INTRODUCTION

This study focuses on the application of the recently introduced GALAS modeling methodology to develop predictive models that allow estimating toxicity of new chemicals to several aquatic species. Experimental data used in the analysis were expressed as the median lethal concentration of test compound in water (LC50) representing a compound’s toxicity to fish and crustaceans. The overall data set collected from literature contained toxicities of 900 compounds to fathead minnows (Pimephales promelas) and almost 600 LC50 values determined for water fleas (Daphnia magna).

The modeling approach utilized herein for aquatic toxicity was validated by applying the same principles to develop a new model that predicts IG50 (50% inhibitory growth concentration) to protozoan Tetrahymena in batches. This model was submitted as an entry for the Environmental Toxicity Prediction Challenge hosted by the CADASTER project. The final model derived using known IG50 values for 644 compounds was identified among the winners achieving RMSE under 0.8 kg units for prediction of a validation set containing 120 chemicals.

EXPERIMENTAL DATA

Experimentally determined LC50 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)–644 compounds
- Water flea (Daphnia magna)–688 compounds

The compiled set of LC50 values for fathead minnows overlapped to a large extent with the respective data set available from the PubChem project [Add 11888] [1].

METHODS

The predictive models described in this study were derived using the recently introduced GALAS (Global, Adjusted Locally According to Similarity) modeling methodology. A schematic view of the modeling process is presented in Fig. 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 LC50 values for the most similar training set compounds

Baseline models are based on fragmental descriptors and are built using a PLS method with multiple bootstrapping. Molecule fragmentation is performed using a predefined list of fragments that include various atom types, functional groups, and interactions describing general chemical composition as well as specific 'toxicophore' fragments.

Local part of the model provides the basis for two essential features:

- Evaluation of the Model Applicability Domain by the means of calculated Reliability Index
- Model Tunability—expansion of Model Applicability Domain by incorporating user-defined data without rebuilding the model

Reliability indices (RI) RI values ranging from 0 to 1 provide a quantitative estimate of the prediction accuracy. RI value for a compound depends on the following aspects:

- Similarity to the training set compounds
- Consistency of experimental data for similar molecules

Toxicophores: The identification of fragments involved in specific mechanisms was performed utilizing the information regarding the chemical’s Mode Of Action (MOA) provided in the PubChem Fathead minnow data set as well as the original work by Verhaar [5] (see Table 1 for examples of fragments associated with different MOAs).

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 [4] stating that:

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

\[ \log \text{LC50 (baseline)} = \log P \]

2) Toxicity of more reactive compounds exceeds the expected baseline levels and the excess LC50 values is designated as Excess Toxicity (TE):

\[ T = \text{LC50 (baseline)} - \text{LC50 (exp)} \]

The global part of the GALAS model provides an estimate of baseline LC50, which can be easily described by molecular structure. The local part corrects the systematic deviations produced by the baseline model which are equivalent to the respective TE values.

TABLE 1. Typical structural fragments for molecules exhibiting certain Modes Of Action (MOA)

<table>
<thead>
<tr>
<th>Fragment</th>
<th>MOA</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>MOA</td>
<td>Baseline toxicity levels are typical of inert compounds such as aliphatic alcohols.</td>
</tr>
<tr>
<td>Polar</td>
<td>MOA</td>
<td>Less inert compounds: primary alcohols, phenols, aromatic amines, are especially toxic.</td>
</tr>
<tr>
<td>Unspecified reactivity</td>
<td>MOA</td>
<td>Molecules with electrophilic or nucleophilic functional groups.</td>
</tr>
<tr>
<td>Specific reactivity</td>
<td>MOA</td>
<td>Specific mechanism of action, e.g., cholesterol esterification by carbamates.</td>
</tr>
</tbody>
</table>

RESULTS

As shown in Table 2 and Fig. 2, the derived model for predicting pLC50 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 those are provided with moderate or good predicted reliability. Similar results were obtained for D. magna model.

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 following data were provided to participants:

- 644 compounds with known IG50 values—training set for model development
- 449 compounds with known IG50 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. IG50 values for these compounds were published in September, 2009 after the competition had ended.

SOFTWARE FOR PREDICTING AQUATIC TOXICITY

The presented predictive models of acute toxicity to fish and water fleas are available as part of ACD/Tox Suite software (www.acdlabs.com/acd_atmos). It provides a standard predictor interface which calculates LC50 for the respective species using built-in libraries, displays experimental data for most similar compounds, and features a Trainable module interface that enables a user to extend the accuracy of predictions by incorporating in-house data. (Fig. 5). Calculations for multiple compounds in Batch mode are also available.

REFERENCES

5. CADASTER project (http://cadaster.eu).