Inhibition of *Tetrahymena pyriformis* growth by Aliphatic Alcohols and Amines: a QSAR Study

Nadia Ziani, Khadidja Amirat & Djelloul Messadi

Laboratoire de Sécurité Environnementale et Alimentaire (LASEA), Université Badji Mokhtar – Annaba, 23 000, Annaba, Algérie

Soumis le : 07.06.2013 R évisé le : 30.04.2014 Accepté le : 04.06.2014

**Abstract**

A Quantitative Structure- Activity Relationship (QSAR) study was undertaken to evaluate the relative toxicity of a mixed series of 21 alcohols (linear and branched-chain) and 9 normal aliphatic amines in terms of the 50% inhibitory concentration (IGC$_{50}$) of *Tetrahymena pyriformis*. The applied simple linear regression approach is based on theoretical 3D (geometrical) molecular descriptors from DRAGON package, and some calculated logP descriptors. The robustness and the predictive performance of the models were verified using both internal (cross-validation by LOO and LMO; bootstrap) and external statistical validations. ClogP turned out to be the best descriptor to model the considered endpoint. It may be replaced with the geometrical descriptor ADDD in a straightforward manner or with differences in the statistical parameters.

**Keywords:** Alcohols and Amines – Aquatic toxicity – QSAR – Geometrical descriptors and hydrophobic character – Simple linear regression.

---

Auteur correspondant : d_messadi@yahoo.fr

©UBMA - 2014
1. INTRODUCTION

The impact of the potential hazard of untested chemicals, a challenge confronting national and international regulatory agencies [1-4] can be measured by experimental investigations, but this approach is both quite expensive and time consuming [5]. An alternative is to rely on QSAR (Quantitative Structure-Activity Relationships) models that describe a mathematical relationship between the structural features of a set of chemicals and the particular activity associated with them [6,7].

Several QSAR models predicting acute chemical toxicity for aquatic environment have been published [8-12]. They are based mainly on the logarithm of the octanol-water coefficient (logP, also referred to as logKow) as this hydrophobicity term reproduces the ability of a substance to enter cells through the lipid membranes and indicates both the toxicant uptake and baseline toxicity.

Albeit the number of compounds with a measured value for the logP was estimated to be 30 000 [13], which seems at a first glance to be high, this is negligible compared to the rapidly increasing number of compounds for which logP values are desired but missing. Furthermore, the experimental determination is tedious, time-consuming and demands a high purity of the solute [14]; none of these preconditions are compatible with high-throughput techniques, there is, therefore, an ongoing interest in methods for the prediction of logP values.

Over recent decades various approaches (fragmental, atom-based, conformation – dependent methods) [15-18] have been developed that are mostly implemented and available as computer programs. However, even in these calculations it is not uncommon to have differences of several order of magnitude [19, 20].

For these reasons logP cannot be considered a univocal descriptor, which brought different authors [19-24] to propose toxicity models based exclusively on other structural theoretical molecular descriptors.

The present paper proposes predictive simple linear regression QSAR models to evaluate the relative toxicity of organic chemicals, in terms (of the logarithm of the inverse) of the 50% inhibitory growth concentration (IGC50) of Tetrahymena pyriformis. Models based on different kinds of logP (calculated values for AlogP, MlogP and ClogP), are compared to the optimal model constructed using a single 3D (geometrical) descriptor calculated from the chemical structure alone.

2. METHODS

2.1 Experimental Data

Two different toxicants were studied: a set of 21 (linear and branched-chain) alcohols and 9 normal aliphatic amines, selected to reflect diversity in chain length and branching.

These toxicants, which are both nonionic and nonreactive, inhibit the growth concentration of Tetrahymena pyriformis the most tested common freshwater hymenostome ciliate, which approximately measures 50 µm in length and 30 µm in width [25].

The ciliates were grown in axenic culture with population density being measured spectrophotometrically as optical density (absorbance) at 540 nm following 48h of incubation.

The set of experimental data was taken from Schultz [26].

2.2 Estimation of octanol /water partition coefficient

ClogP (≡ calculated logP) [17]

The software version of the system developed by Hansch and Leo [27], using the Rekker’s additive scheme [28], is known as ClogP (or calculated logP). It is based on different fragmental constants and correction terms. Fragment constants were derived from solutes where the fragment occurs in isolation. Furthermore, the bonding environment was taken into account (alkyl, benzyl, vinyl, styryl, and aromatic neighbors) resulting in five values per fragment. If a fragment in combination with the bonding environment was missing but at least two values for the same fragment with different neighbors could be found an interpolation was attempted to derive the missing data. The correction factors have been calculated from the corrections required for the specific interactions being modelled. For instance, the interaction of the two hydroxyl groups in diethylene glycol increases the logP value by 0.85 compared with two hydroxyl.
groups that do not interact. This value is then
taken as the correction term for a two-
neighborred hydroxyl group [29].

The decomposition of the molecular structure
into fragments is performed by using a unique
and simple set of rules, thus obtaining a unique
solution; the fragments are either atoms or
polyatomic groups.

\textbf{AlogP} (≡ Ghose and Cripen model based on
atomic increment system) [30]

Several models have been published where the
fragments are defined on a purely atomic level.
This simplifies both the recognition of
fragments and the calculation, as correction
substructures are not applied (see Eq. (1)).

The most frequently used atomic increment
system, AlogP, was developed by Ghose and
Crippen [31]. Atoms are classified by their
neighboring environment and carbon atoms
additionally by their hybridization.

Estimated logP for any compound is given by:

\begin{equation}
\text{AlogP} = \sum_i n_i a_i
\end{equation}

where \(n_i\) is the occurrence of the \(i\)th atom type
and \(a_i\) is the corresponding hydrophobicity
constant.

The AlogP model implemented in DRAGON
has been evaluated on a set of 2648 compounds
with known experimental logP taken from the
NCI open Data Base. The resulted correlation
coefficient \(r\) is 0.915.

\textbf{MlogP} (≡ Moriguchi model based on
structural parameters)

This is a model described by a regression
equation based on 13 structural parameters [32, 33].

The regression coefficients have been evaluated
by a training set of 1230 organic molecules
including general aliphatic aromatic and
heterocyclic compounds containing the
following atoms: C, H, N, O, S, P, F, Cl, Br, I
[30]. The statistical parameters of the model are
\(r = 0.952;\ SE = 0.422;\ F_{13;1216} = 900.4\)

\subsection{2.4 Chemometric Methods}

Models with one variable were performed
by the software MOBYDYGS [35] using the
Ordinary Least Square regression (OLS)
method.

Population of 74 regression models
corresponding to each of the 74 Geometry
descriptors were ordered according to their
decreasing internal predictive performance,
verified by \(Q^2\), optimal non-logP model was
then selected and compared to the three logP –
based models.

The goodness of fit of the calculated models
were assessed by means of the multiple
determination coefficients, \(R^2\), and the standard
deviation error in calculation (SDEC).

\begin{equation}
S D E C = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}
\end{equation}

Cross validation techniques allow the
assessment of internal predictivity (\(Q^2_{\text{LMO}}\ cross
validation; bootstrap) in addition to the
robustness of model (\(Q^2_{\text{LOO}}\ cross
validation).

Cross validation methods consist in leaving out
a given number of compounds from the training
set and rebuilding the model, which is then used
to predict the compounds left out. This
procedure is repeated for all compounds of the
training set, obtaining a prediction for everyone.
If each compound is taken away one at a time
the cross validation procedure is called leave-
one-out technique (LOO technique), otherwise
leave-more-out technique (LMO technique). An LOO or LMO correlation coefficient, generally indicated with $Q^2$, is computed by evaluating the accuracy of these “test” compounds prediction.

$$Q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{PRESS}{TSS}$$  \hspace{1cm} (3)

The “hat” of the variable $y$, as is the usual statistical notation, indicates that it is a predicted value of the studied property, and the sub index “i/i” indicates that the predicted values come from models built without the predicted compound.

$TSS$ is the total sum of squares.

The predictive residual sum of squares ($PRESS$) measures the dispersion of the predicted values. It is used to define $Q^2$ and the standard deviation error in prediction ($SDEP$).

$$SDEP = \sqrt{PRESS/n}$$  \hspace{1cm} (4)

A value $Q^2 > 0.5$ is generally regarded as a good result and $Q^2 > 0.9$ as excellent [36, 37]. However, studies [38, 39] have indicated that while $Q^2$ is a necessary condition for high predictive power a model, is not sufficient.

To avoid overestimating the predictive power of the model LMO procedure (repeated 5000 times, with 5 objects left out at each step) was also performed ($Q^2_{L(O)}$).

In bootstrap validation technique $K$ n-dimensional groups are generated by a randomly repeated selection of n-objects from the original data set. The model obtained on the first selected objects is used to predict the values for the excluded sample, and then $Q^2$ is calculated for each model. The bootstrapping was repeated 8000 times for each validated model.

By using the selected model the values of the response for the test objects are calculated and the quality of these predictions is defined in terms of $Q^2_{ext}$, which is defined as

$$Q^2_{ext} = 1 - \frac{\sum (y_{iext} - \hat{y}_{iext})^2 / n_{ext}}{\sum (y_i - \bar{y})^2 / n_y} = 1 - \frac{PRESS_{ext}}{TSS_{ext}}$$  \hspace{1cm} (5)

Here $n_{ext}$ and $n_y$ are the number of objects in the external set (or left out by bootstrap) and the number of training set objects, respectively.

The data set randomly was divided into a training set (20 objects) used to develop the QSAR models and a validation set (10 objects), used only for statistical external validation.

Other useful parameters are $R^2$, calculated for the validation chemicals by applying the model developed on the training set, and external standard deviation error of prediction ($SDEP_{ext}$), defined as:

$$SDEP_{ext} = \sqrt{\frac{1}{n_{ext}} \sum_{i=1}^{n_{ext}} (y_i - \hat{y}_i)^2}$$  \hspace{1cm} (6)

where the sum runs over the test set objects ($n_{ext}$).

3. RESULTS AND DISCUSSION

The best one dimensional non-logP model was obtained using the average distance-distance degree (ADDD) index. It encodes information on the molecular folding [40, 41] information about molecular diffusion easiness through biological barriers like membranes.

Distance/distance matrices, denoted as $D/D$, were defined as quotient matrices in terms of geometric $r_{ij}$ distances and topological distance $d_{ij}$:

$$[G/D]_{ij} = \begin{cases} 
\frac{r_{ij}}{d_{ij}} & \text{if } i \neq j \\
0 & \text{if } i = j 
\end{cases}$$  \hspace{1cm} (7)

The row sums of these matrices contain information on the molecular folding; in effect, in highly folded structures, they tend to be relatively small as the inter-atomic distances are small while the topological distances increase as the size of the structure increases.

Therefore, the average row sum is a molecular invariant called average distance-distance degree, that is:

$$ADDD = \frac{1}{A} \sum_{i=1}^{A} \sum_{j=1}^{A} \frac{n_{ij}}{d_{ij}}$$  \hspace{1cm} (8)

A being the number of molecule atoms. Table 1 lists the CAS number, -logIGC$_{50}$, AlogP, MlogP, ClogP and ADDD values of the selected aliphatic alcohols and amines.
Table 1. Relative toxicity and molecular descriptor data for the selected aliphatic alcohols and amines

<table>
<thead>
<tr>
<th>Compound</th>
<th>CAS number</th>
<th>(-\log \text{IGC}_{50})</th>
<th>(\text{AlogP})</th>
<th>(\text{MlogP})</th>
<th>(\text{ClogP})</th>
<th>ADDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methanol</td>
<td>67-56-1</td>
<td>-2.77</td>
<td>-0.358</td>
<td>-0.814</td>
<td>-0.764</td>
<td>4.964</td>
</tr>
<tr>
<td>Ethanol</td>
<td>64-17-5</td>
<td>-2.41</td>
<td>0.009</td>
<td>-0.172</td>
<td>-0.235</td>
<td>8.108</td>
</tr>
<tr>
<td>1-propanol</td>
<td>71-23-8</td>
<td>-1.84</td>
<td>0.515</td>
<td>0.347</td>
<td>0.294</td>
<td>11.098</td>
</tr>
<tr>
<td>1-pentanol</td>
<td>71-41-0</td>
<td>-1.12</td>
<td>1.427</td>
<td>1.209</td>
<td>1.352</td>
<td>17.194</td>
</tr>
<tr>
<td>1-hexanol</td>
<td>111-27-3</td>
<td>-0.47</td>
<td>1.883</td>
<td>1.587</td>
<td>1.881</td>
<td>20.305</td>
</tr>
<tr>
<td>1-heptanol</td>
<td>111-70-6</td>
<td>0.02</td>
<td>2.339</td>
<td>1.940</td>
<td>2.410</td>
<td>23.451</td>
</tr>
<tr>
<td>1-nonanol</td>
<td>143-08-8</td>
<td>0.77</td>
<td>3.252</td>
<td>2.591</td>
<td>3.468</td>
<td>29.859</td>
</tr>
<tr>
<td>1-decanol</td>
<td>112-30-1</td>
<td>1.1</td>
<td>3.708</td>
<td>2.894</td>
<td>3.997</td>
<td>33.118</td>
</tr>
<tr>
<td>1-dodecanol</td>
<td>112-53-8</td>
<td>2.07</td>
<td>4.620</td>
<td>3.467</td>
<td>5.055</td>
<td>39.716</td>
</tr>
<tr>
<td>1-tridecanol</td>
<td>112-70-9</td>
<td>2.28</td>
<td>5.077</td>
<td>3.739</td>
<td>5.584</td>
<td>43.053</td>
</tr>
<tr>
<td>2-propanol</td>
<td>67-63-0</td>
<td>-1.99</td>
<td>0.368</td>
<td>0.347</td>
<td>0.074</td>
<td>10.987</td>
</tr>
<tr>
<td>2-methyl-1-butanol</td>
<td>137-32-6</td>
<td>-1.13</td>
<td>1.290</td>
<td>1.209</td>
<td>1.222</td>
<td>16.810</td>
</tr>
<tr>
<td>3-methyl-1-butanol</td>
<td>123-51-6</td>
<td>-1.13</td>
<td>1.223</td>
<td>1.209</td>
<td>1.222</td>
<td>16.833</td>
</tr>
<tr>
<td>3-methyl-2-butanol</td>
<td>598-75-4</td>
<td>-1.08</td>
<td>1.211</td>
<td>1.209</td>
<td>1.002</td>
<td>16.622</td>
</tr>
<tr>
<td>(tert)pentanol</td>
<td>75-85-4</td>
<td>-1.27</td>
<td>1.097</td>
<td>1.209</td>
<td>1.002</td>
<td>16.640</td>
</tr>
<tr>
<td>1-propylamine</td>
<td>107-10-8</td>
<td>-0.85</td>
<td>0.225</td>
<td>0.347</td>
<td>0.394</td>
<td>11.968</td>
</tr>
<tr>
<td>1-hexylamine</td>
<td>11-26-2</td>
<td>-0.34</td>
<td>1.594</td>
<td>1.587</td>
<td>1.981</td>
<td>21.226</td>
</tr>
<tr>
<td>1-heptylamine</td>
<td>111-68-2</td>
<td>0.1</td>
<td>2.050</td>
<td>1.940</td>
<td>2.510</td>
<td>24.393</td>
</tr>
<tr>
<td>1-octylamine</td>
<td>111-86-4</td>
<td>0.51</td>
<td>2.506</td>
<td>2.274</td>
<td>3.039</td>
<td>27.602</td>
</tr>
<tr>
<td>1-undecylamine</td>
<td>7307-55-3</td>
<td>2.26</td>
<td>3.875</td>
<td>3.186</td>
<td>4.626</td>
<td>37.408</td>
</tr>
<tr>
<td>1-butanol*</td>
<td>71-36-3</td>
<td>-1.52</td>
<td>0.971</td>
<td>0.800</td>
<td>0.823</td>
<td>14.138</td>
</tr>
<tr>
<td>1-totanol*</td>
<td>111-87-5</td>
<td>0.5</td>
<td>2.796</td>
<td>2.274</td>
<td>2.939</td>
<td>26.640</td>
</tr>
<tr>
<td>1-undecanol*</td>
<td>112-42-5</td>
<td>1.87</td>
<td>4.164</td>
<td>3.186</td>
<td>4.526</td>
<td>36.401</td>
</tr>
<tr>
<td>2-pentanol*</td>
<td>6032-29-7</td>
<td>-1.25</td>
<td>1.348</td>
<td>1.209</td>
<td>1.132</td>
<td>16.888</td>
</tr>
<tr>
<td>3-pentanol*</td>
<td>584-02-1</td>
<td>-1.33</td>
<td>1.416</td>
<td>1.209</td>
<td>1.132</td>
<td>16.908</td>
</tr>
<tr>
<td>(neo)pentanol*</td>
<td>75-84-3</td>
<td>-0.96</td>
<td>1.108</td>
<td>1.209</td>
<td>1.092</td>
<td>16.710</td>
</tr>
<tr>
<td>1-butylamine*</td>
<td>109-73-9</td>
<td>-0.7</td>
<td>0.681</td>
<td>0.800</td>
<td>0.923</td>
<td>15.015</td>
</tr>
<tr>
<td>1-amylamine*</td>
<td>110-58-7</td>
<td>-0.61</td>
<td>1.137</td>
<td>1.209</td>
<td>1.452</td>
<td>18.101</td>
</tr>
<tr>
<td>1-nonylamine*</td>
<td>112-20-9</td>
<td>1.59</td>
<td>2.962</td>
<td>2.591</td>
<td>3.568</td>
<td>30.837</td>
</tr>
<tr>
<td>1-decylamine*</td>
<td>2016-57-0</td>
<td>1.95</td>
<td>3.418</td>
<td>2.894</td>
<td>4.097</td>
<td>34.110</td>
</tr>
</tbody>
</table>

(a): Chemical Abstract Services registry number ; (*) validation set compound.

Intercepts (\(\beta_0\)) and slopes (\(\beta_1\)) of the calculated one dimensional models are shown in table 2

Table 2. Coefficients for the ordinary least squares calculated models.

<table>
<thead>
<tr>
<th>X</th>
<th>AlogP</th>
<th>MlogP</th>
<th>ClogP</th>
<th>ADDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\beta_0)</td>
<td>-2.144 (±0.139)</td>
<td>-2.230 (±0.134)</td>
<td>-1.998 (±0.088)</td>
<td>-3.333 (±0.134)</td>
</tr>
<tr>
<td>(\beta_1)</td>
<td>0.939 (±0.057)</td>
<td>1.191 (±0.068)</td>
<td>0.815 (±0.033)</td>
<td>0.138 (±0.006)</td>
</tr>
</tbody>
</table>

©UBMA - 2014
Relevant statistical parameters reported in table 3 below clearly show the difference in fitting and prediction performances for the selected logP descriptors, ClogP appears as the best ones.

Another remarkable fact observed is that ClogP and ADDD theoretical molecular descriptor can be interchanged without relevant variations in the statistical results.

Table 3. Summary statistics for the one dimensional calculated models.

<table>
<thead>
<tr>
<th>X</th>
<th>R²</th>
<th>Q²</th>
<th>Q²_{L(5)O}</th>
<th>Q²_{boot}</th>
<th>Q²_{ext}</th>
<th>SDEC</th>
<th>SDEP</th>
<th>SDEP_{ext}</th>
<th>F_{(p=0.000)}</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alog P</td>
<td>0.9372</td>
<td>0.9188</td>
<td>0.9165</td>
<td>0.9063</td>
<td>0.8584</td>
<td>0.365</td>
<td>0.415</td>
<td>0.554</td>
<td>268.4</td>
<td>0.385</td>
</tr>
<tr>
<td>Mlog P</td>
<td>0.9444</td>
<td>0.9291</td>
<td>0.9270</td>
<td>0.9160</td>
<td>0.8998</td>
<td>0.344</td>
<td>0.388</td>
<td>0.461</td>
<td>305.52</td>
<td>0.362</td>
</tr>
<tr>
<td>Clog P</td>
<td>0.9713</td>
<td>0.9630</td>
<td>0.9621</td>
<td>0.9580</td>
<td>0.9352</td>
<td>0.247</td>
<td>0.280</td>
<td>0.371</td>
<td>610.02</td>
<td>0.260</td>
</tr>
<tr>
<td>ADDD</td>
<td>0.9711</td>
<td>0.9624</td>
<td>0.9616</td>
<td>0.9566</td>
<td>0.9326</td>
<td>0.248</td>
<td>0.283</td>
<td>0.378</td>
<td>603.82</td>
<td>0.261</td>
</tr>
</tbody>
</table>

Carbó-Dorca et al. [42] reported a QSAR study where the same data was examined, these authors constructed a predictive model using, as a molecular descriptor, the expectation value of the inter electronic repulsion energy operator presented as a kind of quantum self-similarity measure (QS-SM). The correlation results reached $R^2 = 0.9240$, $Q^2 = 0.9090$ and $SE = 0.415$, which are inferior than the present approach.

The value of $R^2$ attests the good fitting performances of the model. In general, the larger the magnitude of the F ratio, the better the model predicts the property values in the training set. The large F ratio of 603.82 indicates that the model does an excellent job of predicting the $-\log IGC_{50}$ values. The model is robust, the difference between $R^2$ and $Q^2$ is small (<1%). Figure 1 shows a plot contrasting experimental and cross-validated $-\log IGC_{50}$.

The point dispersion is small, although there is one point a little bit far away from the rest (1-propylamine). SDEP is similar to SDEC, so this model has internal predictivity not so dissimilar from fitting power.

The model demonstrates a very good stability in internal validation (difference between $Q^2$ and $Q^2_{L(5)O}$ is 0.33%), while bootstrapping confirms the internal predictivity and stability of the model.

Though small sized the data set underwent statistical validation by preliminary random splitting of the chemicals into training (20 chemicals) and validation (10 chemicals) sets. The small size of the published experimental data set [26] did not allow a more drastic splitting. The information obtained by $Q^2_{ext}$ is somewhat optimistic. In fact with small data sets (20-30 chemicals), completely new chemicals external predictivity can only be verified a posteriori, case-by-case.

Figure-1 Experimental versus cross-validation activity for the training set objects.

4. CONCLUSION

From the results and discussion above we conclude that:

1. Among the logP descriptors selected to model the inhibition of *Tetrahymena pyriformis*

©UBMA - 2014
growth by aliphatic alcohols and amines, Clog P is the best.  
2. Geometrical descriptor ADDD (Average Distance/Distance Degree), a theoretical molecular descriptor, and Clog P can be interchanged without relevant variations in the statistical results.
3. The non-logP model obtained in this study has very good fitting performances, is robust and with acceptable predictive power. The internal validation parameters (QLOO, QLSO and bootstrap) are similar to the fitting parameters.
Notice that chemical 19 (1-propyl amine) with cross validated standardized residuals (not reported here) greater than 3 standard deviation units, is an heavy outlier in all the models considered in this paper.

REFERENCES


[35] MOBYDIGS – Models BY Descriptors In Genetic Selection – ver. 1.1 for Windows, Talete s.r.l., Milano, Italy.


