

In silico Predictive Toxicology: The State of the Art
and Strategies to Predict Human Health Effects

Christoph Helma

Inst. f. Computer Science

Univ. Freiburg

October 19, 2004

Abstract

In silico Predictive Toxicology techniques are a fast and cost efficient alternative (or supplement) to bioassays for the identification of toxic effects at an early stage of product development. This review provides a conceptual description of the most important *in silico* prediction techniques and presents exemplary strategies for the prediction of human health effects. Special emphasis will be given to validation issues and the performance of models for human health related effects.

Keywords Predictive Toxicology, Expert Systems, (Q)SAR, Data Mining, Validation, Human Health Effects

1 Introduction

Unacceptable toxicity is still a major bottleneck in the drug discovery process and it was recently estimated that the attrition rate due to toxicity rose from $\approx 10\%$ to more than 20% during the last ten years [1]. There is therefore an urgent need for techniques, that are capable to identify adverse effects at a very early stage of product development and provide reasonable toxicity estimates for the huge number of untested compounds. Computer based (*in silico*) techniques are particularly appealing for this purpose, because they are extremely fast and cost efficient and can be applied even without a physically available compound.

This review will focus on a conceptual description of the most important *in silico* prediction and validation techniques and present exemplary strategies for the prediction

of human health effects. As the author works predominantly on the development of Data Mining techniques for the prediction of toxicity, there will be certainly a bias towards these techniques within this review.

Readers who are interested in a more thorough introduction to *in silico* Predictive Toxicology can consult three recent textbooks with complementary content: “Predictive Toxicology” [2] focuses on Predictive Toxicology techniques with an emphasis on Data Mining and their implementation in commercial and non-commercial programs, “Predicting Chemical Toxicity and Fate” [3] is geared more towards statistical QSAR techniques and their application for human and environmental endpoints and “(QSAR) Models of Mutagenicity and Carcinogenicity” [4] focuses on the endpoints Mutagenicity and Carcinogenicity.

2 *in silico* Toxicity Prediction Techniques

In silico Toxicity Prediction Techniques may be coarsely classified into methods that model biochemical events that are relevant for toxicity (Molecular Modeling), techniques that mimic human reasoning about toxicological phenomena (Expert Systems) and methods that derive predictions from a training set of experimentally determined data (Data Driven Systems).

2.1 Molecular Modeling

Molecular Modeling techniques assess the interaction of small molecules with biological macromolecules (predominately proteins), by fitting the ligand into the active site of the receptor. Molecular Modeling has been used mainly in pharmaceutical research for the detection and evaluation of new lead compounds. But the same techniques can also be applied for toxicological purposes, if the mechanism is receptor mediated (e.g. cytochrome P450s, estrogen receptor) and the receptor structure is available. Kroemer [5] gives a brief review of Molecular Modeling techniques, software for Molecular Modeling can be found at the Protein Data Bank (PDB) website <http://www.rcsb.org/pdb/software-list.html>.

Molecular Modeling can be used to elucidate mechanisms and biotransformations and to predict receptor-mediated toxicity (e.g. estrogenicity), but the prediction of toxicities with complex and partially unknown mechanisms is beyond their scope.

2.2 Expert Systems

Expert Systems attempt to formalize the knowledge of human experts, who assess the toxicity of a new compound, in a computer program. This approach is intuitively appealing to most users, because it promises easy access to toxicological knowledge, and many of the most successful Predictive Toxicology software tools are in fact Expert Systems (Tables 1 and 2). A concise description of the Expert System approach in toxicology can be found in a review from Parsons and McBurney [6], techniques for the prediction of biotransformations and metabolites have been reviewed by Payne [7].

Program	URL	Reference
HazardExpert	http://www.compudrug.com/	—
Oncologic	http://www.epa.gov/opptintr/cahp/actlocal/can.html	[8]
DEREK	http://www.chem.leeds.ac.uk/luk/derek/index.html	[9]

Table 1: Exemplary Expert Systems for Toxicity Predictions

Program	URL	Reference
MetabolExpert	http://www.compudrug.com/	—
META	http://www.multicase.com/products/prod05.htm	[10]
METEOR	http://www.chem.leeds.ac.uk/luk/meteor/index.html	[11]
TIMES	http://omega.btu.bg/software.php	[12]

Table 2: Exemplary Expert Systems for Metabolite Predictions

The creation of a knowledge base for an Expert System requires extensive literature searches and the developers capability to create general applicable knowledge from specific cases. This enables predictions, even if very few experimental measurements are available (e.g. human health endpoints and biotransformation). The flexibility of the human mind allows also an integration of very diverse chemical and biological information. (e.g. substructures, measured and calculated molecular properties, *in vivo* and *in vitro* measurements).

An inaccurate knowledge base (frequently inappropriate generalizations from a few examples) can be on the other hand a major pitfall of Expert Systems. It is therefore necessary to update the knowledge base regularly and to integrate new scientific evidence as well as feedback from the users.

The formal definition of applicability domains and proper validation (with complete exclusion of all test set information, see validation section below) of Expert Systems can

be problematic, because it is frequently unclear from which compounds the information in the knowledge base originates.

2.3 Data Driven Systems

Data Driven Systems are formalized methods for the extraction of prediction models directly from experimental data. (Quantitative)Structure-Activity Relationships (Q)SARs are typical examples for such a procedure. Classical (Q)SAR analysis uses regression techniques to derive equations from experimental data [13]. These equations can be used for the prediction of further compounds with similar structures (i.e. *congeneric* compounds) and mechanisms. Predictions can be based on quantitative molecular properties (QSAR) as well as on the presence or absence of toxicity inducing substructures (SAR).

As toxicity experiments are frequently too expensive and time consuming to be performed especially for the development of (Q)SAR models, data mining techniques have gained much popularity during the last years. Well known data mining techniques for this purpose are k-nearest neighbors, Bayesian techniques, recursive partitioning (decision and regression trees), rule induction, neural nets and support vector machines. A description of data mining techniques and their applicability in Predictive Toxicology can be found in [15].

An important task for Data Driven systems is the selection of chemical features that are relevant for the toxic effect under investigation. With some toxicological knowledge it is relatively easy to come up with an almost unlimited number of chemical features,

Program	URL	Reference
TOPKAT	http://www.accelrys.com/products/topkat/	[17]
MCASE	http://www.multicase.com/products/prod01.htm	[10]
PASS	http://www.ibmh.msk.su/PASS/index.html	[18]
lazar	http://www.predictive-toxicology.org/lazar/	[19]

Table 3: Exemplary Data Driven Systems for Toxicity Predictions

that might be relevant for toxicity. But it is hard to determine *a priori*, which of these features are really relevant for toxicity and to guarantee, that no important features are missing. Klopman was the first who tried to solve this problem with the CASE and MCASE systems [14], by generating *all* possible linear substructures up to a certain length and evaluating their relevance automatically with statistical procedures. Such a procedure allows the automated detection of toxicologically relevant substructures, but it requires efficient algorithms for feature generation and feature selection [15, 16].

Examples for data driven programs that are capable to predict toxic effects are summarized in Table 3. It is of course also possible to use one of the countless (Q)SAR equations from the literature or to generate a new prediction model with a statistical or data mining package¹.

2.3.1 Toxicity Databases

With Data Driven systems it is in principle to model every conceivable toxic endpoint, if a training set of sufficient size, structural diversity and quality exists. A major bottleneck in the development of (Q)SAR models is still the limited public availability of high-quality

¹R (<http://cran.r-project.org/>) and WEKA (<http://www.cs.waikato.ac.nz/~ml/weka/>) are excellent open-source programs for this purpose.

toxicity data [20]. Two major initiatives have been started to standardize and increase the availability of toxicity data: The Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network (<http://www.epa.gov/nheer1/dsstox/index.html>, [21]) is an initiative of Ann Richard (US EPA) to provide a decentralized set of standardized structure-searchable toxicity databases. The VITIC database www.chem.leeds.ac.uk/luk/vitic/ is also a structure-searchable database that is sponsored by a number of pharmaceutical and chemical companies and coordinated by LHASA Inc. It is planned to release the public part of the VITIC data within the DSSTox project.

2.4 Combination of Predictions

Each Predictive Toxicology technique has its own distinct advantages and weaknesses, an obvious approach to get the best out of several worlds, is to combine predictions from different models. Such an exercise has been reported frequently in the literature and the general experience is, that prediction accuracies can be significantly enhanced in most cases. A similar and equally beneficial strategy is the integration of bioassay results with *in silico* predictions.

2.5 Regulatory Acceptance of *in silico* Predictions

In order to regulate the use of chemicals and pharmaceuticals governmental authorities require information about adverse effects on human and environmental health. Traditionally this information comes predominately 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 methods. For industrial chemicals the Commission of the European Communities proposes to address these concerns by increasing the use of *in silico* and *in vitro* techniques [22] and many regulatory authorities (e.g. Danish EPA, German Federal Institute for Risk Assessment, US EPA, US Food and Drug Administration) use already (Q)SAR models to support their decisions [23]. If *in silico* predictions are used in a regulatory context, it is obvious that they have to meet rigorous quality standards. These standards have been drafted in the *OECD Principles for (Q)SAR Validation* (formerly Setubal Principles) which are currently under revision.

3 Validation of *in silico* Models

3.1 Validation Techniques

A statistically sound and unbiased estimation of the capabilities of Predictive Toxicology systems is crucial for the comparison of different techniques, the proper interpretation of their results and for an application in a regulatory context. As many (Q)SAR studies still contain major methodological flaws in regard to validation, it is important to understand the basic principles which are relatively straightforward: A prediction model is derived from a set of training compounds. This model is used to make predictions for a set of test compounds which are compared with their measured values. For an unbiased estimation of prediction capabilities it is crucial to exclude all information from the test set from model

development [24].

Validation on the training set (or an incomplete exclusion of test set instances) is still a frequent flaw in (Q)SAR validation studies. [24] demonstrates the consequences: It is possible to use algorithms (e.g. 1-nearest neighbor) that give *per definitionem* 100% accurate predictions, for test instances that can be found in the training set. This value is of course not indicative of the performance for unseen compounds. The following paragraphs will present a brief overview of the most popular validation techniques, an excellent treatment of validation techniques for (Q)SAR problems can be found in [24].

3.1.1 Validation with External Test Sets

A clear separation between the training set for model development and a test set for validation (ideally untested at the time of prediction), is clearly very desirable. Such an approach has e.g. been used for the NTP prediction exercises [25] and the Predictive Toxicology Challenge [26]. The main problem with external test sets is a practical one: To obtain statistically sound results a test set of sufficient size and coverage of the “chemical universe” is required. Results that have been obtained with small test sets are likely to be due to chance. Benigni [27] has demonstrated this empirically for several prediction models and test sets: Each model could span almost the entire performance space and the predictive accuracy depended heavily on the composition of the test set.

3.1.2 Validation with Artificial Test Sets

If it is impossible to create an external test set of sufficient size, it is feasible to perform proper validation with artificial test sets. The recommended technique for small training sets is leave-one-out (LOO) cross-validation [24]: One compound is removed as test instance from the training set, a model is created with the remaining training set, and the activity of the test instance is predicted and compared with its real activity. The same procedure is repeated for all compounds of the training set. If the training set size is too big, it is possible to remove a certain fraction (e.g. 1/10th) of the training set compounds at one time, this procedure is called n-fold-cross-validation (10 and 5 fold cross-validation are the most popular choices).

It is important to consider, that feature selection algorithms (and some feature generation algorithms) utilize activity information. It is therefore mandatory that these procedures are performed separately for each fold, otherwise the training data can be severely overfitted despite cross-validation.

3.2 Validation Results

It is beyond the scope of this article to provide an extensive review of results from the literature. The achievable accuracies depend to a large extent on the quality and composition of the training set and the test set and can vary largely from endpoint to endpoint. Rodent carcinogenicity, e.g. is relatively hard to predict and the best accuracies that have been achieved with *in silico* techniques for structurally diverse chemicals lie between 60 and 70%

[28, 25]². Purely *in vitro* techniques can achieve a similar degree of accuracy [25]. It is interesting to see, that the integration of *in vitro* results with *in silico* techniques can lead to significant improvement in predictive accuracy: The integration of mutagenicity data in **lazar** models for carcinogenicity led to $\approx 80\%$ correct predictions for rat carcinogenicity and $\approx 75\%$ accuracy for mouse carcinogenicity (unpublished results of the author).

Other endpoints are much more easy to predict, values between 75% and 80% are e.g. typical for mutagenicity [4]. The following section will present two case studies, where human endpoints (Maximum Recommended Therapeutic Dose and Hepatotoxicity) have been successfully predicted by *in silico* techniques.

4 Prediction of Human Health Effects

In silico tools can help with the prediction of human health effects in two ways. First it is possible to augment the information that is used in the human hazard assessment process with *in silico* data, if measured values are missing. In this case it is desirable to have simultaneous access to models for various mechanistically related toxic endpoints, metabolism and ADME predictions, to have the possibility to search for toxic activities of structurally similar compounds and to have links to the relevant literature. If *in silico* predictions are contradictory or insecure, it might be necessary to perform targeted experiments to solve this unclear issues. A potential pitfall of such an approach is error accumulation. As every extrapolation is associated with an error, it is likely that errors accumulate especially along

²Some classes of chemicals can be predicted with much higher accuracy [4]

a long chain of inference (e.g. *in silico* → *in vitro* → *in vivo* → human).

The second strategy is to predict human health effects directly. This can be done either by implementing an expert system for human health effects or by using data from clinical or epidemiological studies to train a Data Driven System. Two examples shall illustrate the second approach:

In a recent study US FDA researchers trained a MCASE model with Human Maximum Recommended Therapeutic Dose (MRTD) data from clinical studies [29]. In a cross-validation study 63.7% correct predictions were obtained for the 4 classes very active, active, marginally active and inactive, 85.8% of the predictions could differentiate correctly between inactive/moderate vs. active/very active compounds and 94.7% of the predictions had a not more than 1 class difference to the measured value. The accuracy of these predictions is excellent, especially if we consider that the correlation of MRTD with *in vivo* data (rodent maximum tolerated dose) is very poor ($R^2 = 0.20$).

In another investigation Cheng and Dixon trained ensembles of decision trees with (non-idiosyncratic) human hepatotoxicity data compiled from the literature [30]. As the original references did not provide consistent quantitative information on liver toxicity, the authors decided to evaluate the studies on a case-per-case basis for dose-dependent hepatotoxicity. Depending on the results of this expert evaluation the training set compounds were classified as hepatotoxic or non-hepatotoxic. A series of validation experiments with leave-one-out cross-validation, 10-fold cross-validation and an external test set of 54 compounds was performed to evaluate the quality of prediction models. The results of these experiments indicate, that it is possible to predict human hepatotoxicity from chemical

structures alone with more than 80% accuracy.

Both studies clearly demonstrate the possibility of relatively reliable *in silico* predictions for human health effects, and it would be desirable to build models for further human health related endpoints. The main limiting factor in this respect is presently again the limited public availability of clinical and epidemiological data.

5 Conclusion

Present *in silico* tools are mature enough to play an important role in the preclinical assessment of toxic effects. The accuracy of *in silico* predictions can be many cases at least comparable to *in vitro* and *in vivo* alternatives and it is likely that further enhancements are achievable with the integration of biological information.

It is however crucial to know the limitations of *in silico* techniques and to avoid to apply them blindly to every instance. In this respect it is advantageous to use techniques that present the rationales for their predictions in a traceable manner, and that indicate the limitation of their predictions clearly (e.g. compounds that falls beyond the applicability domain of the model, inconclusive evidence for predictions). In the case of unreliable or implausible predictions it is better to perform additional biological experiments to clarify these issues than to trust *in silico* predictions blindly.

References

- [1] Kola I, Landis J: **Can pharmaceutical industry reduce attrition rates?** *Nat Rev* (2004) **3**:711–715.
- [2] Helma C (Ed.): **Predictive Toxicology**. Marcel Dekker, New York (2005).
- ** An introduction to Predictive Toxicology techniques and their implementation in commercial and non-commercial programs.*
- [3] Cronin M, Livingstone D (Eds.): **Predicting Chemical Toxicity and Fate**. CRC Press, Boca Raton, Florida (2004).
- ** An introduction to the (Q)SAR methodology for toxicity and metabolite predictions with many applications for human and environmental endpoints.*
- [4] Benigni R (Ed.): **Quantitative Structure–Activity Relationship (QSAR) Models of Mutagens and Carcinogens**. CRC Press, Boca Raton, Florida (2003).
- ** (Q)SAR techniques and applications for the prediction of carcinogenicity and mutagenicity.*
- [5] Kroemer R: **Molecular modelling probes: Docking and scoring**. *Biochem Soc Trans* (2003) **31**:980–984.
- * A brief survey of Molecular Modelling techniques.*

- [6] Parsons S, McBurney P: **The use of expert systems for toxicology risk prediction.** In: Helma C (Ed.), *Predictive Toxicology*, Marcel Dekker, New York (2005) .
- * *An introduction to the Expert Systems approach in Predictive Toxicology.*
- [7] Payne M: **Computer-based methods for the prediction of chemical metabolism and biotransformation within biological organisms.** In: Cronin M, Livingstone D (Eds.), *Predicting Chemical Toxicity and Fate*, CRC Press, Boca Raton, Florida, 205–227 (2004) .
- * *A review of techniques for metabolite predictions.*
- [8] Woo Y, Lai DY: **OncoLogic: A mechanism-based expert system for predicting the carcinogenic potential of chemicals.** In: Helma C (Ed.), *Predictive Toxicology*, Marcel Dekker, New York (2005) .
- [9] Judson P, Marchant C, Vessey J: **Using argumentation for absolute reasoning about the potential toxicity of chemicals.** *J Chem Inf Comput Sci* (2003) **43**:1364–1370.
- [10] Klopman G, Sedykh A: **META: An expert system for the prediction of metabolic transformations.** In: Helma C (Ed.), *Predictive Toxicology*, Marcel Dekker, New York (2005) .

- [11] Button W, Judson P, Long A, Vessey J: **Using absolute and relative reasoning in the prediction of the potential metabolism of xenobiotics.** *J Chem Inf Comput Sci* (2003) **43**:1371–1377.
- [12] Mekenyan O, Dimitrov S, Pavlov T, Veith G: **A systematic approach to simulating metabolism in computational toxicology. I. the TIMES heuristic modelling framework.** *Curr Pharm Design* (2004) **10**:1273–1293.
- [13] Eriksson L, Johansson E, Lundstedt T: **Regression- and projection-based approaches in Predictive Toxicology.** In: Helma C (Ed.), *Predictive Toxicology*, Marcel Dekker, New York (2005) .
- * *An introduction to statistical (Q)SAR techniques.*
- [14] Klopman G, Ivanov J, Saiakhov R, Chakravarti S: **MC4PC - an artificial intelligence approach to the discovery of Quantitative Structure Toxic Activity Relationships (QSTAR).** In: Helma C (Ed.), *Predictive Toxicology*, Marcel Dekker, New York (2005) .
- [15] Helma C: **Data mining and knowledge discovery in predictive toxicology.** *SAR QSAR Environ Res* (2004) in press.
- * *A mini-review of data mining techniques in Predictive Toxicology.*
- [16] Wegner J, Fröhlich H, Zell A: **Feature selection for descriptor based classification models. 1. theory and ga-sec algorithm.** *J Chem Inf Comput Sci* (2004) **44**:921 – 930.

- [17] Enslein K, Gombar V, Blake B: **Use of SAR in computer-assisted prediction of carcinogenicity and mutagenicity of chemicals by the *TOPKAT* program.** *Mutation Res* (1994) **305**:47–61.
- [18] Poroikov V, Filimonov D: **Pass: Prediction of biological activity for substances.** In: Helma C (Ed.), *Predictive Toxicology*, Marcel Dekker, New York (2005) .
- [19] Helma C: **lazar: Lazy Structure – Activity Relationships for toxicity prediction.** In: Helma C (Ed.), *Predictive Toxicology*, Marcel Dekker, New York (2005) .
- [20] Cronin M: **Toxicological information for use in predictive modeling: Quality, sources, and databases.** In: Helma C (Ed.), *Predictive Toxicology*, Marcel Dekker, New York (2005) .
- * *A description of databases for Predictive Toxicology models.*
- [21] Richard A: **DSSTox web site launch: Improving public access to databases for building structure-toxicity prediction models.** *Preclinica* (2004) **2**:103–108.
- [22] Commission of the European Communities: **White paper: Strategy for a future chemicals policy.** Tech. rep., Commission of the European Communities (2001).
- [23] Cronin M, Jaworska J, Walker J, Comber M, Watts C, Worth A: **Use of QSARs in international decision-making frameworks to predict health effects of chemical substances.** *Environ Health Perspect* (2003) **111**:1391–1401.
- * *A survey of regulatory applications of (Q)SAR models.*

- [24] Hawkins D: **The problem of overfitting.** *J Chem Inf Comput Sci* (2004) **44**:1–12.
- * *An excellent review of validation issues with (Q)SAR models with many examples.*
- [25] Benigni R, Zito R: **The second National Toxicology Program comparative exercise on the prediction of rodent carcinogenicity: Definitive results.** *Mutation Res* (2004) **566**:49–63.
- [26] Helma C, Kramer S: **A survey of the Predictive Toxicology Challenge.** *Bioinformatics* (2003) **19**:1179–1182.
- [27] Benigni R: **QSAR: Validity and validation.** In: **11th International Workshop on Quantitative Structure Activity Relationships in Human Health and Environmental Sciences (QSAR 2004)** (2004) .
- [28] Toivonen H, Srinivasan A, King RD, Kramer S, Helma C: **Statistical evaluation of the Predictive Toxicology Challenge 2000–2001.** *Bioinformatics* (2003) **19**:1183–1193.
- [29] Matthews E, Kruhlak N, Benz R, Contrera J: **Assessment of the health effects of chemicals in humans: I. QSAR estimation of the maximum recommended therapeutic dose (MRTD) and no effect level (NOEL) of organic chemicals based on clinical trial data.** *Current Drug Discovery Technologies* (2004) **1**:61–76.
- * *An application of the MCASE system for the prediction of the human maximum recommended therapeutic dose.*

- [30] Cheng A, Dixon S: **In silico models for the prediction of dose-dependent human hepatotoxicity.** *J Comput Aid Mol Des* (2004) **17**:811–823.

* *Ensembles of decision trees were successfully applied to predict human hepatotoxicity.*

Address

Christoph Helma
Inst. f. Computer Science
University Freiburg
Georges Köhler Allee 79
D-79110 Freiburg
Germany
Phone: ++49-761-203-8013
Fax: ++49-761-203-8007
email: helma@informatik.uni-freiburg.de