candidates:     c:c:c:c:c:c:c c:c:c:c:c:c-N c:c:c:c:c:c-Cl N-c:c:c:c:c-N
N-c:c:c:c:c-Cl Cl-c:c:c:c:c-Cl
Frequent:       c:c:c:c:c:c-N c:c:c:c:c:c-Cl

LEVEL 8
Candidates:     N-c:c:c:c:c:c:c-N N-c:c:c:c:c:c:c-Cl Cl-c:c:c:c:c:c:c-Cl
Frequent:
```

Figure 5. Version space and results for the MOLFEA query *freq*(*f*, *A*) ≥ 2.

input:
the current borders S and G
a constraint of the form $freq(f, D) \leq t$
output:
the updated borders S and G w.r.t. the constraint

$i := 0$; $C_i := \{\top\}$; $F_i := \{\top\}$; $S := \{s \in S \mid freq(s, D) \leq t\}$
while $C_i \neq \emptyset$ **do**
$i := i + 1$;
$C_i := \{f \odot a \mid f \in F_{i-1}, a$ is an atom, and $\odot$ a bond $\}$
$C_i := \{a \odot f \mid (a \odot f) \in C_i$ and $f \in F_i\}$
$C_i := C_i - \{c \in C_i \mid c_i$ is not more general than any element of $S\}$
$F_i := \{c \in C_i \mid freq(c, D) > t\}$
$I_i := C_i - F_i$
endwhile
$G := \{c \mid c$ is maximally general in $\bigcup_i I_i$ and c is more general than an element in $S\}$

Figure 6. Algorithm 2 for solving maximum frequency queries.

(http://www.daylight.com) [OpenBabel http://openbabel.sourceforge.net is an Open Source alternative.] Those candidates satisfying the frequency threshold are retained in F_i and used to generate the candidates in the next iteration. The process is continued until no more candidates can be generated. At this point, the *S* set is computed by taking the maximally specific elements among the solutions computed that are more specific than an element of *G*. The algorithm for addressing a maximum frequency threshold of the form (*freq*(*f*, *D*) < *t*) can be obtained in a dual manner, it is listed in algorithm 2 (Figure 6).

In summary, MOLFEA is an inductive database for finding substructures in data sets with molecular structures. The two most important features are (1) the possibility of querying for patterns (fragments) in sets of compounds and (2) the organization of solution fragments in version spaces, that is, by storing the most general and the most specific solution fragments only.

2.3. Molecular Properties. Initial three-dimensional structures were calculated by the rule-based program CORINA.[11] Semiempirical quantum mechanical optimizations with the MNDO-PM3 Hamiltonian were performed with MOPAC (http://www.ccl.net/cca/software/LINUX/mopac7/README.shtml) to refine structures and calculate energetic and electronic properties. Lipophilicity (logP) was calculated with KOWWIN,[12] hydrophilic and lipophilic surface areas by NACCESS http://wolf.bms.umist.ac.uk/naccess/. All format conversions were done by Babel http://openbabel.sourceforge.net/.

2.4. Machine Learning. We compared in our experiments three machine learning algorithms in their implementation in the Weka workbench:[2] The decision tree learner C4.5[13] in the Weka implementation (J48), the rule learner PART,[14] and support vector machines (SVM).[15]

C4.5[13] is a classical decision tree algorithm. The algorithm is designed to construct small trees with the most relevant attributes near the root. C4.5 is known to perform well over a wide range of learning tasks and data sets. The advantages (fast performance) and disadvantages (fragmentation of cases into various subtrees and replication of tests in subtrees) both stem from the tree representation. Regarding comprehensibility, decision trees are in principle interpretable, but experts often tend to prefer rule sets over decision trees.

PART[14] is one of the best performing rule learning algorithms available. It differs from conventional rule learning schemes in that it does not require a separate, complex optimization stage, where the rule set is tuned after the induction of individual rules. Rules are perhaps the most popular representation of models in machine learning, since they can readily be interpreted by domain experts. On the negative side, comprehensibility comes often at the expense of predictivity, and rule learners are usually computationally more expensive than decision tree learners.

Support vector machines[15] are linear classifiers that map the input space of the learning examples implicitly into a higher-dimensional feature space using a Kernel function. In the feature space, the examples (e.g. mutagenic and nonmutagenic compounds) are separated by a decision boundary. The SVM algorithm looks for the decision boundary that provides the best separation between both classes. SVMs have the advantage that the risk of overfitting is reduced and that they are able to model very complex decision boundaries. The disadvantage is that they are hard to interpret in the case of nonlinear Kernel functions (e.g., polynomial Kernels of degree greater than one). Also, the proper choice of a Kernel function and parameter settings is far from obvious.

2.5. Validation. 10-Fold cross-validation was used to validate the results of machine learning experiments. This means that the whole data set is divided into 10 parts. One part is removed as a test set, the remaining 9 parts are the learning set for the model. Predictions from the model for the test set are compared with the real values from the test set to estimate the predictive accuracy. The whole process is repeated 10 times, so that each part has served once as a test set and predictions for all compounds in the data set are available. For each validation run we report accuracy, sensitivity, and specificity using the definitions of Table 1.

Table 1. Definition of Variables for 10-Fold Cross-Validation[a]

	actual value		
	mutagen	nonmutagen	
mutagen	TP	FP	predicted
nonmutagen	FN	TN	value

[a] TP: true positives. FP: false positives. TN: true negatives. FN: false negatives. *C*: correct predictions (TP + TN). *A*: all predictions (TP + FP + TN + FN). Accuracy: *C*/*A*, i.e., the percentage of correct predictions. Sensitivity: TP/(TP + FN), i.e., the proportion of mutagens that are correctly predicted. Specificity: TN/(TN + FP), i.e., the proportion of nonmutagens that are correctly predicted.

Table 2. Performance of SAR Models for *Salmonella* Mutagenicity Using Fragments from Unoptimized Structuresy

threshold	algorithm	training set accuracy	mean accuracy	cross-validation mean sensitivity	mean specificity
0.01	J48	86.6959	75.0000	0.749	0.750
0.03	J48	86.9883	75.1462	0.746	0.756
0.05	J48	86.4035	77.0468	0.740	0.800
0.10	J48	82.4561	74.7076	0.755	0.739
0.01	PART	90.9357	77.4854	0.769	0.780
0.03	PART	88.8889	72.6608	0.720	0.733
0.05	PART	88.3041	76.3158	0.772	0.753
0.10	PART	84.5029	73.6842	0.743	0.730
0.01	SMO,E1	95.3216	77.6316	0.784	0.768
0.03	SMO,E1	86.5497	76.9006	0.778	0.759
0.05	SMO,E1	85.6725	77.4854	0.778	0.771
0.10	SMO,E1	77.7778	72.2222	0.708	0.736
0.01	SMO,E2	98.6842	75.1462	0.749	0.753
0.03	SMO,E2	97.0760	73.8304	0.743	0.733
0.05	SMO,E2	95.9064	73.3918	0.725	0.741
0.10	SMO,E2	91.2281	72.5146	0.725	0.724

3. RESULTS

The experiments were organized as follows:

The first experiment described chemical structures in terms of MOLFEA generated fragments and used three machine learning algorithms (C4.5, PART, SVM) to learn SAR models. To identify relevant fragments we used the MOLFEA query ($freq(f, mutagens) \geq t) \wedge freq(f, nonmutagens) \leq t$), i.e., we were looking for fragments that occur more than *t* times in mutagenic compounds and less than *t* times in nonmutagenic compounds. The complete version spaces of four frequency thresholds were used in our experiments: 0.01 (1%, 6 compounds), 0.03 (3%, 20 compounds), 0.05 (5%, 34 compounds), and 0.10 (10%, 68 compounds); the complete data set contained 341 mutagenic and 343 nonmutagenic compounds.

The machine learning algorithms were used with their default settings in the WEKA-Workbench. Support vector machines were used with a linear and quadratic kernel. The results of these experiments are summarized in Table 2.

In the next step we tried to improve and correct the chemical structures. Although we are using only linear fragments, it is important to obtain a consistent representation (e.g. in respect to aromaticity, protonation status, tautomers, ...) of all molecules in the training set to obtain correct results. As a manual inspection of structures is very error-prone and basic quality checks[8] have been done already on the initial structures, we performed semiempirical quantum mechanical optimizations (which was also needed for the calculation of molecular properties). Although the optimizations were performed in the gas phase and 3D structures were converted back to 2D for fragment generation, we hoped that this procedure would give a more consistent representation of the compounds than the original data set (e.g. in respect to aromaticity, protonation status, tautomers, ...). The parameters for fragment generation and machine learning were identical with the first experiment, and the results can be found in Table 3.

Table 3. Performance of SAR Models for *Salmonella* Mutagenicity Using Fragments from MOPAC Optimized Structures

threshold	algorithm	training set accuracy	mean accuracy	cross-validation mean sensitivity	mean specificity
0.01	J48	86.8421	75	0.714	0.785
0.03	J48	87.4269	73.3918	0.720	0.747
0.05	J48	86.9883	75.8772	0.743	0.774
0.10	J48	82.6023	73.9766	0.734	0.744
0.01	PART	91.3743	73.6842	0.731	0.741
0.03	PART	90.3509	74.1228	0.755	0.727
0.05	PART	89.1813	74.7076	0.720	0.774
0.10	PART	86.4035	72.076	0.708	0.733
0.01	SMO,E1	95.3216	76.6082	0.752	0.780
0.03	SMO,E1	86.8421	74.269	0.746	0.739
0.05	SMO,E1	85.0877	78.5088	0.775	0.794
0.10	SMO,E1	77.4854	73.6842	0.743	0.730
0.01	SMO,E2	98.6842	76.462	0.763	0.765
0.03	SMO,E2	96.9298	74.5614	0.731	0.759
0.05	SMO,E2	95.7602	73.538	0.725	0.744
0.10	SMO,E2	91.3743	72.5146	0.725	0.724

Table 4. Molecular Properties Used in SAR Models for *Salmonella* Mutagenicity

parameter	program
accessible surface (all atoms)	NACCESS
accessible surface (nonpolar atoms)	NACCESS
accessible surface (polar atoms)	NACCESS
dipole	MOPAC
electronic energy	MOPAC
electronegativity	MOPAC
heat of formation	MOPAC
HOMO	MOPAC
(HOMO − LUMO)/2	MOPAC
hybridization dipole	MOPAC
ionization potential	MOPAC
largest interatomic distance	MOPAC
logP	KOWWIN
LUMO	MOPAC
molecular weight	MOPAC
point charge dipole	MOPAC
total energy	MOPAC

Table 5. Performance of SAR Models for *Salmonella* Mutagenicity Using Molecular Properties

algorithm	training set accuracy	mean accuracy	cross-validation mean sensitivity	mean specificity
J48	82.1637	63.1579	0.699	0.563
PART	71.7836	64.6199	0.731	0.560

In the third experiment we developed SARs using molecular properties (Table 4) alone. As our implementation of support vector machines had performance problems with this type of data (real numbers instead of binary fragment fingerprints), only C4.5 and PART were applied to learn the models. The results are summarized in Table 5.

The final step was to use both fragment data and molecular properties. Again, C4.5 and PART were used for the machine learning experiments. Tables 6 and 7 summarize the results;

Table 6. Performance of SAR Models for *Salmonella* Mutagenicity Using Fragments from Unoptimized Structures and Molecular Properties

threshold	algorithm	training set accuracy	mean accuracy	cross-validation mean sensitivity	mean specificity
0.01	J48	93.2749	75	0.734	0.765
0.03	J48	92.5439	73.8304	0.711	0.765
0.05	J48	94.0058	74.4152	0.728	0.759
0.10	J48	90.0585	71.345	0.711	0.715
0.01	PART	95.614	75.5848	0.731	0.780
0.03	PART	94.0058	75.731	0.763	0.750
0.05	PART	95.614	74.269	0.749	0.736
0.10	PART	96.0526	71.7836	0.734	0.700

Table 7. Performance of SAR Models for *Salmonella* Mutagenicity Using Fragments from MOPAC Optimized Structures and Molecular Properties

threshold	algorithm	training set accuracy	mean accuracy	cross-validation mean sensitivity	mean specificity
0.01	J48	92.9825	75.4386	0.720	0.788
0.03	J48	93.5673	73.8304	0.723	0.753
0.05	J48	92.8363	72.6608	0.702	0.750
0.10	J48	90.6433	73.2456	0.714	0.750
0.01	PART	95.0292	74.5614	0.737	0.753
0.03	PART	94.2982	72.807	0.725	0.730
0.05	PART	95.3216	74.1228	0.766	0.715
0.10	PART	96.345	74.5614	0.731	0.759

Table 6 uses the original structures for fragment generation, whereas Table 7 uses optimized structures.

4. DISCUSSION

4.1. Performance. The first question under investigation was whether molecular properties or molecular fragments are better descriptors for mutagenicity SARs with noncongeneric compounds. A comparison of Table 5 with Tables 2 and 3 clearly indicates that models based on molecular fragments give much more accurate predictions (up to 28% above default) than models based on our set of molecular properties (up to 14 above default). It is of course possible (and quite likely) that we did not choose the "correct" set of molecular properties that are relevant for *Salmonella* mutagenicity. [Please note that the machine learning techniques in our investigation are, at least theoretically, robust towards correlated and unnormalized data.] Under practical circumstances it is however impossible to determine the relevant properties a priori, because of the complexity of the involved biochemical processes.

Although our set of molecular properties is less useful than molecular fragments, they might provide information, which can be used in conjunction with fragments. Therefore we performed a set of experiments using both types of descriptors. The results are summarized in Tables 6 and 7. Comparing them with Tables 2 and 3 shows that there is no dramatic improvement in terms of predictive accuracy in the cross-validation experiments. In fact, the highest accuracies were obtained with fragments alone. It is interesting to note that the accuracies on the training set are generally higher with molecular properties, which indicates an overfitting of the data. It is also possible that the addition of molecular properties provides only redundant data. Many global properties are influenced by the presence of certain substructures, which might be already represented by one of the fragments. The calculation of logP, for example, checks for predefined substructures to calculate the overall value for the whole molecule.[12] Although quantum-mechanical optimizations do not consider substructures, the presence of certain structural features influences electronic properties implicitly.

Summing up, our experiments indicate that accurate mutagenicity SAR models can be generated from molecular fragments alone. If this holds true for other toxic endpoints, then further molecular properties have to be clarified by further investigations.

The next question was if semiempirical quantum mechanical optimizations help to obtain a more consistent representation of chemical structures and improve the derived SAR models. The results are ambiguous: Although the highest predictive accuracies (78.5%) were obtained with MOPAC optimized structures, the overall picture is different. In 8/16 cases unoptimized structures gave better results, in 6/16 cases optimization gave an improvement (mainly for support vector machines), in 2/16 cases the accuracies were identical. Generally the difference between both models is only a few percentage points. Considering the computational overhead for quantum mechanical calculations and the possibility of introducing errors by format conversions and nonparametrized elements, it is unlikely that this procedure improves the performance of the derived SAR models.

We did not obtain a value for the frequency threshold for fragment generation which is optimal under all circumstances. Thresholds of 0.05 gave generally good results in our case, but we recommend to perform experiments with new data sets. Frequency thresholds between 0.01 and 0.05 seem to be good starting points.

For the learning algorithm there is also not a single solution that is optimal in all cases. Most of the times linear support vector machines and the PART rule learner perform better than C4.5 and support vector machines with quadratic kernels. Quantitatively the highest predictive accuracies (78.5%) were obtained with linear support vector machines and optimized structures. For unoptimized structures the performance of PART and support vector machines were equivalent (77.5%). The differences between algorithms are generally not very pronounced. The choice of the machine learning algorithm can depend therefore more on the scope of the investigation (e.g. prediction of untested compounds, development of mechanistic hypothesis, computer aided drug design) than on performance issues.

In all models the *sensitivity* (i.e. proportion of correctly predicted mutagens) and *specificity* (i.e. proportion of correctly predicted nonmutagens) values are very similar. This indicates well balanced models that are capable of predicting mutagenicity and nonmutagenicity with a similar degree of confidence. As the predictive accuracy is substantially higher than the default guess of the majority class (nonmutagens) of the database (50.15%) and close to the experimental reproducibility of the *Salmonella* assay (~85%[16]) we conclude that we were able to extract significant knowledge out of empirical data.

It is hard to perform a quantitative comparison of our results with other investigations in this area, because the data sets are not identical. Table 8 summarizes the results with

Table 8. Performance of SAR Models for *Salmonella* Mutagenicity Reported in the General Literature

author	citation	method	mean accuracy (%)
Perotta et al.	17		73.9[a]
Klopman and Rosenkranz	18	CASE	72
Klopman and Rosenkranz	18	MULTICASE	80
Klopman and Rosenkranz	18	CASE/GI	47

[a] Leave-one-out validation.

noncongeneric mutagenicity data from the literature. It indicates clearly that the quantitative performance of our models is at least competitive with the best published results so far.

4.2. Comparison with Other Fragment Based Approaches. At a first glance our strategy may seem to be very similar to the CASE[19] and MULTICASE[20] programs from Gilles Klopman and the reimplementations of the same concept by Malacarne et al.[21] In fact we derived a lot of inspiration by this pioneering work, but from the computer science point of view, there are substantial differences that we want to discuss briefly:

Both CASE and MULTICASE are monolithic programs that integrate Fragment Generation and Prediction. We could not find any detailed information about the fragment generation algorithm in the literature, but we assume that the procedure is similar in both programs. CASE/MULTICASE fragments are in principle linear, but they support some branches at their backbone. The classification algorithm makes both programs different: CASE uses a Bayesian approach to aggregate the contributions of all fragments, whereas MULTICASE employs a more structured divide-and-conquer strategy to distinguish between major biophores that provide a primary classification and modulators (properties and fragments) that are capable to up- and downregulate the activity of the primary biophore.

We use in contrast a flexible modular strategy and different algorithms: descriptor generation and classification are completely decoupled, and it is possible to use various techniques for both of the steps. For feature generation we have investigated as a starting point linear fragments and a few molecular properties, but there are a lot of further possibilities. Apart from MOLFEA extensions to 3D-fragments[22] and arbitrary substructures that are currently investigated it is possible to use all kinds of chemicals descriptors from computational chemistry[23] as well as predefined structural features (e.g. structural alerts[24]). In principle it is even possible to use spectroscopic data or biological activities (e.g. from surrogate assays) to characterize chemicals.

The classification step is also extremely flexible. For this study we have selected three techniques (C4.5, PART, and SVMs) that seemed to be particularly promising for our purpose, but this is not the only possibility. Depending on the scope of the study (and personal preferences) it is possible to choose from a variety of statistical (e.g. multiple linear regression, principal component regression), probabilistic (e.g. naive Bayes), or connectionist (e.g. neural nets) techniques for the classification of new instances and/or the prediction of biological activities (e.g. LC_{50}'s, EC_{50}'s).

4.3. Interpretation and Application of SAR Models. As an illustration how to interpret and use SAR Models from data mining and machine learning experiments, we will use the model created by linear support vector machines from fragments with a threshold of 0.05 as an example.

```
1.6274328192175296 * c:c:c:c:c:c:c:c:c
1.4455302626881337 * C-Cl
1.3226667063998578 * C-C-C-C-N-C
1.3105243804180 45  * C-C-C-O
0.9516054404252757 * C-C=C
0.8654786477941714 * c:c:c:c:c:n
0.8243351055367271 * C-C-C-C=C
0.8197902253156605 * C-C-C-N-C
0.7969086522621357 * c:c:c-C=O
0.7819601605449131 * C-N-C
0.7796980414561107 * N-N
0.7498673413287917 * C-C-C-C-O
0.7276759799450657 * C-C-N-N
0.727514353351238  * N-O
0.7167965121501293 * C-O-C
0.6784153103780268 * C
0.6744410897500348 * C-N-c:c:c:c:c:c
0.6744410897500348 * C-N-c:c:c:c:c
0.6716119052528489 * c:c-N
0.5686660779334143 * C-C-N
```

Figure 7. The 20 strongest activating fragments for *Salmonella* mutagenicity derived from linear support vector machines. Fragments are written in SMARTS notation: uppercase letters: aliphatic atoms, lowercase letters: aromatic atoms, - single bond, : aromatic bond, = double bond; baseline value: −0.24.

```
-1.479449078121286  * Cl-C-Cl
-1.4528269249653274 * C-C-C=C-C
-1.0145947939115687 * C-N-c:c
-1.0145947939115687 * C-N-c:c:c
-0.9492959881012157 * C-C
-0.9474885039899876 * C-C-N-C
-0.9402207493855474 * C-O-C=O
-0.937214552267573  * c:c:c:c:c:c-S
-0.937214552267573  * c:c:c:c:c-S
-0.937214552267573  * c:c:c:c-S
-0.9115486314638905 * C-C-C-C=O
-0.8877782140374197 * C-C-C-C
-0.8678653536715137 * c:c:c:c:c:c:c:c:c:c
-0.8568018292049271 * c:c:n:c:c
-0.7574483341970001 * Cl
-0.7529686472886363 * O-C=O
-0.7447971289365931 * C-C-C-N
-0.7285699786145916 * O
-0.7168970154384797 * C-C-C-C-C
-0.6759056107684382 * c:n
```

Figure 8. The 20 strongest deactivating fragments for bacterial mutagenicity derived from linear support vector machines. Fragments are written in SMARTS notation: uppercase letters: aliphatic atoms, lowercase letters: aromatic atoms, - single bond, : aromatic bond, = double bond; baseline value: −0.24.

Figure 7 lists the 20 most important fragments contributing to mutagenicity, and Figure 8 lists the 20 most important fragments indicating nonmutagenicity. The total number of attributes in this SAR model is 171. Each relevant fragment has an associated weight factor, which may be positive (i.e. the fragment contributes to mutagenicity, biophores in CASE[25] terminology) or negative (i.e. the fragment reduces mutagenicity, biophobes in CASE[25] terminology). [In our model we did not consider fragment frequencies in molecules, so fragments are either present (1) or absent (0) in a given compound.] To make a prediction for a new compound, it is necessary to generate the fragments for this molecule, to look for the weights associated with the presence

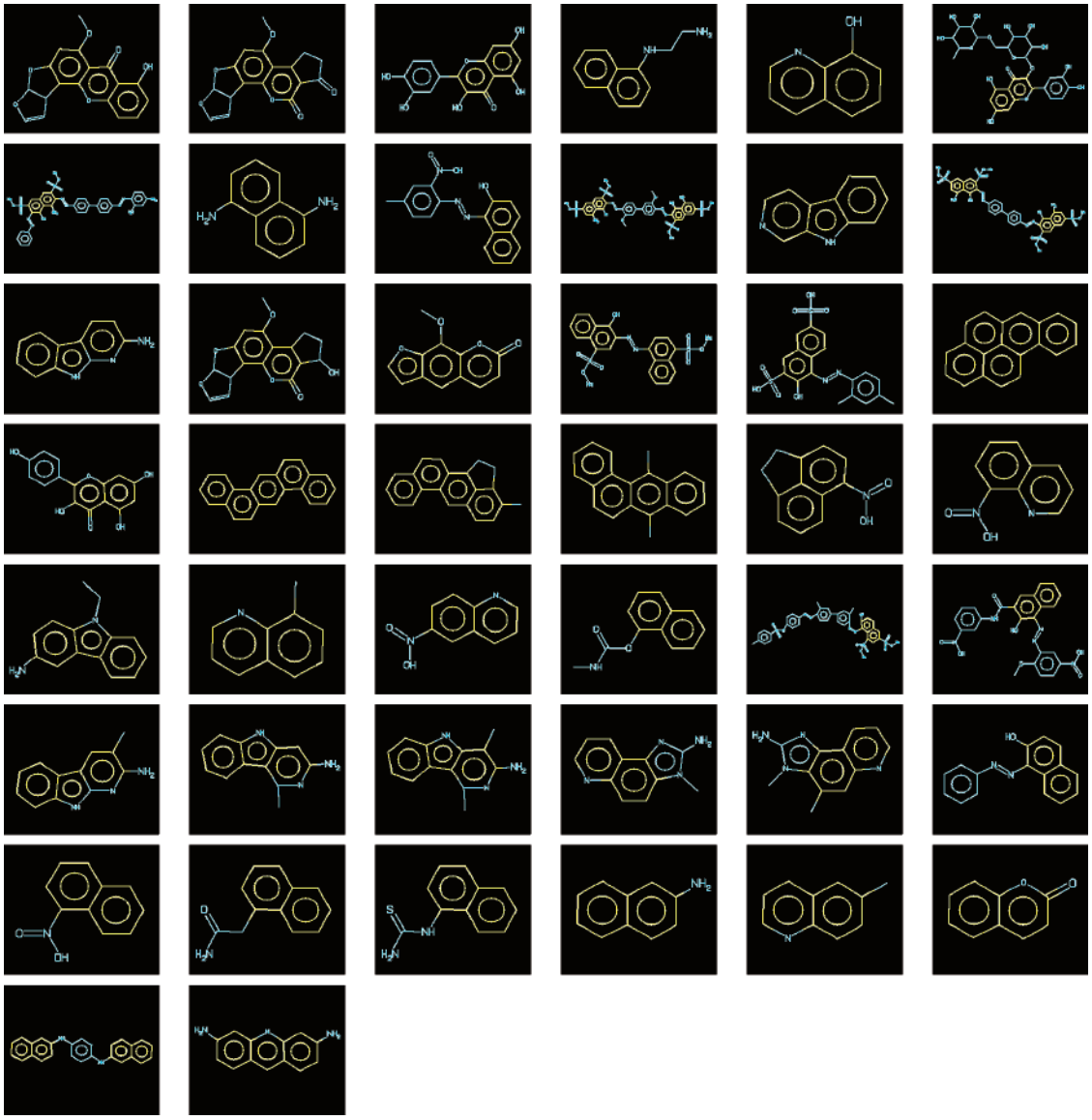


Figure 9. Mutagenic compounds containing the fragment c:c:c:c:c:c:c:c:c:c. Atoms matching this fragment are marked in yellow.

Table 9. Interpretation of Example Fragments

fragment	interpretation
c:c:c:c:c:c:c:c:c	9 connected aromatic carbons, indicates condensated aromatic ring systems
C−Cl	chlorinated aliphatic carbon
Cl−C−Cl	dichlorinated aliphatic carbon, compensates the effect of C−Cl
C−N-c:c	aliphatic carbon attached via nitrogen to a carbon in an aromatic system
c:c:c:c:c:n	aromatic system with 5 carbons and 1 nitrogen (usually pyridine)

of these fragments, to sum them up, and add them to the baseline value (−0.24). The resulting value is used to predict the mutagenic activity: positive values indicate mutagenicity, negative values indicate nonmutagenicity. This means that the presence of one or more activating fragments is not a sufficient criteria for mutagenicity. Activating fragments may be compensated by inactivating fragments and vice versa.

Table 9 gives a few examples of how to interpret some of the fragments in this model in a chemical context. As this requires some experience it is often advantageous to match interesting fragments on chemical structures visually (using e.g. the Daylight's depictmatch web interface http://http://www.daylight.com/cgi-bin/contrib/depictmatch.cgi). Figure 9 for example depicts all mutagenic molecules that contain the strongest activating fragment c:c:c:c:c:c:c:c:c of our model. The areas where this fragment matches are marked in yellow.

To demonstrate the practical utility of our approach we will show with a hypothetical example, how to use the derived SAR models to modify a mutagenic compound into a nonmutagenic one. We have randomly selected one compound from our data set (melphalan, CAS 148-82-3, Figure 10), and we will use the same model as before. [This is just a hypothetical example to illustrate the concept. We do not claim any particular biological or chemical relevance for this example.]

Figure 10 lists the activating and deactivating fragments that are found in this compound, together with their weights. The resulting value is almost equal to 1—a strong indication of mutagenicity (positive values indicate mutagenicity, negative values nonmutagenicity). In a first attempt to reduce mutagenicity we can remove two chlorines (the most important contributers to mutagenicity in this compound),

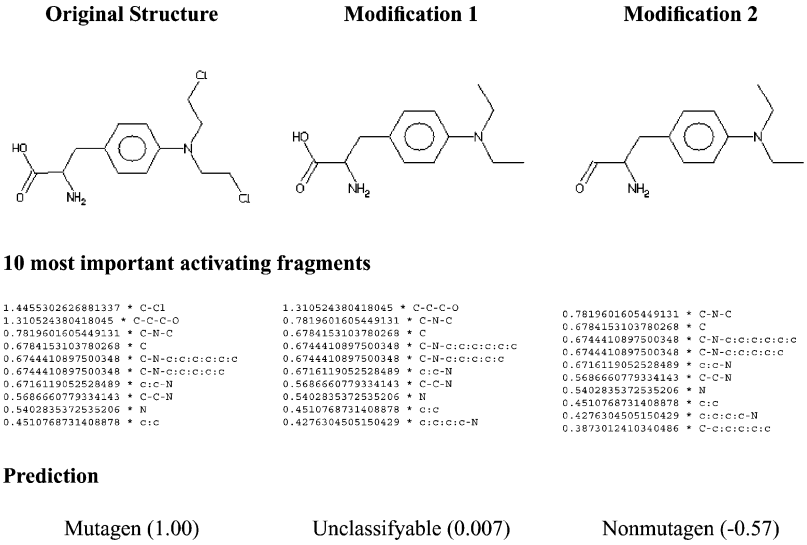


Figure 10. The consequences of removing activating fragments from melphalan (CAS 148-82-3).

leading to the second structure in Figure 10. This causes a substantial drop of the predicted value to 0.007. As this value is very close to zero, we cannot make a very reliable prediction, but because of the positive value we will still have to classify this compound as a mutagen—although with low confidence. As a next step, we may want to change the carbonyl group to an aldehyde. This leads to the third structure in Figure 10 with a predicted value of −0.57—probably a nonmutagen.

In a similar spirit we can add some of the deactivating structures from Figure 8 to remove mutagenic activity. The application of this concept under real world conditions is a bit more complicated, because it is necessary to conserve (or improve) structural features, that are responsible for desired (e.g. pharmacological, ADME) effects. If we have several SAR models for different endpoints, it is possible to combine them with a scoring function, that considers the balance between (desired and undesired) effects.

5. CONCLUSION

With this investigation we have demonstrated the utility of the inductive database MOLFEA for the generation of descriptors for SAR studies with mutagenic compounds. These descriptors can be used e.g. by machine learning techniques to create reliable SARs for noncongeneric compounds. We have further demonstrated how to interpret and use SAR models induced by support vector machines in a practical context.

Initial experiments with other endpoints indicate that a similar procedure can be used for other data sets as well, but a definitive answer requires more experiments with diverse data sets. We are presently extending the MOLFEA framework for the identification of three-dimensional fragments[22] and arbitrary substructures (graphs) within molecules. This should enable us to deal more efficiently with receptor interactions and stereochemistry. As soon as more public data from toxicogenomics, -proteomics, and -metabolomics experiments become available it will be possible to include more biological information in SAR studies. This is also an open research issue, but the reward will be models that consider individual susceptibilities than models based on present data.

ACKNOWLEDGMENT

Thanks to our students Steven Ganzert, Daniel Moeller, Stefan Mutter, and Marcin Nadolny for their initial experiments with this data set.

Supporting Information Available: The data sets used in this article are available from http://www.predictive-toxicology.org/data/cpdb_mutagens/. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Livingstone, D. J. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.
(2) Witten, I. H.; E. Frank. *Data Mining*; Morgan Kaufmann Publishers: San Francisco, CA, 2000.
(3) Kramer, S.; Frank, E.; Helma, C. Fragment generation and support vector machines for inducing SARs. *SAR QSAR Environ. Res.* **2002**, *13*, 509−523.
(4) Helma, C.; Kramer, S.; DeRaedt, L. The molecular feature miner molfea. In Hicks, M., Ed.; *Proceedings of the Beilstein Workshop 2002: Molecular Informatics: Confronting Complexity*; Beilstein Institut: Frankfurt, 2003.
(5) Gold, L. S.; Zeiger, E. *Handbook of Carcinogenic Potency and Genotoxicity Databases*; CRC Press: 1997.
(6) Ames, B. N.; Durston, W. E.; Yamasaki, E.; Lee, F. D. Carcinogens are mutagens: A simple test system combining liver homogenates for activation and bacteria for detection. *Proc. Natl. Acad. Sci. U.S.A.* **1973**, *70*, 2281−2285.
(7) Weininger, D. SMILES, a chemical language and information system 1. Introduction and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(8) Helma, C.; Gottmann, E.; Kramer, S. Knowledge discovery and data mining in toxicology. *Stat. Methods Med Res.* **2000**, *9*, 329−358.
(9) DeRaedt, L. A perspective on inductive databases. *SIGKDD Explorations* **2002**, *4*, 69−77.
(10) Kramer, S.; De Raedt, L.; Helma, C. Molecular feature mining in HIV data. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-01)*; 2001; pp 136−143.
(11) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput. Method.* **1990**, *3*, 537−547.
(12) Meylan, W. M.; Howard, P. H. Atom/Fragment contribution method for estimating octanol−water partition coefficients. *J. Pharm. Sci.* **1995**, *84*, 83−92.
(13) Quinlan, J. R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann: San Mateo, CA, 1993.
(14) Frank, E.; Witten, I. H. Generating accurate rule sets without global optimization. In *Proceedings of the Fifteenth International Conference on Machine Learning*; Morgan Kaufmann Publishers: San Francisco, CA, 1998.

(15) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer: Berlin, Heidelberg, New York, 1995.
(16) Benigni, R.; Giuliani, A. Computer assisted analysis of interlaboratory Ames test variability. *J. Toxicol. Environ. Health* **1988**, *1*, 135−146.
(17) Perotta, 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.
(18) Klopman, G.; Rosenkranz, H. S. Testing by artificial intelligence: Computational alternatives to the determination of mutagenicity. *Mutat. Res.* **1992**, *272*, 59−71.
(19) 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.
(20) Klopman, G. MultiCASE: A hierarchical computer automated structure evaluation program. *Quant. Struct.-Act. Relat.* **1992**, *11*, 176−184.
(21) 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.
(22) Hill, A. Erweiterung des Molecular Feature Miners für 3-dimensionale fragmente, Master's Thesis, Universität Freiburg, 2002.
(23) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; VCH: Weinheim, 2000.
(24) Ashby, J.; Paton, D. The influence of chemical structure on the extent and sites of carcinogenesis for 522 rodent carcinogens and 55 different human carcinogen exposures. *Mutat. Res.* **1993**, *286*, 3−74.
(25) Klopman, G.; Rosenkranz, H. S. Approaches to SAR in carcinogenesis and mutagenesis. Prediction of carcinogenicity/mutagenicity using MULTI-CASE. *Mutat. Res.* **1994**, *305*, 33−46.

CI034254Q