

Evaluation of (Q)SAR models for the prediction of mutagenicity potential

Stephanie Ringeissen, Reine Note, Catherine Dochez, Nicole Flamand,
Gladys Ouedraogo-Arras, and Jean-Roch Meunier

L'Oréal Recherche

Corresponding author: Stephanie Ringeissen
1 Avenue E. Schueller, 93600 Aulnay-sous-Bois, France
sringeissen@rd.loreal.com

Abstract

Developing alternative methods to *in vivo* testing is critical to the cosmetic industry based on ethical reasons, the REACH regulation and the 7th Amendment of the European Directive on Cosmetics. A number of (Q)SAR models are commercially available, and building a strategy based on more than one such system is relevant considering the differences in models and applicability domains. The predictive performance of such models has to be assessed on a regular basis, given the chemical diversity and reactivity of new substances, and regular updates in the software versions. Three commercially available computer-assisted prediction models were evaluated for mutagenicity. Those systems include Derek for Windows, a knowledge-based expert system; MULTICASE, a fragment-based statistical system; and TIMES, a 3D QSAR system comprising a metabolism simulator. The selected test set contains chemicals with AMES data on different strains of *Salmonella*, in the presence or the absence of metabolic activation (S9), and pre-incubation. Applicability domains and predictive performance are compared. Such predictive systems provide a valuable support for the screening and categorization of chemicals, and the understanding of mechanistic rationale. Because our chemical space is not fully covered by these systems, there is a need for expanding their applicability domains by integrating in-house data.

Keywords: mutagenicity, (Q)SAR, evaluation, *in silico*, applicability domain

Introduction

A number of *in silico* models predictive of human-health related endpoints are commercially available. Building an *in silico* strategy based on more than one such system is relevant considering the differences in the models (SAR versus QSAR, expert system versus artificial intelligence-based systems) and their applicability domain (chemical space and toxicological endpoint coverage). The predictive performance of such models has to be assessed on a regular basis, given the chemical diversity and reactivity of new chemical entities of interest to industry, and regular updates in the software versions. A number of evaluation exercises have been published over the past years (Patlewicz 2007, Patlewicz 2003, Snyder 2005, Crettas 2005, Snyder 2004, Cariello 2002, Hayashi 2005).

In this context three major commercially available computer-assisted prediction models were evaluated with an external dataset for their ability to predict bacterial mutagenicity:

- Derek for Windows (DfW), a knowledge-based expert system

- MC4PC, a statistically driven fragment-based approach
- TIMES, an hybrid expert system combining structural alerts and 2D/3D-QSAR

Materials & methods

Dataset for models evaluation

A test set of in-house data was compiled with 200 mutagen chemicals and 98 non mutagen chemicals. These chemicals were tested with the AMES test or the mini-mutagenicity test (MMT) (Flamand 2001) on 6 different *Salmonella typhimurium* strains, in the presence or the absence of metabolic activation (rat S9), or pre-incubation (PI). Negative chemicals were found to be non mutagen when tested under all test conditions (Table 1A). Positive chemicals were found to be mutagen in at least one strain and one condition (Table 1B).

A total of 84 out of the 200 known mutagenic chemicals are mutagens only in the presence of rat S9 (PI conditions included); 103 are mutagens without S9. 13 chemicals are reported with equivocal results as to whether the parent and/or metabolite(s) are

mutagen.

Chemical domain

The test set was imported into a **Leadscope Database Manager** (Leadscope Inc). This program provides a classification of compounds using over 27,000 chemistry building blocks which represent functional groups among other features.

In silico models

- **Derek for Windows** (DfW, Lhasa Ltd., UK) version 9.0.0, mutagenicity module.

DfW toxicity assessments are in part based on alerts, chemical features determined to be associated with toxicity. DfW mutagenicity predictions were considered positive if they were associated with a level of likelihood of "equivocal" or higher. All other predictions were considered negative.

- **MC4PC** (MULTICASE Inc.) version 1.8, with two modules.

The two *salmonella* Ames mutagenicity modules A2I (2191 compounds) and the more recent one A2H (5864 compounds) are composite modules that include data from all testing conditions. Data sources include NTP, US EPA GENETOX and FDA. MC4PC is based on the identification

of biophores and consideration of additional descriptors called modulators (eg log P, water solubility, presence of deactivating fragments). When the system gave an "active" or "inactive" call, we considered these calls as "positive" and "negative" respectively. "Inconclusive" calls were considered "out of domain". When the prediction is annotated with warnings ("w") for the presence of unknown structural fragments, we consider it as an indication of the molecule being "out of domain". All "out of domain" calls were excluded from statistical analysis.

- **TIMES** (TIssue MEtabolism Simulator, LMC, Bulgaria) version 2.24.9. updated (to be released) TIMES mutagenicity model includes a mammalian metabolism simulator and combines alerting groups with physico-chemical parameters and 2D/3D molecular descriptors. Data sources include NTP data (1341 compounds) and proprietary data from BASF (1626 compounds). A multi-step domain is incorporated into the model.

Two rules were applied for accepting/rejecting predictions:

- **Rule A:** Molecules were considered "positive" when they were predicted "Mutagen, and considered "negative" when predicted "Not



Fig. 1. Chemical categories present in the test set as generated with the Leadscope Database Manager program

Table 1. Experimental mutagenicity data for a non mutagen compound (A) and mutagen compounds (B). (0) Negative, (1) positive, (2) equivocal, (-) no data.

	TA1535			TA1537			TA1538			TA98			TA100			TA102		
Compound	-S9	+S9	+PI	-S9	+S9	+PI	-S9	+S9	+PI	-S9	+S9	+PI	-S9	+S9	+PI	-S9	+S9	+PI
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	TA1535			TA1537			TA1538			TA98			TA100			TA102		
Compound	-S9	+S9	+PI	-S9	+S9	+PI	-S9	+S9	+PI	-S9	+S9	+PI	-S9	+S9	+PI	-S9	+S9	+PI
2	0	0	-	0	1	-	1	1	-	1	1	-	1	1	-	-	-	-
3	0	0	-	1	0	-	1	0	-	2	0	-	0	0	-	-	-	-
4	0	0	-	0	0	-	0	1	-	0	1	-	0	0	-	-	-	-
5	-	-	-	-	-	-	-	0	1	-	0	1	-	-	-	-	-	-
6	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	-	-	-
7	-	-	-	-	-	-	0	1	-	0	1	-	-	-	-	-	-	-
8	0	0	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0
9	0	0	-	0	0	-	0	1	0	0	0	-	0	0	-	-	-	-
10	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
11	0	0	-	0	0	-	1	0	-	0	0	-	2	0	-	-	-	-
12	0	0	0	0	0	0	1	1	0	1	1	0	0	0	-	-	-	-
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0
15	0	0	-	1	1	-	0	0	0	0	0	0	1	1	-	0	0	0
16	-	-	-	-	-	-	-	-	-	-	-	-	1	1	-	-	-	-
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
18	0	0	0	0	1	-	0	0	0	0	2	2	-	-	-	-	-	-
19	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0

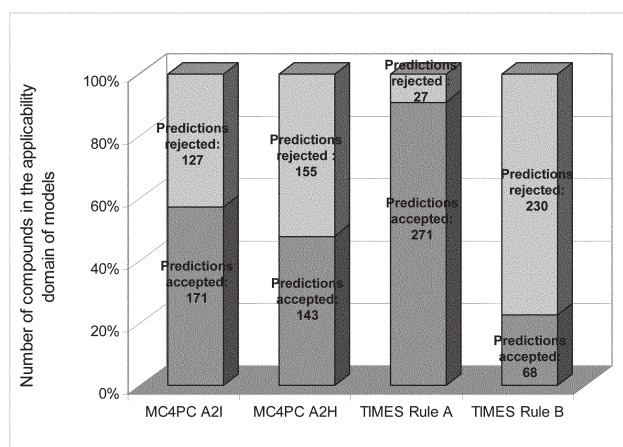


Fig. 2. Coverage of the test set with the MC4PC and TIMES models (numbers of accepted predictions on the bars of the histogram correspond to predictions included in statistical analysis)

mutagen (whether "in domain or "out of domain). Any call annotated with "Can't predict was considered "out of domain, and therefore excluded from statistical analysis.

- **Rule B:** Molecules were considered "positive when they were predicted "Mutagen/In domain, and considered "negative when predicted "Not mutagen/In domain. Any call annotated with "Can't predict was considered "out of domain, and therefore excluded from statistical analysis.

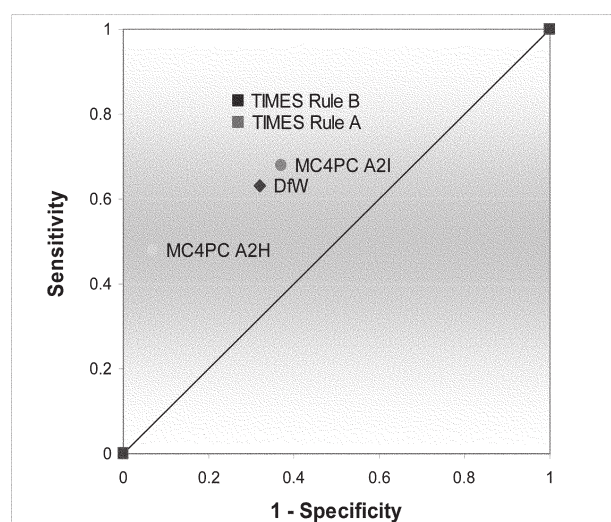


Fig. 3. ROC graph of mutagenicity predictions for the models evaluated.

Results

For each model, applicability domain (AD) and predictive performance are discussed.

Chemical domains

The test set includes a large proportion of aromatic amines since they represent 184 out of the 200 known mutagens and 57 out of the 98 known non-mutagens (refer to Fig. 1 which illustrates the chemical categorisation provided by the Leadscape Database Manager program.

Table 2. Performance of models in terms of (i) specificity and false positives (referred to the total number of non mutagens) and (ii) sensitivity and false negatives (referred to the total number of mutagens)

	False negatives	Sensitivity	False positives	Specificity
MC4PC A2I	38/122	69%	18/49	63%
MC4PC A2H	45/86	48%	4/57	93%
DfW	75/125	63%	31/98	68%
TIMES Rule A	41/187	78%	23/84	73%
TIMES Rule B	7/42	83%	7/26	73%

Table 3. Example of differences observed in terms of biophores between the A2I and A2H MC4PC modules.

Compound	Biophores identified with model A2H	Biophores identified with model A2I
X	None	cH =cH -cH =c. -cH =
Y	cH =cH -c."-CH -c. =cH -cH =cH -c =c. -	None

Table 4. DfW mutagenicity alerts triggered in the test set

Mutagenicity: Number/Name of the alert	Number of times the alert was fired	Number of times the alert was correctly fired	Number of times the alert was incorrectly fired	Number of references	Number of examples
007, N-Nitro or N-nitroso compound	1	1	0	2	4
028, Mono- or di-alkylhydrazine	1	1	0	9	5
039, Arylhydrazine	1	1	0	1	2
312, alpha,beta-Unsaturated imine	1	1	0	5	0
327, Quinone	1	1	0	12	3
329, Aromatic nitro compound	19	19	0	18	6
330, Aromatic azo compound	13	8	5	5	0
351, Aromatic amine or amide	11	11	0	23	5
352, Aromatic amine or amide	77	55	12	25	5
353, Aromatic amine or amide	10	3	7	31	0
354, Aromatic amine or amide	35	28	7	23	5

It is important to know when a test compound is adequately represented in the AD of a given model. MC4PC and TIMES domains have been studied in the present work (Fig. 2). As observed for MC4PC modules and TIMES Rule B, the test set chemical space is not adequately covered by the training sets of these systems. TIMES Rule A provides a better coverage but at the cost of less reliable predictions.

DfW does not use a training set and therefore cannot be validated in the same way. However, the AD can be defined by the scope of the alerts implemented in the system. Such an evaluation could be carried out based on predictive performance. Also, one must be aware that the absence of predictions does not mean that the test compound is negative.

Model performances

Fig. 3 displays the Receiver Operating Characteristic (ROC) graph which allows to compare

simultaneously the performance of the various systems (Crettaz 2005). The top left corner is the ideal performance, the bottom right corner the worst performance. It appears that QSAR-based models perform better than SAR-based models.

False predictions, sensitivity and specificity are detailed for each model in Table 2.

One reason for the improved specificity but decreased sensitivity (as shown with the increased rate of false negatives) of MC4PC recent A2H module when compared to the older one A2I could be for instance the experimental conditions of the test used. For instance, current AMES tests tend to involve higher concentrations of chemicals as well as PI conditions (10 out of 200 chemicals of the test set are mutagens only under PI conditions). Chemicals tested non mutagen in old protocols may give positive results in recent ones.

Table 3 gives an example of two molecules which

contain biophores only in one of the two MC4PC modules tested, thus showing some of the differences observed between the 2 modules.

DfW performance was average, with a sensitivity of 63%. A total of 156 molecules fire at least one DfW alert: 142 for one alert and 14 for two alerts with a large proportion of the test set (133 chemicals) firing for the DfW "Aromatic amine or amide alerts (Table 4). It is noteworthy that alert 353 overfires 7/10 times, which highlights the need for improving this rule.

TIMES performs well both in terms of sensitivity and specificity with a better overall performance for Rule B than Rule A. Rule B provides predictions only for a reduced number of chemicals. Alternatively one could accept predictions provided by Rule A but at the cost of a decreased reliability on predictions.

Discussion

It is recognised that there is an added-value in evaluating with external test sets the ability of models to predict a given endpoint (Benigni 2007). This is why the present evaluation of commercial (Q)SAR models for non-congeneric sets of chemicals was undertaken with a set of in-house data. Models evaluated in this study provide a valuable support for the screening and categorization of chemicals in addition to further understanding of mechanistic rationale. Because our chemical space is not fully covered by these systems, there is a need for expanding their AD by integrating in-house data.

Given (i) the strengths and weaknesses of the different models used and (ii) the current regulatory context, it is important to:

- rely on a battery of tools that cover complementary chemical spaces and provide different rules/algorithms
- consider human expert analysis (to check for instance the relevance of an alert) and other sources of information, as proposed with the implementation of Integrated or Intelligent Testing Strategies, also known as ITS (<http://se.setac.org/files/setac-eu-0032-2007.pdf>)

Given the complexity of the mechanisms involved, results should be analysed in a context-dependent environment, on a case-by-case basis.

Since QSAR-based tools appeared to perform better on the test set used, it would be interesting to evaluate local QSARs such as those dedicated to congeneric chemicals (eg aromatic amines) to get a greater insight into the validity of such an approach.

To follow upon this work, it will be interesting to

- investigate the performance of the TIMES metabolism simulator to predict chemicals which are mutagens only in the presence of S9

- expand - if possible - AD of models to improve predictions accuracy on chemicals of interest to our industry
- refine DfW alerts and develop customized rules
- assess the performance of local QSARs dedicated to aromatic amines using the test set used in this study.

Acknowledgement

We are grateful to Dr Romualdo Benigni for his useful comments in preparing this document.

References

- Benigni, R. et al. (2007) Collection and evaluation of (Q)SAR models for mutagenicity and carcinogenicity EUR 22772 EN.
- Cariello, N.F. et al. (2002) Comparison of the computer programs DEREK and TOPKAT to predict bacterial mutagenicity. *Deductive Estimate of Risk from Existing Knowledge. Toxicity Prediction by Komputer Assisted Technology, Mutagenesis*, 17, 321-329.
- Crettaz, P. and Benigni, R. (2005) Prediction of the rodent carcinogenicity of 60 pesticides by the DEREKfW expert system, *J. Chem. Inf. Model.*, 45, 1864-1873.
- Flamand, N. et al. (2001) Mini mutagenicity test: a miniaturized version of the Ames test used in a prescreening assay for point mutagenesis assessment, *Toxicol. In Vitro*, 15, 105-114.
- Hayashi, M. et al. (2005) In silico assessment of chemical mutagenesis in comparison with results of salmabella microsome assay on 909 chemicals, *Mut. Res.*, 588, 129-135.
- Patlewicz, G. et al. (2007) *Regul. Toxicol. Pharmacol.*, 48, 225-239.
- Patlewicz, G. et al. (2003) Quantitative structure-activity relationships for predicting mutagenicity and carcinogenicity, *Environ. Toxicol. Chem.*, 22, 1885-1893.
- Snyder, R.D. and Smith, M.D. (2005) Computational prediction of genotoxicity: room for improvement, *Drug Discovery Today*, 10, 1119-1124.
- Snyder, R.D. et al. (2004) Assessment of the sensitivity of the computational programs DEREK, TOPKAT, and MCASE in the prediction of the genotoxicity of pharmaceutical molecules, *Environ. Mol. Mutagen.*, 43, 143-158.

