Running Title: lazar Carcinogenicity Predictions

Lazy Structure-Activity Relationships (lazar) for the Prediction of Rodent Carcinogenicity and *Salmonella* Mutagenicity.

Christoph Helma[1,2]

[1] in silico toxicology, Talstraße 20, D-79102 Freiburg, Germany

[2] Machine Learning Lab, University Freiburg, Georges Köhler Allee 79,

D-79110 Freiburg, Germany

Phone: +49-160-94623203

Email: helma@in-silico.de

September 27, 2005

Abstract

`lazar` is a new tool for the prediction of toxic properties of chemical structures. It derives predictions for query structures from a database with experimentally determined toxicity data. `lazar` generates predictions by searching the database for compounds that are similar *with respect to a given toxic activity* and calculating the prediction from their activities. Apart form the prediction, `lazar` provides the rationales (structural features and similar compounds) for the prediction and a reliable confidence index that indicates, if a query structure falls within the applicability domain of the training database.

Leave-one-out (LOO) crossvalidation experiments were carried out for 10 carcinogenicity endpoints ($\{female|male\}$ $\{hamster|mouse|rat\}$ $carcinogenicity$ and aggregate endpoints $\{hamster|mouse|rat\}$ $carcinogenicity$ and $rodent$ $carcinogenicity$) and *Salmonella* mutagenicity from the Carcinogenic Potency Database (CPDB). An external validation of *Salmonella* mutagenicity predictions was performed with a dataset of 3895 structures. Leave-one-out and external validation experiments indicate that *Salmonella* mutagenicity can be predicted with 85% accuracy for compounds within the applicability domain of the CPDB. The LOO accuracy of `lazar` predictions of rodent carcinogenicity is 86%, the accuracies for other carcinogenicity endpoints vary between 78 and 95% for structures within the applicability domain.

Key Words Applicability Domain, Carcinogenic Potency Database, Data Mining, `lazar`, Predictive Toxicology, (Quantitative) Structure-Activity Relationships

Abbreviations

CCRIS Chemical Carcinogenesis Research Information System

CPDB Carcinogenic Potency Database

DSSTox Distributed Structure-Searchable Toxicity Project

`lazar` Lazy Structure-Activity Relationships

LOO Leave-One-Out Crossvalidation

k-nn k-Nearest-Neighbours

(Q)SAR (Quantitative) Structure-Activity Relationships

Background

Chemical and pharmaceutical industries, regulatory agencies and research institutions need techniques that are capable of identifying adverse effects at a very early stage of product development and provide reasonable toxicity estimates for the huge number of untested compounds. This information comes traditionally from *in vivo* testing, but the public pressure to reduce animal experiments and the lack of important toxicity information for many old compounds has led to an increased acceptance of alternative (*in vitro* and *in silico*) methods. Computer based (*in silico*) techniques are particularly appealing for this purpose, because they are extremely fast and cost efficient and can be applied even when a compound is not physically available.

The problem of predicting toxic activities from chemical structures can be approached from different directions [1], e.g. with statistical *(Quantitative)Structure-Activity Relationships ((Q)SAR)* techniques [2], by developing expert systems [3] or with the application of data mining and machine learning techniques [4]. This paper presents a new approach for this purpose that uses an *Inductive Database* [5, 6]. Inductive Databases can be queried not only for data (as traditional databases), but also for regularities and patterns within the data. `lazar` (*Lazy Structure Activity Relationships*) is a special-purpose extension of this concept, because it allows to specify a chemical structure and to query for its potential biological activities. This paper presents the algorithms that are used by the `lazar` engine to solve queries for toxic activities and presents an exemplary validation study for rodent carcinogenicity and *Salmonella* mutagenicity.

Improving predictive accuracy (as determined by cross-validation or validation with an external test set) has been for long the main driving force for the development of new *in silico* prediction techniques. The quest for higher and higher predictive accuracies leads however frequently to overfitted models that perform well on cross-validation or on a particular test set but fail completely on unknown compounds [7–9]. The working hypothesis during `lazar` development was that inaccurate predictions are frequently not the result of poor algorithms, but of insufficient information in the database or of inaccurate experimental

measurements. The main goal was to develop a system that is capable of

- working with databases of structurally diverse (*non-congeneric*) compounds (that do not act by a common biochemical mechanism)

and of providing for each prediction

- the rationales that led to the prediction and
- an indication of the reliability of the prediction.

These features shall prohibit the naive trust in every prediction and ensure that predictions are amenable to critical evaluations from toxicological experts. Please note that the algorithm presented in this article differs substantially from previous versions of `lazar` [10, 11] that were based on Bayesian classification/regression.

Methods

The `lazar` algorithm

Overview

`lazar` does not create a global (Q)SAR model that is valid for all instances, but it derives its prediction specifically for a *query* structure with a modified *k-nearest-neighbour (k-nn)* algorithm. For this purpose `lazar` searches a database with chemical structures and experimental data (*training set*) for compounds that are similar to the query structure (*neighbours*) and calculates a prediction from the experimental measurements of the neighbours. In contrast to traditional k-nn techniques `lazar` considers chemical similarities not as absolute values, but as values that have to be determined *with respect to a given biological activity* (Figures 1,2). The prediction of the toxicity of a query compound requires four steps that will be described in detail in the next sections:

1. Determination of features that characterise the structures of the query compound and the compounds in the training set

2. Selection of features that are relevant for the toxic endpoint under investigation

3. Identification of neighbours in the training set

4. Calculation of qualitative (*classification*) or quantitative (*regression*) predictions [1]

A formal representation of the complete algorithm is summarised in Figures 3 and 4 and a screenshot of the web interface can be found in Figures 1 and 2.

FIGURE 1 - 2

FIGURE 3 - 4

Toxicity related chemical similarity

Similarity searching in chemical databases is an important topic in chemoinformatics research, an excellent review of this subject can be found in an article by Willet et al. [12]. Most of these techniques do not work directly with chemical graphs, but with a limited number of predefined substructures (*fragments*). Most similarity indices rely on the number of fragments that are shared between the structures and the number of fragments that occur only in a single structure. These numbers are summarised into a single index value, e.g. the Tanimoto index (Equation 1).

For the determination of toxicity related chemical similarities it is important to consider only those fragments, that are relevant for the toxic endpoint under investigation (i.e. only those parts of the chemical structures that are involved in chemical reactions and transport processes that lead to toxicity or to detoxification). The crucial task is therefore to identify these fragments in an efficient and reliable manner.

The classical strategy to derive toxicity related substructures is to consult the literature and domain experts for the biochemical mechanisms that lead to a particular toxic effect and to define *structural alerts* for a particular endpoint. It is however likely that a predefined set of structural alerts is incomplete (or maybe wrong), because many toxicity mechanism are still poorly understood or even unknown. This work introduces an alternative approach for the determination of toxicity related chemical similarities that relies on fragment languages (e.g. linear fragments, trees, subgraphs). With the help of Data Mining algorithms

[1] A regression algorithm is available in the current `lazar` version, but it will not be a subject of this article.

it is possible identify relevant fragments from a given language automatically from the training data. This procedure saves error prone and expensive human work and some algorithms can even guarantee that no relevant feature of the given language can be missed.

Determination of features `lazar` uses at present predominantly the language of *linear fragments* for the identification of toxifying and detoxifying substructures [13]. Linear fragments are defined as chains of heavy (non-hydrogen) atoms with connecting bonds, without branches or cycles. All linear fragments that are present in the query structure or in one of the training structures are determined exhaustively by a simplified version of the MOLFEA algorithm [13]. This step does not consider biological activities, the relevant features are identified by the feature selection process described below. As all possible linear fragments are evaluated, no relevant linear fragment can been missed.

Although linear fragments seem to be limited at a first glance (no explicit consideration of branches or cycles), they perform remarkably well on a variety of toxicity endpoints. A possible reason is that a lot of chemical information is implicitly contained in these fragments [2] and the "chemical context" is considered by the neighbourhood based prediction algorithm. `lazar` has furthermore the possibility to derive linear fragments not only from the table of elements, but also from arbitrary SMARTS (`http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html`) expressions. With such an alphabet we have the facility to consider chemically relevant concepts like local chemical properties (e.g. H-bond donor/acceptor), branching, presence in rings, rotable bonds or even stereochemistry.

We have also explored extensions of the fragment language towards 3D fragments [14] and arbitrary subgraphs [15]. Up to now, the authors experience with various public toxicity datasets did not require the necessity to implement such a computationally expensive framework. It is however important to note that the feature selection and prediction algorithms presented below are independent of the fragments that characterise the chemical structures. `lazar` may use therefore also other chemical features like multiple neighborhoods of atoms [16]/ augmented atoms [17], spectra or results from short term assays. An extension towards quantitative molecular descriptors (e.g. HOMO, LUMO, logP) is also relatively straight-

[2] An aromatic atom e.g. is an indication of a ring system.

forward.

Linear fragments and structural alerts (both are presently implemented in `lazar`) can be used in conjunction. Statistical criteria (see below) can be used to decide if a fragment is relevant in regard to toxicity or not. This is a valuable tool for *hypothesis testing*.

Feature selection The goal of the feature selection step is the identification of fragments that are relevant for the toxic activity under investigation.

The relevance of fragments for a given toxic activity can be determined with simple statistical tests. `lazar` uses the chi-square test to identify fragments that occur significantly more frequent in toxic compounds than in non-toxic compounds (or vice versa) and to calculate their statistical significance p_f. Very significant features have a higher impact on chemical similarities than features with low significance (Equation 1). For efficiency reasons fragments below a predefined threshold ($p_f < 0.95$ [3]) are discarded from further calculations.

As it may happen that a query structure has to be removed from the training structures (e.g. for validation purposes), activity information in the training database may change when multiple compounds are predicted. It is therefore essential that relevant features are identified for each query compound separately. Precomputing relevant fragments in a single preprocessing step can lead to overly optimistic validation results.

Calculation of activity related similarity As soon as all relevant (i.e. statistically significant) fragments have been identified for the training set D and the query structure s_q, it is possible to determine the similarities between the query structure s_q and all training structures $s_t \in D$. If s_q and s_t contain the same set of fragments, they will be considered as equal (with respect to the given activity) with a similarity of 1, if they share no common fragment, they will be considered as unequal with a similarity of 0. Taking into account the statistical significances p_f of the significant fragments F we can define a similarity index $sim(s_q, s_t, D)$ (weighted Tanimoto index) for structures s_q and s_t with respect to the training database D as:

[3] No optimisation of this parameter was performed to avoid overfitting.

$$sim(s_q, s_t, D) = \frac{\sum_{f \in F} \{{p_f}^4 | f \subseteq s_q \wedge f \subseteq s_t\}}{\sum_{f \in F} \{{p_f}^4 | f \subseteq s_q \vee f \subseteq s_t\}} \quad (1)$$

with

$p_f|f \subseteq s_q \wedge f \subseteq s_t \ldots$ significance of fragment f that occurs in s_q *and* s_t

$p_f|f \subseteq s_q \vee f \subseteq s_t \ldots$ significance of fragment f that occurs in s_q *or* s_t [4]

$F \ldots$ set of significant features

Prediction

To obtain the prediction for a query structure, toxicity related similarities are computed for each compound in the training set. For efficiency reasons only instances of the training set with a similarity above a predefined threshold ($sim > 0.3$ [5]) are considered as *neighbours* to the query structure. Predictions are derived from all neighbours (N) of a query structure.

Classification To classify a query structure `lazar` uses a weighted majority vote from all neighbours. For this purpose we can define a confidence measure $conf$ that indicates the expected class and the reliability of the prediction as

$$conf = \frac{\sum_{n \in N} \{sim_n | t_n = "active"\}^4 - \sum_{n \in N} \{sim_n | t_n = "inactive"\}^4}{|N|} \quad (2)$$

with

$sim_n|t_n = "active" \ldots$ similarity of active neighbour n

$sim_n|t_n = "inactive" \ldots$ similarity of inactive neighbour n [6]

$N \ldots$ set of neighbours

$|N| \ldots$ number of neighbours

[4] The exponent of 4 ensures that the contribution of fragements decreases exponentially with their significance. No optimisation of this parameter was performed to avoid overfitting.

[5] No optimisation of this parameter was performed to avoid overfitting.

[6] The exponent of 4 reduces the weight of dissimilar neighbours. No optimisation of this parameter was performed to avoid overfitting.

A query structure is classified as active, if $conf > 0$ and as inactive, if $conf < 0$. This confidence measure considers contradictory examples in the training set as well as the similarities of these instances to the query structure. It is therefore a parameter that indicates the *applicability domain* of the test set.

Implementation

`lazar` was implemented in C++ using the Openbabel (`http://openbabel.sourceforge.net/`) and Gnu Scientific (GSL) (`http://www.gnu.org/software/gsl/`) Libraries. InChI codes (main layer) [18], an unique identifier for the connectivity of chemical structures, was used for the identification if identical structures. `lazar` was compiled with `gcc` on various Linux distributions, porting to other platforms should be possible, but has not been tested so far. `lazar` is available on request from the author, a web interface for `lazar` can be found at `http://www.predictive-toxicology.org/lazar/`. Figures 1 and 2 show a screenshots of the web interface.

Carcinogenic Potency Database (CPDB)

The *Carcinogenic Potency Database (CPDB)* `http://potency.berkeley.edu/cpdb.html` contains detailed results and analyses of more than 5000 chronic, long term carcinogenesis bioassays reported in over 1200 papers in the general literature and more than 400 Technical Reports of the National Cancer Institute/National Toxicology Program. For the purpose of this investigation the latest CPDB Summary Table provided by the Distributed Structure-Searchable Toxicity (DSSTox) project `http://www.epa.gov/nheerl/dsstox/` was used (`CPDBAS_v2a_1451_1Mar05.sdf`. It contains data for 1447 compounds with variable fractions of missing values for each endpoint.

Definition of endpoints

For the purpose of this study the following toxicity endpoints have been evaluated.

- $Rodent\ Carcinogenicity$,
- $\{Hamster|Mouse|Rat\}\ Carcinogenicity$,

- $\{Male|Female\}$ $\{Hamster|Mouse|Rat\}$ $Carcinogenicity$ and
- *Salmonella* $Mutagenicity$

An insufficient number of experimental results prevented reliable predictions for the remaining species in the CPDB (Cynomolgus, Dog, Rhesus) as well as the prediction of organ specific effects. Classifications for rodent carcinogenicity endpoints were obtained from the source data by applying the following criteria:

$\{Hamster|Mouse|Rat\}$ $Carcinogenicity$ Positive classification (1) if a TD_{50} value is available, negative classification (0) if no positive results are available (NP), inadequate studies have been excluded.

$Rodent$ $Carcinogenicity$ Positive classification (1) if compound is carcinogenic in at least one rodent species (see before), negative (0) if compound has at least one negative carcinogenicity classification (see before) and no positive classification.

$\{Male|Female\}$ $\{Hamster|Mouse|Rat\}$ $Carcinogenicity$ Positive classification (1) if the given sex/species has at least one target site, negative classification (0) if no target sites have been identified (NP), inadequate studies have been excluded.

$Salmonella$ $Mutagenicity$ CPDB mutagenicity classifications (pos/neg) were used without further modifications.

Validation

Leave-one-out crossvalidation

Leave-one-out (LOO) crossvalidation was used for all experiments. This means that all compounds from the training set are sequentially used as a query structure to determine the concordance between the prediction and the database activity. To enable an unbiased performance estimate the query compound (and all identical structures) are completely removed from the trainingset before its prediction is calculated. This implies of course that feature significances have to be reevaluated for each query structure. After a prediction has been obtained, the query structure and all identical structures are returned to the training set. The

process is repeated, until all compounds from the training set have served as query structures once [19]. For all validation experiments *sensitivities*, *specificities*, *positive/negative predictivity* and *predictive accuracies* are summarised in Tables 1 - 11.

Validation with an external testset

As almost all public carcinogenicity data of sufficient quality has been integrated into the CPDB it is at present impossible to find an external testset of sufficient size and quality to assess carcinogenicity predictions. Fortunately the situation has improved recently for *Salmonella* mutagenicity as a new dataset with 4337 compounds [20] was published in 2005. 3895 structures from this dataset have no mutagenicity information in the CPDB and were therefore used as an external testset (Kazius/Bursi testset). The results of this external validation experiment are summarised in Table 12.

Results

Leave-one-out crossvalidation

Tables 1 - 11 summarise the results of LOO validation. The first column contains the results that have been obtained without a consideration of the applicability domain (i.e. all predictions are accepted). The predictive accuracies can vary between 67% (Rat Carcinogenicity) and 86% (Male Hamster Carcinogenicity).

Unknown fragments (i.e. fragments that occur in the query structure, but not in the training set) have been identified in a substantial number of predictions (e.g. 43% of Rodent Carcinogenicity predictions, 50% of *Salmonella* Mutagenicity predictions). As no information about these fragments is available from the training set, it is up to the expert user to determine their toxicological relevance. The accuracy of predictions with/without unknown fragments is 67/72% for Rodent Carcinogenicity and 76/80% for *Salmonella* mutagenicity. This indicates that a subset of the unknown fragments has indeed toxicological relevance and can be responsible for misclassifications.

Another reason for misclassifications is harder to detect: These are structures that are too dissimilar to the training structures to make reliable predictions, although they share all their fragments with the training

set (i.e. they fall beyond the applicability domain of the training set). As the `lazar` confidence index incorporates the distances to similar training set structures (neighbours) as well as contradictory results within the training set it can be used as an indicator of the training sets applicability domain.

The second column in Tables 1 - 11 contains the results for structures within the applicability domain of the training set. For the purpose of this investigation a cutoff value of 0.05 was selected for the confidence index[7]. Predictions with a confidence index below this threshold were not accepted, because they fall beyond the applicability domain of the training set. As expected, the predictive accuracies rise to 78% (Female Rat Carcinogenicity) - 95% (Hamster Carcinogenicity) and the majority of crossvalidation results is better than 85%.

TABLES 1 - 11

Figures 5 and 6 provide a more detailed picture of the relationship between predictive accuracy and prediction confidence. Predictions are sorted according to their (absolute) confidence and cumulative prediction accuracies are plotted against the confidence index in a procedure similarly to lift charts [21]. These figures indicate also a good correlation between predictive accuracies and the `lazar` confidence index [8].

FIGURES 6, 5

Validation with an external testset

Table 12 summarises the results of *Salmonella* mutagenicity predictions for the external testset. The accuracy of predictions without consideration of the applicability domain is considerably lower (69%) than the LOO estimate for the same endpoint (78%). The results for structures within the applicability domain of the training set is however much more homogeneous (external validation: 85%, LOO: 87%) and show no signs of statistical significance as determined with the chi-square test (chi-square = 0.2647, p-value = 0.61). A plot of confidence indices vs. cumulative accuracies shows again a good correlation between both values (Figure 7).

TABLE 12

[7] This value is adjustable to account for variable application scenarios.

[8] Plots for the remaining endpoints show a similar shape. The high variability at the left hand side of the charts is the consequence of small sample sizes.

FIGURE 7

Discussion

Performance of the `lazar` algorithm

Similarity and neighbourhood based techniques have a long and successful history in Chemoinformatics [12]. Despite their conceptual simplicity they are frequently capable of outperforming much more complex QSAR techniques. Similarity based predictions are also appealing from a technical point of view, as in contrast to other QSAR methods (e.g. regression and projection based techniques) very few model assumptions are required. The rationale behind these techniques is in addition very close to the reasoning of human experts about toxicity, who also argue frequently with compounds that belong to the same chemical class and act by similar mechanisms. We assume therefore that it is relatively easy for a trained toxicologist to interpret and evaluate the results of similarity based predictions, e.g. by inspecting the proposed neighbours and searching for additional information about these compounds, if necessary.

Existing similarity based techniques can consider activity-specific similarities only by using predefined libraries of structural alerts for toxicity, but their definition and formal representation is laborious and error-prone. `lazar` overcomes this limitation by determining relevant fragments and activity related similarities automatically from experimental data.

The LOO results in Tables 1 - 11 indicate that `lazar` is capable of predicting a variety of carcinogenicity endpoints and to identify structures that fall beyond the applicability domain of the training set in a reliable manner. If the applicability domain is considered predictive accuracies can exceed 85% for almost all carcinogenicity endpoints.

An analysis of predictions for the external test set [20] substantiates the importance of considering the applicability domain (Table 12). At a first glance the accuracy for external predictions (69%) is substantially lower than the LOO results (78%, Table 2). This seems to support a common conception in the (Q)SAR community that LOO (and crossvalidation schemes in general) gives overly optimistic results. It is however very likely that an external test set of sufficient size contains a rather large fraction of poorly

predictable compounds that fall beyond the applicability domain of the training set. Table 12 provides evidence that this is indeed the case for the Kazius/Bursi validation set. If structures with a confidence < 0.05 are not accepted as reliable predictions, the predictive accuracy reaches 85% for external predictions. LOO with consideration of the applicability domain leads to almost the same value (87%, Table 2). Both results show no statistically significant differences (chi-square = 0.2647, p-value = 0.61).

This result is not only an indication of the good performance of the `lazar` algorithm, but also another indication that LOO provides indeed a reliable estimate of (Q)SAR predictions [19], if

- *all* information from the test structure has been removed from the test set[9] and
- only predictions that fall within the applicability domain of the training set are accepted.

It is presently impossible to perform a direct comparison of `lazar` with other carcinogenicity prediction techniques, as none of the other techniques were evaluated with the CPDB, and it seems that the size and composition of training and test sets has a major impact on the validation results [9]. The author has however used various combinations of MOLFEA derived linear fragments in conjunction with *Support Vector Machines (SVM)* to predict *Salmonella* mutagenicity [22] for an old version of the CPDB. The best results of this investigation are comparable (predictive accuracy: 0.785) to the `lazar` predictions without consideration of the applicability domain (predictive accuracy: 0.782) in terms of accuracy, but the impact of various MOLFEA and SVM parameters on predictive accuracy did not show a consistent trend. It is also likely that the crossvalidation results of the former investigation are too optimistic because class sensitive feature selection was performed prior to crossvalidation. Bayesian prediction techniques as they were implemented in previous `lazar` versions [10, 11] perform similarly, but they make heavy use of *a priori* probabilities, which leads to poor results on test sets with different fractions of actives and inactives (unpublished results).

Generally, the prediction of carcinogenic activity from chemical structures alone is known as a hard problem and many predictions fail to exceed the default probabilities [7–9, 23, 24]. Benigni and Zito [8]

[9] If feature selection and/or parameter optimisations are performed these steps have to be recalculated after the test structure has been removed from the training set. This may cause methodological problems for expert derived *structural alerts*, because he/she cannot forget information that has been derived from the test structure.

consider 65% as a reasonable upper limit of current technologies for rodent carcinogenicity predictions. Compared to these numbers, the performance of `lazar` is indeed an improvement.

Inspection of misclassifications

Despite the favourable `lazar` validation results it is obvious from Figures 5 - 7 that there are still structures that are misclassified despite high confidence indices. The purpose of this section is to discuss the most problematic misclassifications of the LOO results for the endpoints Rodent Carcinogenicity and *Salmonella* Mutagenicity as well as for the external validation exercise (Table 13).

Rodent Carcinogenicity (LOO)

The most problematic misclassifications of LOO crossvalidation are most likely the result of inconsistencies in the database. Quercetin (CAS 117-39-5) and sodium saccharin (CAS 128-44-9) e.g. are labelled as carcinogens in the database, although other compounds with the same parent structure (quercetin dihydrate, saccharin and calcium saccharin) are inactive. In these cases the "inactive" `lazar` prediction is probably correct. Retinol acetate (CAS 127-47-9, Vitamin A, prediction: inactive) has been found to induce tumors in the adrenal gland of rats at high doses in a single study, but it is likely not a carcinogen at physiological concentrations.

Salmonella Mutagenicity (LOO)

The Chemical Carcinogenesis Research Information System (CCRIS `http://toxnet.nlm.nih.gov/`) lists several positive results in a variety of *Salmonella* strains for Chlorodibromomethane (CAS 124-48-1, `lazar` classification: active), although the CPDB classification is inactive. The misclassifications of 2-Mercaptobenzothiazole (CAS 149-30-4) and its dimer Benzothiazyl disulfide (CAS 120-78-5) for *Salmonella* mutagenicity depend on each other, because both compounds are classified differently in the CPDB, despite their structural similarity. The classification of Benzothiazyl disulfide as *Salmonella* mutagen, is however very questionable, because a recent evaluation (`http://www.epa.gov/chemrtk/bnzthict/c13324tc.htm`) reports 9 negative and a single positive result. If we assume a negative

category for Benzothiazyl disulfide, 2-Mercaptobenzothiazole will also be correctly classified as negative, because it is its closest neighbour.

Salmonella Mutagenicity (external testset)

The most problematic misclassification from the external testset is Azauraxil (CAS 461-89-2, prediction: inactive). CCRIS lists no positive *Salmonella* mutagenicity findings for Azauraxil, but the structurally identical IPO 3834 was active in strains TA1538 and TA98 with metabolic activation by rat liver S9 (but not with mouse liver S9). The remaining 17 assays were negative. Diazoxon (CAS 962-58-0, prediction: inactive) was active in TA100 without metabolic activation, but inactive in TA100 with and in TA98 with and without metabolic activation. The structurally similar Diazinon (CAS 333-41-5) was negative in all assays. According to the CCRIS Benzo(b)chrysene (CAS 214-17-5, prediction: active) was tested only in a single *Salmonella* strain (TA100) and structurally related PAHs like Dibenz(a,h)anthracene and Benzo(a)pyrene are well known mutagens.

This brief discussion clearly indicates that many of the most problematic misclassifications can be attributed to inconsistencies in the database. The experimental findings for many of these compounds are frequently limited and sometimes contradictory.

There is still the possibility for systematic errors, because linear fragments cannot account for all structural differences. If e.g. the nonmutagenicity of Benzo(b)chrysene is experimentally confirmed, there might be a problem to differentiate it from other PAHs like Dibenz(a,h)anthracene and Benzo(a)pyrene. In this case it will be necessary to substitute linear fragments with a richer fragment language (e.g. subgraphs), but the main classification algorithm can remain the same.

Conclusions

`lazar` is a new tool for the prediction of toxic properties of chemical structures. It derives predictions for query structures from a database with experimentally determined toxicity data. For this purpose, `lazar` searches the database for compounds that are similar *with respect to a given toxic activity* and calculates the prediction from their activities. `lazar` is able to determine wheter a query structure falls within the

applicability domain of the training set, by assigning a confidence index to each prediction.

Leave-one-out crossvalidation and validation with an external testset of almost 4000 compounds, indicate that `lazar` is capable of achieving predictive accuracies of more than 85% for most of the investigated carcinogenicity and mutagenicity endpoints and that it is capable of discriminating reliably between trustworthy and not trustworthy predictions. It is interesting to note that the crossvalidation and external validation results are in good agreement for structures within the applicability domain of the training set.

As high prediction accuracies are achievable for compounds within the applicability domain of the test set, it may be justified to conclude that the poor performance of previous attempts to predict rodent carcinogenicity is not primarily the consequence of poor prediction techniques, complex biological mechanisms [25] or unreliable data [7, 26], but rather the consequence of an insufficient coverage of the chemical space in the training sets. This hypothesis is in accordance with Benigni and Giulianis [27] observation that it is in fact possible to obtain reliable carcinogenicity predictions for certain types of (congeneric) compounds (e.g. aromatic amines and nitroaromatics).

A web interface for the *Carcinogenic Potency Database (CPDB)* can be accessed at `http://www.predictive-toxicology.org/lazar`. The source code for the command line version of the complete program can be obtained on request from the author.

Acknowledgements

This work was funded by a grant of the Centre for Documentation and Evaluation of Alternatives to Animal Experiments (ZEBET). The author wants to thank V. Horal-Gurfinkel and A. Maunz for the `lazar` web interface and A. Karwath, A. Benigni and L. DeRaedt for helpful discussions.

References

[1] Helma, C. (Ed.) Predictive Toxicology. Taylor & Francis, Boca Raton (2005).

[2] Eriksson, L., Johansson, E. and Lundstedt, T. *Regression- and projection-based approaches in Predictive Toxicology*. In Helma, C. (Ed.) Predictive Toxicology. Taylor & Francis, Boca Raton (2005) pp. 177–222.

[3] Parsons, S. and McBurney, P. *The use of expert systems for toxicology risk prediction*. In Helma, C. (Ed.) Predictive Toxicology. Taylor & Francis, Boca Raton (2005) pp. 135–176.

[4] Kramer, S. and Helma, C. *Machine learning and data mining*. In Helma, C. (Ed.) Predictive Toxicology. Taylor & Francis, Boca Raton (2005) pp. 223–254.

[5] Imielinski, T. and Mannila, H. *A database perspective on knowledge discovery*. Communications of the ACM, 39 (1996) 58–64.

[6] DeRaedt, L. *A perspective on inductive databases*. SIGKDD Explorations, 4 (2002) 69–77.

[7] Toivonen, H., Srinivasan, A., King, R. D., Kramer, S. and Helma, C. *Statistical evaluation of the Predictive Toxicology Challenge 2000–2001*. Bioinformatics, 19 (2003) 1183–1193.

[8] Benigni, R. and Zito, R. *The second National Toxicology Program comparative exercise on the prediction of rodent carcinogenicity: Definitive results*. Mutation Res., 566 (2004) 49–63.

[9] Benigni, R. *Structure–activity relationship studies of chemical mutagens and carcinogens: Mechanistic investigations and prediction approaches*. Chemical Reviews, in press (2005).

[10] Helma, C. *Data mining and knowledge discovery in predictive toxicology*. SAR QSAR Environ. Res., 15 (2004) 367–383.

[11] Helma, C. `lazar`: *Lazy Structure – Activity Relationships for toxicity prediction*. In Helma, C. (Ed.) Predictive Toxicology. Taylor & Francis, Boca Raton (2005) pp. 479–499.

[12] Willett, P., Barnard, J. and Downs, G. *Chemical similarity searching*. J. Chem. Inf. Comput. Sci., 38 (1998) 983–996.

[13] Kramer, S., De Raedt, L. and 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.

[14] Hill, A. *Erweiterung des Molecular Feature Miners für 3-dimensionale Fragmente*. Master's thesis, Universität Freiburg (2002).

[15] Molzberger, L. *Development of a method to search efficiently for frequent substructures in large molecule databases*. Master's thesis, Universität Freiburg (2004).

[16] Poroikov, V. and Filimonov, D. *Pass: Prediction of biological activity for substances*. In Helma, C. (Ed.) Predictive Toxicology. Taylor & Francis, Boca Raton (2005) pp. 459–478.

[17] Varnek, A. and Solov'ev, V. *"in silico" design of potential anti-HIV actives using fragment descriptors*. Comb. Chem. High. Throughput Screen., 8 (2005) 403–416.

[18] Coles, S., Day, N., Murray-Rust, P., Rzepa, H. and Zhang, Y. *Enhancement of the chemical semantic web through the use of InChI identifiers*. Org. Biomol. Chem., 3 (2005) 1832–1834.

[19] Hawkins, D. *The problem of overfitting*. J. Chem. Inf. Comput. Sci., 44 (2004) 1–12.

[20] Kazius, J., McGuire, R. and Bursi, R. *Derivation and vlaidation of toxicophores for mutagenicity prediction*. J. Med. Chem., 48 (2005) 312–320.

[21] Witten, I. and Frank, E. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, San Francisco, California (2000).

[22] Helma, C., Kramer, T., Kramer, S. and DeRaedt, 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., 44 (2004) 1402–1411.

[23] Benigni, R. *Qsar prediction of rodent carcinogenicity for a set of chemicals currently bioassayed by the us national toxicology program*. Mutagenesis, 6 (1991) 423–425.

[24] Benigni, R. *Predicting chemical carcinogenesis in rodents: The state of art in light of a comparative exercise*. Mutation Res., 334 (1995) 103–113.

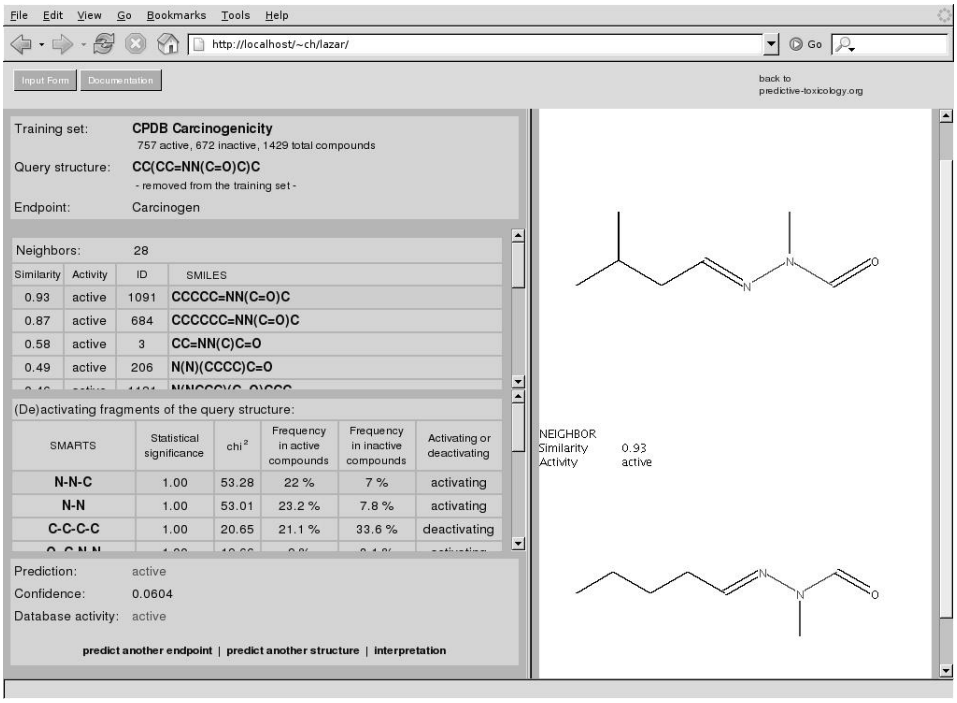
[25] Woo, Y. and Lai, D. Y. *Mechanism of action of chemical carcinogens and their role in Structure-Activity Relationship (SAR) analysis and risk assessment*. In Benigni, R. (Ed.) Quantitative Structure–Activity Relationship (QSAR) Models of Mutagens and Carcinogens. CRC Press, Boca Raton (2003) pp. 41–80.

[26] Gottmann, E., Kramer, S., Pfahringer, B. and Helma, C. *Data quality in predictive toxicology: Reproducibility of rodent carcinogenicity experiments*. Environ. Health Perspect., 109 (2001) 509–514.

[27] Benigni, R. and Giuliani, A. *Putting the Predictive Toxicology Challenge into prespective: Reflections on the results*. Bioinformatics, 19 (2003) 1194–1200.

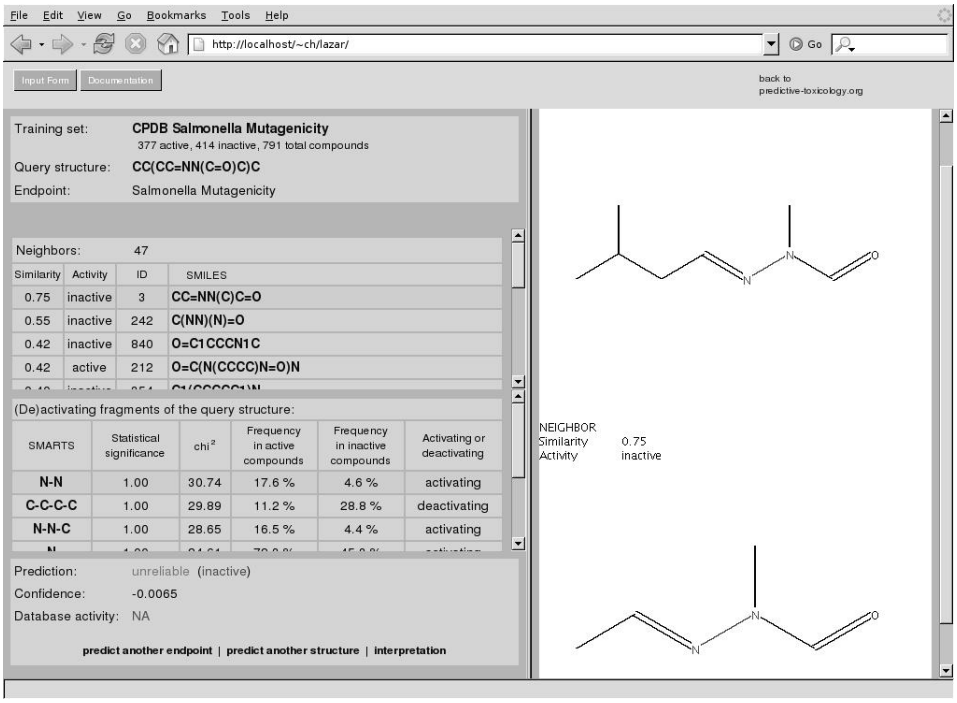
Figures

Figure 1: `lazar` screenshot of the prediction of rodent carcinogenicity for 3-Methylbutanal methylformylhydrazone



The query structure and the first neighbour are depicted in the right frame. Fragments can be highlighted in both structures. Note the difference between neighbours and fragments for *Salmonella* mutagenicity (Figure 2), this is the result of activity specific similarities.

Figure 2: `lazar` screenshot of the prediction of *Salmonella* mutagenicity for 3-Methylbutanal methylformylhydrazone



This prediction is unreliable, because the query structure falls beyond the applicability domain of the training set ($Confidence < 0.05$). The query structure and the first neighbour are depicted in the right frame. Fragments can be highlighted in red in both structures. Note the difference between neighbours and fragments for rodent carcinogenicity (Figure 1), this is the result of activity specific similarities.

Figure 3: The main `lazar` algorithm for classification and regression

Determination of Neighbours

Require: Query structure s_q, training database $D = \{(s_1, t_1), (s_2, t_2), \ldots, (s_n, t_n)\}$ with training structures s_i and training activities, t_i
Neighbours $N = \{\}$
for all $s_i \in D$ **do**
$sim_i = similarity(s_q, s_i, D)$
if $sim_i > 0.3$ **then**
$N = N \cup (s_i, t_i, sim_i)$
end if
end for

Classification

Require: Neighbours N

$$conf = \frac{\sum_{n \in N} \{sim_n | t_n = "active"\}^4 - \sum_{n \in N} \{sim_n | t_n = "inactive"\}^4}{|N|}$$

if $conf > 0$ **then**
$t_q = "active"$
else if $conf < 0$ **then**
$t_q = "inactive"$
end if

Figure 4: Determination of activity specific chemical similarities

Require: Query structure s_q, training structure s_t, training database D, fragment language and/or predefined fragments L
  Significant fragments $F = \{\}$
  for all $\{f \in L | f \subseteq s_q \vee f \subseteq s_t\}$ **do**
    $p_f = significance(f, D)$ {determined e.g. by the χ^2- or sign-test.}
    if $p_f \geq 0.95$ **then**
      $F = F \cup f$
    end if
  end for

$$sim(s_q, s_t, D) = \frac{\sum_{f \in F} \{p_f^4 | f \subseteq s_q \wedge f \subseteq s_t\}}{\sum_{f \in F} \{p_f^4 | f \subseteq s_q \vee f \subseteq s_t\}}$$

Figure 5: Confidence vs. predictive accuracy for rodent carcinogenicity

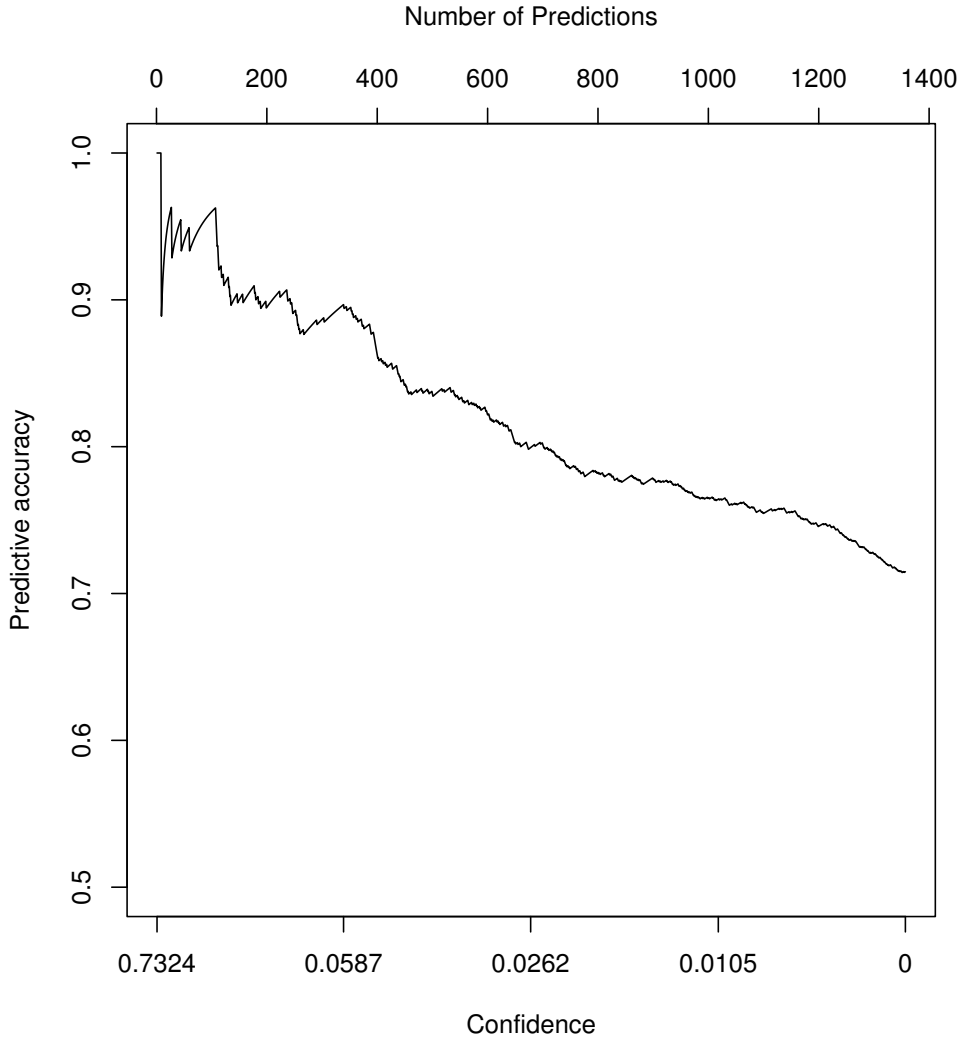


Figure 6: Confidence vs. predictive accuracy for *Salmonella* mutagenicity

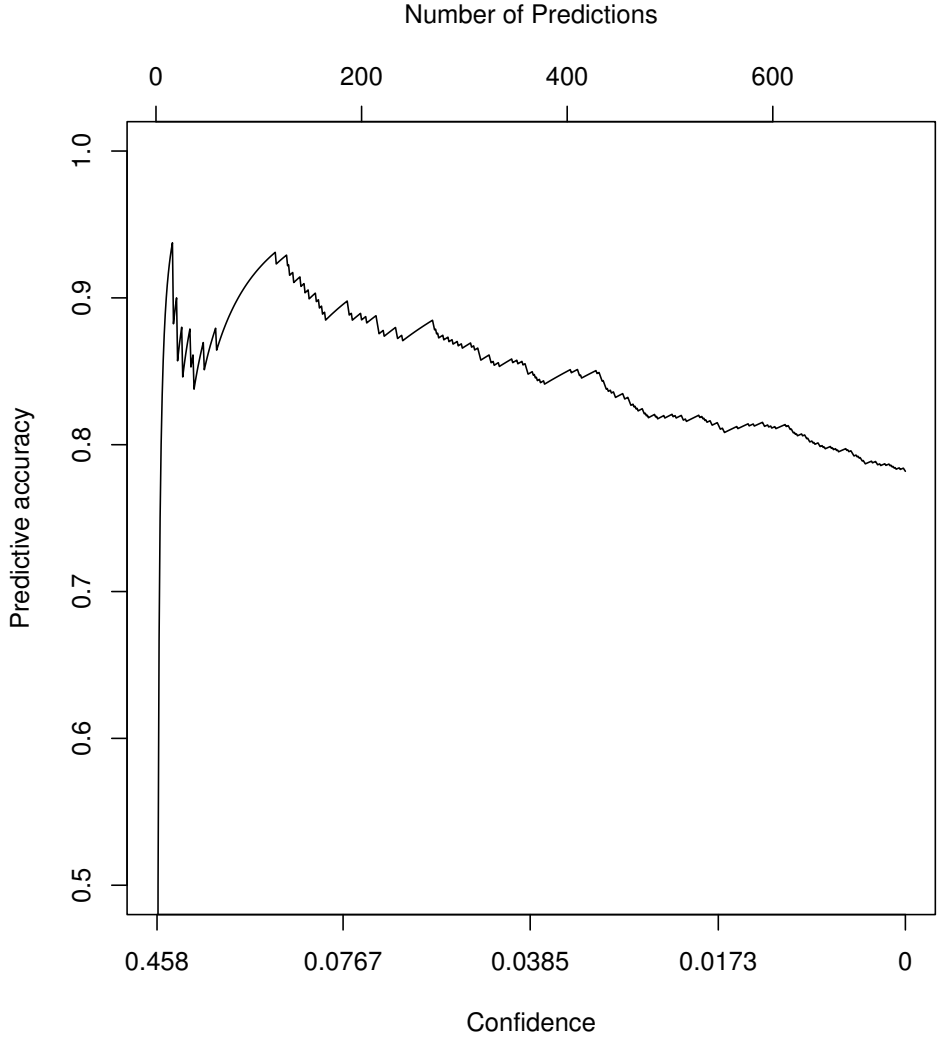
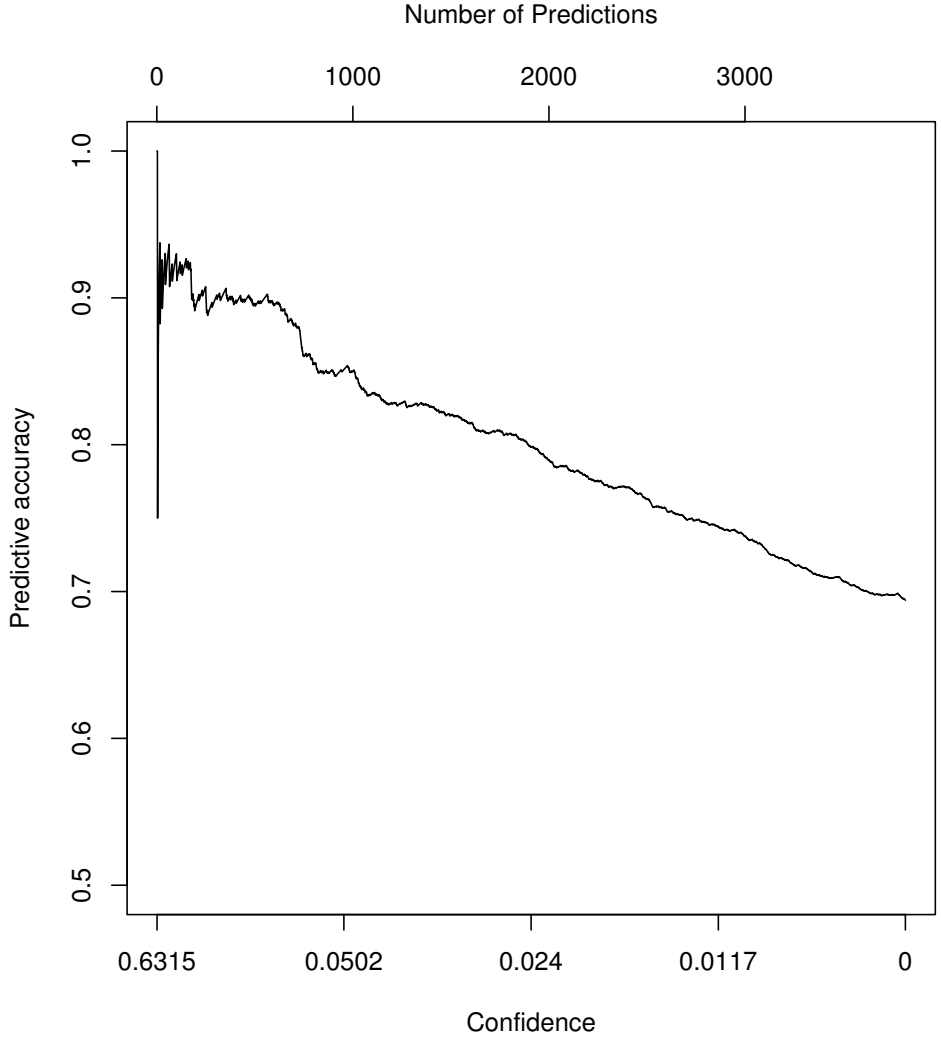


Figure 7: Confidence vs. predictive accuracy for external predictions of *Salmonella* mutagenicity



Tables

Table 1: Leave-one-out crossvalidation of rodent carcinogenicity predictions.

Confidence		$\geq 0^a$	$> 0.05^b$
True positive predictions	tp	513	209
True negative predictions	tn	457	139
False positive predictions	fp	190	26
False negative predictions	fn	197	31
True positive rate (Sensitivity)	$tp/(tp+fn)$	0.7225	0.8708
True negative rate (Specificity)	$tn/(tn+fp)$	0.7063	0.8424
Positive predictivity	$tp/(tp+fp)$	0.7297	0.8894
Negative predictivity	$tn/(tn+fn)$	0.6988	0.8176
Accuracy (Concordance)	$(tp+tn)/(tp+tn+fp+fn)$	0.7148	0.8593

[a] Without consideration of the applicability domain.
[b] Predictions within applicability domain.

Table 2: Leave-one-out crossvalidation of *Salmonella* mutagenicity predictions.

Confidence		$\geq 0^a$	$> 0.05^b$
True positive predictions	tp	283	147
True negative predictions	tn	287	109
False positive predictions	fp	85	20
False negative predictions	fn	74	19
True positive rate (Sensitivity)	$tp/(tp+fn)$	0.7927	0.8855
True negative rate (Specificity)	$tn/(tn+fp)$	0.7715	0.8450
Positive predictivity	$tp/(tp+fp)$	0.7690	0.8802
Negative predictivity	$tn/(tn+fn)$	0.7950	0.8516
Accuracy (Concordance)	$(tp+tn)/(tp+tn+fp+fn)$	0.7819	0.8678

[a] Without consideration of the applicability domain.
[b] Predictions within applicability domain.

Table 3: Leave-one-out crossvalidation of hamster carcinogenicity predictions.

Confidence		$\geq 0^a$	$> 0.05^b$
True positive predictions	tp	30	26
True negative predictions	tn	22	11
False positive predictions	fp	4	1
False negative predictions	fn	7	1
True positive rate (Sensitivity)	$tp/(tp+fn)$	0.8108	0.9630
True negative rate (Specificity)	$tn/(tn+fp)$	0.8462	0.9167
Positive predictivity	$tp/(tp+fp)$	0.8824	0.9630
Negative predictivity	$tn/(tn+fn)$	0.7586	0.9167
Accuracy (Concordance)	$(tp+tn)/(tp+tn+fp+fn)$	0.8254	0.9487

[a] Without consideration of the applicability domain.
[b] Predictions within applicability domain.

Table 4: Leave-one-out crossvalidation of mouse carcinogenicity predictions.

Confidence		$\geq 0^a$	$> 0.05^b$
True positive predictions	tp	220	94
True negative predictions	tn	388	137
False positive predictions	fp	97	14
False negative predictions	fn	153	23
True positive rate (Sensitivity)	$tp/(tp+fn)$	0.5898	0.8034
True negative rate (Specificity)	$tn/(tn+fp)$	0.8000	0.9073
Positive predictivity	$tp/(tp+fp)$	0.6940	0.8704
Negative predictivity	$tn/(tn+fn)$	0.7172	0.8562
Accuracy (Concordance)	$(tp+tn)/(tp+tn+fp+fn)$	0.7086	0.8619

[a] Without consideration of the applicability domain.
[b] Predictions within applicability domain.

Table 5: Leave-one-out crossvalidation of rat carcinogenicity predictions.

Confidence		$\geq 0^a$	$> 0.05^b$
True positive predictions	tp	317	164
True negative predictions	tn	412	106
False positive predictions	fp	149	18
False negative predictions	fn	212	24
True positive rate (Sensitivity)	$tp/(tp+fn)$	0.5992	0.8723
True negative rate (Specificity)	$tn/(tn+fp)$	0.7344	0.8548
Positive predictivity	$tp/(tp+fp)$	0.6803	0.9011
Negative predictivity	$tn/(tn+fn)$	0.6603	0.8154
Accuracy (Concordance)	$(tp+tn)/(tp+tn+fp+fn)$	0.6688	0.8654

[a] Without consideration of the applicability domain.
[b] Predictions within applicability domain.

Table 6: Leave-one-out crossvalidation of female hamster carcinogenicity predictions.

Confidence		$\geq 0^a$	$> 0.05^b$
True positive predictions	tp	21	18
True negative predictions	tn	18	16
False positive predictions	fp	6	3
False negative predictions	fn	4	1
True positive rate (Sensitivity)	$tp/(tp+fn)$	0.8400	0.9474
True negative rate (Specificity)	$tn/(tn+fp)$	0.7500	0.8421
Positive predictivity	$tp/(tp+fp)$	0.7778	0.8571
Negative predictivity	$tn/(tn+fn)$	0.8182	0.9412
Accuracy (Concordance)	$(tp+tn)/(tp+tn+fp+fn)$	0.7959	0.8947

[a] Without consideration of the applicability domain.
[b] Predictions within applicability domain.

Table 7: Leave-one-out crossvalidation of male hamster carcinogenicity predictions.

Confidence		$\geq 0^a$	$> 0.05^b$
True positive predictions	tp	24	20
True negative predictions	tn	20	13
False positive predictions	fp	3	2
False negative predictions	fn	4	1
True positive rate (Sensitivity)	$tp/(tp+fn)$	0.8571	0.9524
True negative rate (Specificity)	$tn/(tn+fp)$	0.8696	0.8667
Positive predictivity	$tp/(tp+fp)$	0.8889	0.9091
Negative predictivity	$tn/(tn+fn)$	0.8333	0.9286
Accuracy (Concordance)	$(tp+tn)/(tp+tn+fp+fn)$	0.8627	0.9167

[a] Without consideration of the applicability domain.
[b] Predictions within applicability domain.

Table 8: Leave-one-out crossvalidation of female mouse carcinogenicity predictions.

Confidence		$\geq 0^a$	$> 0.05^b$
True positive predictions	tp	158	83
True negative predictions	tn	451	157
False positive predictions	fp	57	13
False negative predictions	fn	150	21
True positive rate (Sensitivity)	$tp/(tp+fn)$	0.5130	0.7981
True negative rate (Specificity)	$tn/(tn+fp)$	0.8878	0.9235
Positive predictivity	$tp/(tp+fp)$	0.7349	0.8646
Negative predictivity	$tn/(tn+fn)$	0.7504	0.8820
Accuracy (Concordance)	$(tp+tn)/(tp+tn+fp+fn)$	0.7463	0.8759

[a] Without consideration of the applicability domain.
[b] Predictions within applicability domain.

Table 9: Leave-one-out crossvalidation of male mouse carcinogenicity predictions.

Confidence		$\geq 0^a$	$> 0.05^b$
True positive predictions	tp	135	71
True negative predictions	tn	431	153
False positive predictions	fp	60	16
False negative predictions	fn	145	30
True positive rate (Sensitivity)	$tp/(tp+fn)$	0.4821	0.7030
True negative rate (Specificity)	$tn/(tn+fp)$	0.8778	0.9053
Positive predictivity	$tp/(tp+fp)$	0.6923	0.8161
Negative predictivity	$tn/(tn+fn)$	0.7483	0.8361
Accuracy (Concordance)	$(tp+tn)/(tp+tn+fp+fn)$	0.7341	0.8296

[a] Without consideration of the applicability domain.
[b] Predictions within applicability domain.

Table 10: Leave-one-out crossvalidation of female rat carcinogenicity predictions.

Confidence		$\geq 0^a$	$> 0.05^b$
True positive predictions	tp	183	93
True negative predictions	tn	448	132
False positive predictions	fp	72	26
False negative predictions	fn	151	38
True positive rate (Sensitivity)	$tp/(tp+fn)$	0.5479	0.7099
True negative rate (Specificity)	$tn/(tn+fp)$	0.8615	0.8354
Positive predictivity	$tp/(tp+fp)$	0.7176	0.7815
Negative predictivity	$tn/(tn+fn)$	0.7479	0.7765
Accuracy (Concordance)	$(tp+tn)/(tp+tn+fp+fn)$	0.7389	0.7785

[a] Without consideration of the applicability domain.
[b] Predictions within applicability domain.

Table 11: Leave-one-out crossvalidation of male rat carcinogenicity predictions.

Confidence		$\geq 0^a$	$> 0.05^b$
True positive predictions	tp	203	123
True negative predictions	tn	440	132
False positive predictions	fp	86	21
False negative predictions	fn	166	28
True positive rate (Sensitivity)	$tp/(tp+fn)$	0.5501	0.8146
True negative rate (Specificity)	$tn/(tn+fp)$	0.8365	0.8627
Positive predictivity	$tp/(tp+fp)$	0.7024	0.8542
Negative predictivity	$tn/(tn+fn)$	0.7261	0.8250
Accuracy (Concordance)	$(tp+tn)/(tp+tn+fp+fn)$	0.7184	0.8388

[a] Without consideration of the applicability domain.
[b] Predictions within applicability domain.

Table 12: Validation of *Salmonella* mutagenicity predictions for an external testset

Confidence		$\geq 0^a$	$> 0.05^b$
True positive predictions	tp	1467	546
True negative predictions	tn	1183	282
False positive predictions	fp	492	103
False negative predictions	fn	676	39
True positive rate (Sensitivity)	$tp/(tp+fn)$	0.6846	0.9333
True negative rate (Specificity)	$tn/(tn+fp)$	0.7063	0.7325
Positive predictivity	$tp/(tp+fp)$	0.7489	0.8413
Negative predictivity	$tn/(tn+fn)$	0.6364	0.8785
Accuracy (Concordance)	$(tp+tn)/(tp+tn+fp+fn)$	0.6941	0.8536

[a] Without consideration of the applicability domain.
[b] Predictions within applicability domain.

Table 13: Misclassified instances with high prediction confidences

Compound Name	CAS	Confidence	Classification `lazar`	Classification CPDB	Remarks
Rodent Carcinogenicity (LOO)					
Quercetin	117-39-5	-0.6605	inactive	active	Quercetin dihydrate inactive
Sodium saccharin	128-44-9	-0.4938	inactive	active	Saccharin and Calcium saccharin inactive
Retinol acetate	127-47-9	-0.3734	inactive	active	Vitamin A probably not carcinogenic at physiological doses
Salmonella mutagenicity (LOO)					
Chloro-dibromo-methane	124-48-1	0.4580	active	inactive	Active in several *Salmonella* strains with and without metabolic activation (CCRIS)
2-Mercapto-benzothiazole	149-30-4	0.3445	active	inactive	Classification based on Benzothiazyl disulfide (see below)
Benzothiazyl disulfide	120-78-5	-0.3222	inactive	active	Classification based on 2-Mercaptobenzothiazole (see above)
Salmonella mutagenicity (Kazius/Bursi testset)					
Azaurazil	461-89-2	-0.5479	inactive	active	Azaurazil negative, IPO 3834 inactive in 17 from 19 *Salmonella* assays (CCRIS)
Diazoxon	962-58-3	-0.4249	inactive	active	Active in TA100 without metabolic activation, TA100 with and TA98 with/without metabolic activation negative, Diazinon negative
Benzo(b)-chrysene	214-17-5	0.4061	active	inactive	Tested only in a single strain (TA100), related PAHs like Dibenz(a,h)anthracene and Benzo(a)pyrene are active

CCRIS . . . Chemical Carcinogenesis Research Information System `http://toxnet.nlm.nih.gov/`