Toxicity Prediction in StarDrop

Introduction

Whether compounds are intended as drugs, cosmetics, agrochemicals or for other industrial application, it is essential to understand their potential to cause toxic effects. This can be taken into account when prioritising compounds for further research or considering the most appropriate downstream experiments to confirm their safety. The ability to predict toxicities based on chemical structure alone would allow these factors to be considered prior to synthesis, allowing the safest options to be pursued and saving time and resources wasted on synthesis and testing of unsuitable compounds.

However, the chemical and biological mechanisms leading to toxicities are complex and varied. Therefore, prediction of toxicities from chemical structure is a major challenge and even the best models have a high degree of uncertainty in the predictions they make. This is especially true of 'global' models that attempt to cover a wide range of chemical diversity. Therefore, when considering the results from toxicity predictions, it is essential to give them an appropriate degree of weight in the decision-making process; rejecting an area of chemistry purely on the basis of a predicted toxicity would run the risk of missing valuable opportunities, unless viable alternatives are available that are more likely to be benign.

Therefore, StarDrop provides an ideal environment for application of toxicity models to the selection of compounds. StarDrop's probabilistic scoring algorithm (1) allows predictive models of toxicity to be used, while explicitly taking into consideration the uncertainty in each prediction. This ensures that uncertain predictions are not given undue weight, relative to other data, when prioritising compounds that are more likely to have an appropriate *balance* of properties.

This paper describes the generation and validation of Quantitative Structure Activity Relationship (QSAR) models of key toxicity endpoints, based on data made available by the US Environmental Protection Agency (EPA) as part of its Toxicity Evaluation Software Tool (T.E.S.T.) toolkit (2) . The models were built with StarDrop's Auto-Modeller module and are available to all StarDrop users free-of-charge to download from Optibrium's on-line community at http://www.optibrium.com/community.

OECD Principles

The Organisation for Economic Cooperation and Development (OECD) agreed the following statement of principles for validation of QSAR models at the 37th Joint Meeting of the Chemicals Committee and Working Party on Chemicals, Pesticides and Biotechnology in November 2004.

"To facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information:

1) a defined endpoint
2) an unambiguous algorithm
3) a defined domain of applicability
4) appropriate measures of goodness-of–fit, robustness and predictivity
5) a mechanistic interpretation, if possible" (3)

We will consider each of these principles and the associated guidelines published by the OECD (4) as we discuss the data and methods used to build and validate the models described in this paper.

Overview

In the following section we will describe the data used to build the models herein, obtained from the EPA and previously used to build the models in the EPA's T.E.S.T. software. In the Methods section, we will provide an overview of the methods used to build models of this data in StarDrop and under Results we will summarise the validation results for the resulting models and compare these with the results for the T.E.S.T. models. Finally, we will summarise the outcome and draw some conclusions.

Data

The data used to build these models were obtained from the EPA website and are provided with version 4.0 of the EPA's T.E.S.T (2). Data sets for the following properties were used to build QSAR models in StarDrop:

- The negative logarithm of the molar concentration of the test chemical in water that causes 50% of fathead minnow to die after 96 hours (Fathead minnow -log(LC50))
- The negative logarithm of the molar concentration of the test chemical in water that causes 50% of *Daphnia magna* to die after 48 hours (Daphnia magna -log(LC50))
- The negative logarithm of the molar concentration of the test chemical in water that causes 50% growth inhibition to *Tetrahymena pyriformis* after 48 hours (Tetrahymena pyriformis -log(IGC50))
- The negative logarithm of the amount of chemical in mol/kg body weight that causes 50% of rats to die after oral ingestion (Rat oral -log(LD(50))
- The ratio of the chemical concentration in fish as a result of absorption via the respiratory surface to that in water at steady state (Bioconcentration Factor)
- Whether or not a chemical causes developmental toxicity effects to humans or animals (Developmental tox.)
- Whether or not a compound is positive for mutagenicity, i.e. if it induces revertant colony growth in any strain of *Salmonella typhimurium* (Ames mutagenicity)

All of these represent common approaches to determining well-defined endpoints, provided as examples in the OECD guidance document on the validation of QSAR models (4). Full details of how the datasets were generated can be found in the User's Guide for T.E.S.T. (version 4.0) (5). The data sets, divided into training and independent prediction sets, can be downloaded from http://www.epa.gov/nrmrl/std/cppb/qsar/DataSets.zip.

The sizes of the training and prediction data sets for each property are summarised in Table 1. The training and prediction data sets used in each case are identical to those used in T.E.S.T. 4.0 to permit direct comparison of the T.E.S.T. models with those generated in StarDrop. The data set splits were generated randomly by the T.E.S.T. project with the exception of the Developmental tox endpoint, which was originally generated as part of the CAESAR project using rational design (6).

Table 1 Summary of data set sizes.

Property	Training set size	Test set size
Fathead minnow -log(LD50)	652	164
Daphnia magna -log(LD50)	269	68
Tetrahymena pyriformis -log(LD50)	867	217
Rat oral -log(LD50)	5935	1484
Bioconcentration Factor	541	136
Developmental tox.	227	58
Ames mutagenicity	4594	1149

Methods

StarDrop's Auto-Modeller was used to build all of the models in this study. Full details of the methods employed by the Auto-Modeller can be found in Chapter 6 of the StarDrop Reference Guide (7)and the references therein. However, these are briefly summarized here.

Descriptors

2D SMARTS based descriptors, which are counts of atom type and functionalities, and whole molecule properties such as logP, molecular weight, and polar surface area (a total of 330 descriptors) were calculated. All of the descriptors are listed in detail in Appendix 10.3 of the StarDrop Reference Guide (7).

The calculated descriptors were subjected to a descriptor pre-selection step that removed descriptors with low variance and low occurrence. Specifically, descriptors with a standard deviation less than 0.0005 and descriptors represented by less than 4% of the compounds in the training set were excluded. Also, highly correlated descriptors are excluded (when the pairwise correlation exceeded 0.95 in the training set), such that just one of the pair remained.

Modelling Methods

For regression models (Fathead minnow -log(LD50), Daphnia magnia -log(LD50), Tetrahymena pyriformis -log(LD50), Rat oral -log(LD50) and Bioconcentration factor) the following methods were applied to the training set in order to build predictive models:

- Partial Least Square (PLS) (8)
- Radial Basis Function fitting (RBF) (7)
- Radial Basis Function fitting with Genetic Algorithm descriptor selection (RBF-GA) (7)
- Gaussian Processes (GP) with the following methods for hyperparameter determination (9)
    - Fixed (GPFixed)
    - 2DSearch (GP2DSearch)
    - Forward Variable Selection (GPFVS)
    - Rescaled Forward Variable Selection (GPRFVS)
    - Optimised (GPOpt)
    - Nested Sampling (GPNest)

In the case of Rat oral -log(LD50), the data set was too large for the GP methods, other than GPFixed, to be computationally tractable. Similarly, in the case of Bioconcentration Factor, the GPNest method was intractable. In these cases, the remaining methods were all used to train models of the respective properties.

For classification models (Developmental tox. and Ames mutagenicity) a range of decision tree (DT) algorithms based on the C4.5 algorithm introduced by Quinlan (10) were employed. In addition, for the Developmental tox. dataset, a Gaussian Processes classification algorithm was applied (11); however, this method was intractable for the large Ames mutagenicity data set.

All of these algorithms are well-defined, as required by the OECD principles for QSAR model validation.

For each model the 'chemical space' which is represented by the training set is captured (also described as the 'domain of applicability'). The position of a new compound relative to the chemical space of a model is reflected in the reported confidence in the prediction for the new compound. The StarDrop models employ the Hotelling's T^2 method for representing the chemical space of the model (for more information see Section 2.6 of the StarDrop Reference Guide (7)). Thus, the models generated by the StarDrop Auto-Modeller conform to the OECD principle that a QSAR model should have a defined domain of applicability. Furthermore, the GP and DT algorithms also provide an estimate of the uncertainty in the prediction of each individual compound within the chemical space.

Validation

Each model was validated by application to the independent prediction set for the relevant property. This is a true measure of the predictive performance of a model, which is preferable to an internal measure of performance such as goodness of fit to the training set or cross validation.

The predictive performance of each regression model for a numerical property was assessed by calculation of the coefficient of determination:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i^{pred} - y_i^{obs})^2}{\sum_{i=1}^{N}(y_i^{obs} - \overline{y_i^{obs}})^2},$$

and the root-mean-square error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}(y_i^{pred} - y_i^{obs})^2}$$

where y_i^{pred} and y_i^{obs} are respectively the predicted and observed values of the property for compound *i* and *N* is the number of compounds in the prediction set. For each of the numerical properties, the model with the highest R^2 and lowest RMSE was selected.

The coefficient of determination ranges from 0 to 1 and the closer it is to 1 the better the model describes the proportion of the variation in the observed property values that is explained by the fitted regression, e.g. if we have $R^2=0.85$ this means that 85% of the variation in the property is explained by the model. Note that this definition of R^2 is different from the Pearson correlation coefficient,

$$R^2_{\text{Pearson}} = \frac{\left(\sum_{i=1}^{N}(y_i^{pred} - \overline{y_i^{pred}})(y_i^{obs} - \overline{y_i^{obs}})\right)^2}{\sum_{i=1}^{N}(y_i^{pred} - \overline{y_i^{pred}})^2 \sum_{i=1}^{N}(y_i^{obs} - \overline{y_i^{obs}})^2},$$

which is a measure of how well the predicted versus observed values fit to a straight line, but not the ideal line. However, as the R^2_{Pearson} value was used in the validation of the models provided by the T.E.S.T. package (unfortunately also denoted R^2 in the user guide for this package), this was also calculated for each model generated using StarDrop. Similarly the mean-absolute-error (MAE),

$$MAE = \frac{1}{N}\left|y_i^{pred} - y_i^{obs}\right|,$$

the slope of the best fit line constrained to pass through the origin, *k*, and

$$\frac{R^2_{\text{Pearson}} - R^2_0}{R^2_{\text{Pearson}}},$$

where R^2_0 is the correlation coefficient of the best fit line forced to go through the origin, were calculated for comparison with the models in the T.E.S.T. package. Generally, a regression model is considered to have acceptable predictive power if $R^2_{\text{Pearson}} > 0.6$, $\frac{R^2_{\text{Pearson}} - R^2_0}{R^2_{\text{Pearson}}} < 0.1$ and $0.85 \leq k \leq 1.15$ (12).

For models of classification properties, the performance on the independent test set was evaluated by calculating the concordance, specificity and sensitivity. The concordance is the fraction of all compounds that are correctly predicted, the sensitivity is the fraction of observed active compounds that are correctly predicted (true positives) and the specificity is the fraction of experimentally inactive compounds that are correctly predicted (true negatives). Furthermore, the kappa statistic was calculated, which summarises all of this information by assessing the agreement between observed and predicted classifications with adjustment for chance:

κ = (Observed Agreement- Chance Agreement)/(Total – Chance Agreement).

Generally, a classification model is considered to have acceptable predictive power if $\kappa > 0.6$. For more details, see Chapter 6.8.7 of the StarDrop Reference Guide (7).

This validation procedure conforms with the OECD principle that each model should have “appropriate measures of goodness-of–fit, robustness and predictivity”.

Results

A summary of the results for the best StarDrop models for the numerical properties, along with a comparison with the best models in the T.E.S.T. v4.0 package are shown in Table 2.

Table 3 shows a similar comparison for the categorical properties.

All of the StarDrop models for numerical properties meet the minimum standards with R^2 and $R^2_{Pearson}$ greater than 0.6 and in some cases significantly exceed this goal. In the case of the categorical properties, the Developmental tox. model achieves the goal of $\kappa>0.6$, while the Ames mutagenicity model is marginal with $\kappa=0.58$.

The performance of the best StarDrop model for each property is equal to or better than the corresponding T.E.S.T. model, with only two exceptions: For Daphnia magna -log(LC50) the StarDrop model achieves marginally worse results in some metrics than the T.E.S.T. model, e.g. the RMSE differs by 0.01 log units; and for Ames mutagenicity the StarDrop model achieves a concordance of 0.79 versus 0.80 for the T.E.S.T. model. Neither of these differences is significant.

Glowing Molecule™

The StarDrop models also benefit from the Glowing Molecule visualization that highlights regions of a compound that have a significant influence on a predicted property (an example of this is shown in Figure 1). This provides a link between the predicted toxicity and the chemical mechanism giving rise to this result, helping to guide the redesign of compounds with improved safety and fulfils the OECD principle to provide "a mechanistic interpretation, if possible" (although, of course, this does not give any insight into the biological mechanism for the predicted toxicity).

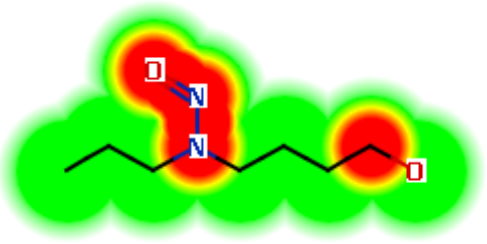

Figure 1 The Glowing Molecule highlights regions of a molecule that have a significant impact on the predicted toxicity. In this case the nitrosamine is, unsurprisingly, identified as the biggest issue in this prediction of Ames mutagenicity.

Table 2 Summary of StarDrop results for best models of numerical properties and comparison with best T.E.S.T. models.

Property	Best StarDrop Model							Best T.E.S.T. Model					
	Method	R^2	RMSE	$R^2_{Pearson}$	$\frac{R^2_{Pearson} - R^2_0}{R^2_{Pearson}}$	k	MAE	Method	RMSE	$R^2_{Pearson}$	$\frac{R^2_{Pearson} - R^2_0}{R^2_{Pearson}}$	k	MAE
Fathead minnow -log(LC50)	RBF	0.72	0.77	0.73	0.14	0.95	0.57	Consensus	0.78	0.72	0.15	0.95	0.53
Daphnia magna -log(LC50)	GPRFVS	0.68	0.91	0.69	0.34	0.97	0.71	Hierarchical	0.90	0.66	0.21	0.97	0.68
Tetrahymena pyriformis -log(IGC50)	GPRFVS	0.88	0.37	0.89	0.04	1.01	0.28	Consensus	0.39	0.89	0.07	1.00	0.29
Rat oral -log(LD50)	RBF	0.64	0.57	0.64	0.03	0.95	0.42	Consensus	0.60	0.60	0.28	0.95	0.44
Bioconcentration Factor	RBF	0.76	0.65	0.76	0.07	0.90	0.46	Consensus	0.78	0.63	0.09	0.89	0.49

Table 3 Summary of StarDrop results for best models of categorical properties and comparison with best T.E.S.T. models.

Property	Best StarDrop Model					Best T.E.S.T. Model				
	Method	Concordance	Sensitivity	Specificity	κ	Method	Concordance	Sensitivity	Specificity	κ
Developmental Tox.	Decision Tree	0.86	0.98	0.59	0.63	Consensus	0.76	0.90	0.41	0.35
Ames Mutagenicity	Decision Tree (pruned)	0.79	0.82	0.77	0.58	Consensus	0.80	0.82	0.78	0.59

Conclusions

The models described in this paper were all built using StarDrop's Auto-Modeller without manual intervention, reinforcing previous examples that illustrated the Auto-Modeller's capability to build and validate models that are comparable with those built 'manually' using other methods (13). The resulting models have acceptable predictive power. However, as for all QSAR models of toxicity, significant uncertainties remain in their predictions and this should be taken into account when making decisions based on their output.; for example, using probabilistic scoring (1) to consider the results in the context of other optimisation objectives.

The models herein conform to the OECD principles for QSAR model validation:

1) The properties modelled correspond to well-defined endpoints
2) The algorithms applied are well defined
3) Each model is associated with a defined domain of applicability
4) Each model is validated against an independent test set using well defined measures of predictive performance
5) The Glowing Molecule output provides a guide to the structural features of a compound that most strongly correlate with the predicted toxicity (although an interpretation of the biological mechanism is not possible)

Finally, from inspection of the toxicity data sets, it is apparent that the majority of the compounds are not 'drug-like'. Therefore, some of the models may be more relevant for application to potential environmental pollutants than in drug discovery; this information is captured by the domain of applicability of the models. However, even where the predictions are highly uncertain, application to potential drugs may nevertheless be useful to reveal the presence of potential toxicophores.

Bibliography

1. *Beyond Profiling: Using ADMET models to guide decisions.* **Segall, M. D., et al.** 2009, Chemistry and Biodiversity, pp. 2144-2151.

2. **US Environmental Protection Agency.** Quantitative Structure Activity Relationship. *Environmental Protection Agency.* [Online] 2011. [Cited: May 20, 2011.] http://www.epa.gov/nrmrl/std/cppb/qsar/#TEST.

3. **OECD.** OECD PRINCIPLES FOR THE VALIDATION, FOR REGULATORY PURPOSES, OF (QUANTITATIVE) STRUCTURE-ACTIVITY RELATIONSHIP MODELS. *www.oecd.org.* [Online] November 2004. [Cited: May 20, 2011.] http://www.oecd.org/dataoecd/33/37/37849783.pdf.

4. **OECD**. GUIDANCE DOCUMENT ON THE VALIDATION OF (QUANTITATIVE)STRUCTURE-ACTIVITY RELATIONSHIPS [(Q)SAR] MODELS. *www.oecd.org.* [Online] March 30, 2007. [Cited: May 20, 2011.] http://www.oecd.org/officialdocuments/displaydocument/?doclanguage=en&cote=env/jm/mono(2007)2.

5. **US Environmental Protection Agency.** User's Guide for T.E.S.T. (version 4.0). *www.epa.org.* [Online] 2011. [Cited: May 20, 2011.] http://www.epa.gov/nrmrl/std/cppb/qsar/testuserguide.pdf.

6. *CAESAR models for developmental toxicity.* **Cassano, A., et al.** Suppl. 1, 2010, Chemistry Central Journal, Vol. 4, p. S4.

7. **Optibrium Ltd.** *StarDrop Reference Guide Version 5.0.* Cambridge : Optibrium Ltd., 2011. Manual.

8. **Wold, S., Sjostrom, M. and Eriksson, L.** Partial Least Squares Projectoins to Latent Structures (PLS) in Chemistry. [book auth.] P von Rague Schleyer, et al. *The Encyclopedia of Computational Chemistry.* Chichester, UK : John Wiley and Sons, 1999, pp. 1-16.

9. *Gaussian processes: a method for automatic QSAR modeling of ADME properties.* **Obrezanova, O, et al.** 2007, J. Chem. Inf. Model., Vol. 47, pp. 1847-1857.

10. **Quinlan, JR.** *C4.5: Programs for Machine Learning.* San Francisco : Morgan Kauffman Publishers, 1993.

11. *Gaussian processes for classification: QSAR modeling of ADMET and target activity.* **Obrezanova, O and Segall, M.D.** 6, 2010, J. Chem. Inf. Model., Vol. 50, pp. 1053-61.

12. *Rational Selection of Training and Test sets for the Development of Validated QSAR Models.* **Golbraikh, A., et al.** 2-4, 2003, Vol. 17, pp. 241-253.

13. *Automatic QSAR modeling of ADME properties: blood-brain barrier penetration and aqueous solubility.* **Obrezanova, O, et al.** 2009, J. Comput. Aided Mol. Des., Vol. 22, pp. 431-440.

Appendix - Regression Plots for Numerical Models

The regression plots are shown for each of the numerical models below. The red line on each is the identity line, i.e. the ideal line for predicted against observed, for comparison.