Novel QSAR Models for Predicting Toxicity of Chemicals to Aquatic Organisms and Identifying the Mode Of Action

Kiril Lanevskij[1,2], Liutauras Juska[1,2], Remigijus Didziapetris[1], Pranas Japertas[1]

[1] ACD/Labs, Inc., A.Mickeviciaus g. 29, LT-08117 Vilnius, Lithuania,
[2] Department of Biochemistry and Biophysics, Vilnius University, M.K.Ciurlionio g. 21/27, LT-03101 Vilnius, Lithuania.

INTRODUCTION

This study presents the application of recently introduced GALAS modeling methodology for estimating toxicity of chemicals to several aquatic species. Experimental data were expressed as median lethal concentration of test compound in water (LC_{50}) and the data set contained 904 LC_{50} values for fishes (*Pimephales promelas*) and 589 LC_{50} values for crustaceans (*Daphnia magna*). The utilized modeling approach was validated by applying the same principles to develop a predictive model for IGC_{50} (50% inhibitory growth concentration) to protozoan *Tetrahymena pyriformis*. This model was submitted as an entry for environmental toxicity prediction challenge hosted by CADASTER project. The model derived using known IGC_{50} values for 644 compounds was identified among the winners achieving RMSE under 0.8 log units for blind validation set of 120 chemicals. It is also demonstrated that the chemicals' Mode Of Action (MOA) can be determined using a simple set of structural fragments associated with certain MOA classes.

EXPERIMENTAL DATA

Experimentally determined LC_{50} values were collected from earlier modeling works and original publications. Initially, the data set involved two aquatic species frequently used for testing:

•Fathead minnow (*Pimephales promelas*)–904 compounds
•Water flea (*Daphnia magna*)–589 compounds

The compiled set of LC_{50} values for fathead minnows overlapped to a large extent with the respective data set available from the PubChem project (AID 1188) [1].

METHODS

1. LC_{50} modeling

The predictive models described in this study were derived using the recently introduced GALAS (**G**lobal, **A**djusted **L**ocally **A**ccording to **S**imilarity) modeling methodology. [2] A brief outline of the modeling process is presented in Scheme. 1.

Each GALAS model consists of two parts:
1) Global (baseline) model that reflects general trends in the variation of the property of interest.
2) Similarity-based routine that performs local correction of baseline predictions taking into account the differences between baseline and experimental LC_{50} values for the most similar training set compounds

Baseline and Excess Toxicity:
The approach outlined above is particularly suitable for modeling toxicity-related properties as it closely resembles the well-known 'excess toxicity' concept [3] stating that:

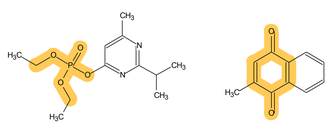
1) Inert chemicals exhibit baseline toxicity levels that are consistent with their lipophilicities:

$$\log LC_{50}(\text{baseline}) \approx f(\log P)$$

2) Toxicity of more reactive compounds exceeds the expected baseline levels and this ratio of LC_{50} values is designated as Excess Toxicity (T_e):

$$T_e = \frac{LC_{50}(\text{baseline})}{LC_{50}(\text{exp.})}$$

The global part of the GALAS model provides an estimate of baseline LC_{50}, which can be easily described by molecular structure. In the current study global model is based on a predefined set of fragmental descriptors (including 'toxicophores' described in the next section). and built using PLS statistical method. The local part corrects the systematic deviations produced by the baseline model which are equivalent to the respective T_e values.



A predefined list of fragments

Variables

Structures

PLS + Multiple Bootstrapping

SD

Similarity analysis

Name	pLC50
Prop-2-enal	3.52
4-Octylaniline	3.23
Dibenzofuran	2.05
1,4-Xylene	1.08
Benzamide	-0.74

Multiple predictors Coefficients (a_i)

Self-training Library

Baseline $pLC_{50} = \sum_i a_i x_i + c$

Correction ΔpLC_{50}

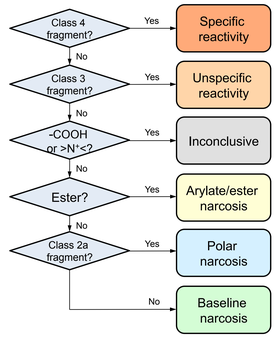
$$pLC_{50} = \text{Baseline } pLC_{50} + \Delta pLC_{50}$$

SCHEME 1. An outline of the model development process

2. Predicting Mode of Action

The identification of structural fragments involved in specific mechanisms of toxic action was performed utilizing the information regarding the chemicals' Mode Of Action (MOA) provided in the PubChem Fathead minnow data set [1] as well as the original work by Verhaar [3] . Examples of fragments associated with different MOAs are listed in Table 1.

Analysis of PubChem data yielded a set of 33 fragments that allowed discriminating between classes of compounds acting by different mechanisms. Classification was performed using a different approach compared to [3]. First, compounds containing a specifically acting substructure (Class 4) or reactive electrophile/proelectrophile moieties (Class 3) were identified. Compounds without such reactive groups were not classified if they contained a carboxylic acid group or were permanently charged cations. The reason is the lack of evidence regarding their mechanism of action, as for most compounds belonging to this group MOA was not determined in the PubChem data set. Finally, the least hazardous classes were assigned based on the presence of the corresponding functional groups (see Scheme 2 for details).



SCHEME 2. Decision tree for MOA classification

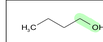
Fragment	MOA	Comment
	1. Baseline narcosis	Baseline toxicity levels are typical of inert compounds such as aliphatic alcohols
	2a. Polar narcosis	Less inert compounds: primary alkylamines, phenols, aromatic amines
	2b. Arylate/ester narcosis	Less inert compounds: esters of alkyl- or arylcarboxylic acids
	3. Unspecific reactivity	Molecules with electrophile or proelectrophile functionality
	4. Specific reactivity	Specific mechanism of action, e.g., cholinesterase inhibition by carbamates

TABLE 1. Typical structural fragments associated with certain Modes Of Action (MOA)

LC_{50} MODELING RESULTS

As shown in Table 2 and Fig. 1, the derived model for predicting LC_{50} of chemicals to fishes produces sufficiently accurate predictions for test set compounds falling within the Model Applicability Domain (i.e., obtaining *RI* ≥ 0.3). Even lower RMSE values are observed if predictions are further filtered by Reliability Index values. Notably, the majority of compounds in the test set obtain predictions of at least borderline reliability, and more than half of these are provided with moderate or high reliability predictions (*RI* ≥ 0.5). Similar results were obtained for *D. magna* model (data not shown).

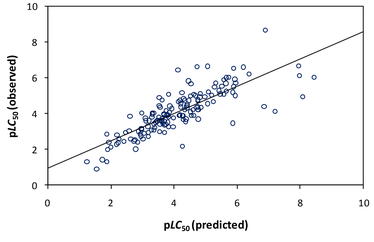


FIGURE 1. Observed vs predicted pLC_{50} (*P. promelas*) for test set compounds with *RI* ≥ 0.3.

METHOD VALIDATION

The described methodology was validated by applying the same considerations to build a new model for predicting toxicity against *T. pyriformis* as an entry for the Environmental Toxicity Prediction Challenge organized in 2009 by the CADASTER project [5]. The participants were provided with the following data:

Subset	R^2	RMSE	% of Test set covered
RI ≥ 0.3	0.656	0.797	85.7%
RI ≥ 0.5	0.795	0.501	59.4%
RI ≥ 0.7	0.880	0.363	28.0%

TABLE 2. Performance of pLC_{50} (*P. promelas*) model on the internal validation set (N = 175).

- 644 compounds with known IGC_{50} values–training set for model development
- 449 compounds with known IGC_{50} values–test set for internal validation and preliminary ranking
- 120 compounds with no published data–blind validation set for final ranking.

Training and test sets were taken from [6-7], while data for the blind set were provided by prof. T.W. Schultz. IGC_{50} values for these compounds were published in September, 2009 after the competition had ended.

VALIDATION RESULTS

The GALAS model developed in the current study was identified among first-pass winners of the competition with RMSE for predicting blind data set compounds (Fig. 2) not significantly different from the best entry. Fig. 3 shows a clear negative correlation between RMSE of predictions and calculated *RI* values for the same data set. It is evident that molecules with large prediction errors (> 1 log unit) typically obtain low or borderline *RI* values, while predictions marked 'highly reliable' are of comparable accuracy to experimental data. Presented results demonstrate that GALAS modeling methodology is suitable for producing 'state-of-the-art' quality models. Additionally, the Reliability Index can serve as a useful tool for discriminating confident predictions from those expected to be less accurate.

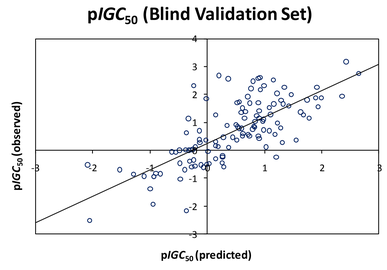


FIGURE 2. Observed vs predicted pIGC_{50} for blind data set (N = 120; R^2 = 0.545; RMSE = 0.794).

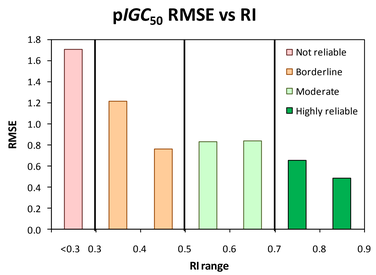


FIGURE 3. Average RMSE of pIGC_{50} predictions within different *RI* ranges.

MOA PREDICTION RESULTS

Of the overall 617 compounds with toxicity data available in the PubChem data set, MOAs were reported for 462 molecules. For the remaining compounds MOA could not be determined due to insufficient of conflicting evidence. The classification scheme used in PubChem could be referred as "Extended Verhaar scheme" since in the original publication [3] the class of "Less inert compounds" was considered without further distinction to "Polar narcosis" and "Arylate and ester narcosis" modes of action.

The proposed classification scheme demonstrated very good discriminating power and achieved the overall accuracy of 86% for the five-class classification problem (see Table 3). Notably, only the compounds with undetermined MOA obtained "Inconclusive" class labels, so that all 462 molecules with MOA data could be classified. It can be concluded that derived fragment set could be useful both for *in silico* evaluation of the mode of toxic action, and as a descriptor set for building QSAR models.

Overall classification accuracy:

%correct = 85.9% (397/462)

Cramér's contingency coefficient:

***V* = 0.827**

$$\left(V = \sqrt{\frac{\chi^2}{N(k-1)}}\right)$$

Exp.\Pred.	Baseline	Polar	Ester	Unspecific	Specific
Baseline	**202**	19	9	10	3
Polar	1	**47**	0	1	0
Ester	0	0	**25**	1	0
Unspecific	12	3	1	**79**	2
Specific	2	0	0	1	**44**

TABLE 3. Statistical performance characteristics of the developed classification scheme on PubChem data set.

REFERENCES

[1] NCBI PubChem DB (http://pubchem.ncbi.nlm.nih.gov/).
[2] Sazonovas A et al. *SAR QSAR Environ Res.* **2010,** 21, 127.
[3] Lipnick RL. *Sci Total Environ.* **1991,** 109–110, 131.
[4] Verhaar HJM et al. *Chemosphere.* **1992,** 25, 471.
[5] CADASTER project (http://www.cadaster.eu)
[6] Zhu H et al. *J Chem Inf Model.* **2008,** 48, 766.
[7] Tetko I et al. *J Chem Inf Model.* **2008,** 48, 1733.

ACD/Labs
Advanced Chemistry Development, Inc.
Tel: (416) 368-3435
Fax: (416) 368-5596
Toll Free: 1-800-304-3988
Email: info@acdlabs.com
www.acdlabs.com