CROATICA CHEMICA ACTA
CCACAA, ISSN-0011-1643, ISSN-1334-417X
Croat. Chem. Acta **82** (1) (2009) 311-316.
CCA-3321

Original Scientific Paper

Predictions of Toxicity to Chlorella vulgaris and the Use of Momentum-space Descriptors*

Jabir H. A. Al-Fahemi,[a], David L. Cooper,[a],*** and Neil L. Allan[b]**

[a]*Department of Chemistry, University of Liverpool, Liverpool L69 7ZD, UK*
[b]*School of Chemistry, University of Bristol, Cantock's Close, Bristol BS8 1TS, UK*

RECEIVED JUNE 2, 2008; REVISED AUGUST 27, 2008; ACCEPTED SEPTEMBER 3, 2008

Abstract. The toxicity to Chlorella vulgaris, expressed as log(1/EC$_{50}$), of two sets of aromatic compounds has been examined. For the first set, which consists of 13 mono- and di-substituted nitrobenzenes, it was found that one- or two-descriptor models provide useful correlations. A six-descriptor regression model for the log(1/EC$_{50}$) values of a larger set that consists of 49 aromatic compounds has also been developed. Overall, it was found that a combination of a small number of trivial 'feature count' classical descriptors (numbers of atoms of a particular type) and less traditional quantities (entropy-like momentum-space descriptors) have potential benefits for useful QSAR models.

Keywords: toxicity, Chlorella vulgaris, momentum-space descriptors, QSAR

INTRODUCTION

Determining the toxic effects of chemicals in natural fresh waters is of particular significance because of the continuing widespread release of industrial, agricultural and domestic chemicals into such environments. Indeed, legislation is increasingly requiring detailed assessment of the impact of such chemicals on a wide range of aquatic life, including unicellular organisms, invertebrates and fish. Algae are of particular importance for such studies not only because of their widespread distribution and their fundamental importance to aquatic ecosystems, but also because they can react rapidly to environmental change, not least because of a short life cycle. Given that algal tests have traditionally tended to be both time-consuming and expensive, there has been significant interest in developing reliable short-term toxicity assays. Of particular relevance to the present computational study is the experimental work of Cronin and co-workers[1–5] who have demonstrated the utility of a novel procedure for the rapid testing of toxicity to the unicellular green alga Chlorella vulgaris.

The experimental technique of Cronin and co-workers[1–5] relies on the presence in living organisms, including Chlorella vulgaris, of nonspecific esterases which are able to hydrolyze non-fluorescent fluorescein diacetate to the fluorescent compound fluorescein. In essence, algae are exposed for 15 minutes to the chemical being tested and the measurements of the resulting fluorescence are compared to those for an appropriate control. Concentrating on a small number of mechanistically interpretable descriptors, such as measures of hydrophobicity, electrophilicity and molecular size, Cronin and co-workers have also developed QSARs for their various data.[1–5]

It has been established in previous work by ourselves and by others that a number of momentum-space quantities are useful as molecular descriptors for quantum molecular similarity studies,[6] as well as for QSAR and QSPR applications.[7–11] Accordingly, we present here an exploratory study of their potential utility also for the prediction of toxicity to Chlorella vulgaris, thereby also complementing the various QSAR studies of Cronin and co-workers.[1–5] We examine first the set of 13 mono- and di-substituted nitrobenzenes that they considered in Ref. 1 and then a larger set that includes the same group of 13 nitrobenzenes as well as the 14 anilines that they considered in Ref. 5 plus a range of other aromatic species.[4]

METHODOLOGY

AM1 geometry optimizations were carried out for each molecule[12] and the momentum-space (p-space) total electron density, $\rho(\boldsymbol{p})$, was calculated from the Fourier

* Dedicated to Professor Zvonimir Maksić on the occasion of his 70th birthday.
** Current affiliation: Umm Al-Qura University, Makka, Saudi Arabia.
*** Author to whom correspondence should be addressed. (E-mail: dlc@liverpool.ac.uk)

transform of the resulting wave function.[13] Families of p-space descriptors which we have successfully used previously[9–11] include moments of momentum, $\langle p^m \rangle$, defined by

$$\left\langle p^m \right\rangle = \int p^m \rho\left(\boldsymbol{p}\right) \mathrm{d}\boldsymbol{p}$$

as well as entropy-like quantities, defined as

$$S_m = -\int p^m \rho\left(\boldsymbol{p}\right) \ln \rho\left(\boldsymbol{p}\right) \mathrm{d}\boldsymbol{p}$$

and

$$\begin{aligned} \acute{S}_m &= -\int p^m \sigma\left(\boldsymbol{p}\right) \ln \sigma\left(\boldsymbol{p}\right) \mathrm{d}\boldsymbol{p} \\ &\equiv \left(\left\langle p^m \right\rangle \ln N + S_m\right) / N \end{aligned}$$

in which $\sigma(\boldsymbol{p}) = \rho(\boldsymbol{p})/N$ is the so-called shape function and N is the total number of electrons explicitly treated in the AM1 calculations. The entropy-like descriptors S_m and $\acute{S}_m$ (with m values of −2, 0 and +2) have previously proved especially useful in studies of octanol/water partition coefficients[9] and blood-brain barrier penetration.[11] For the present study, we considered a total of eight quantum-mechanical descriptors: $\langle p^{-2} \rangle$, $\langle p^2 \rangle$, and values of S_m and $\acute{S}_m$ with m values of −2, 0 and +2, but only the entropy-like quantities turned out to be useful. Other descriptors considered in the present work, as in some of our earlier studies, are the relative molecular mass (M_r) and entirely trivial 'feature count' structural quantities, namely the numbers of atoms of a particular type (n_X) and the number of covalent bonds (n_bond).

Multiple linear regression (MLR) models in the present study were constructed using SPSS[14] using the 'simultaneous method' (which SPSS calls the 'enter method') in which descriptors are automatically rejected from the full set if they are too strongly correlated with the others. Measures of the success of such MLR models include the value of the adjusted correlation coefficient, R^2(adj.), which takes account of the number of observations and the number of variables, as well as the standard error (Δ) and the 'analysis of variance' or Fisher-F statistic, which should be large. As a further indication of the statistical significance of a given MLR correlation we report values of a standard statistical measure which is known as the significance of the Fisher-F statistic and which we denote p_F. The value of this quantity, which should be very small for a statistically significant model, can be considered a measure of the probability that an apparently good correlation has arisen by chance. Qualitatively, p_F can be thought of as the likelihood that random sorting of the experimental log(1/EC$_{50}$) values, but with the values of the molecular descriptors left in the original order, leads to a better correlation.

For a particular descriptor to be considered a useful predictor in the final MLR model, conventional guidelines are that the coefficient for that descriptor divided by its standard error (known as its t value) should lie outside the range −2 to +2. Statistical outliers in our models were identified as molecules with absolute standardized residuals greater than 2, as listed in the table of Casewise Diagnostics in SPSS.[14]

We also consider an alternative method of model construction, known as the 'stepwise method', in which SPSS considers each of the chosen descriptors in turn. Provided that a descriptor contributes to the success of the model then it is included, but all of the other descriptors that are currently included are then reassessed, to determine whether they are still successfully contributing to the model and therefore should be retained. This procedure often yields very small sets of predictor variables, but it is well known that it does not always guarantee the best QSAR model.[15,16]

RESULTS

Series A − Nitrobenzenes

Cronin and co-workers[1] have used their novel 15-minute assay to determine the toxicities of 13 mono- and disubstituted nitrobenzenes to Chlorella vulgaris and have then constructed QSAR models using as descriptors octanol/water partition coefficients, log P, and the energy of the lowest unoccupied molecular orbital, E_LUMO. They reported an R^2(adj.) value of 0.767, standard error (Δ) of 0.442 and F-statistic value of 20.8. Omitting 4-chloronitrobenzene as an outlier with a high standardized residual, they achieved a better fit, with R^2(adj.) = 0.861, Δ = 0.353 and F = 35.2. They noted that this molecule is the only one in their data set that is substituted at the 4-position and they commented that such systems are known to be more difficult to model.

We have examined the same log(1/EC$_{50}$) data for the 13 molecules considered in Ref. 1. Starting from a pool of descriptors that consists of values of $\langle p^{-2} \rangle$, $\langle p^2 \rangle$, and values of S_m and $\acute{S}_m$ (with m values of −2, 0 and +2), as well as the relative molecular mass (M_r) and the various trivial 'feature count' descriptors (n_X and n_bond),[17] the following one-descriptor regression model was found to be the preferred model using the 'stepwise' method:

$$\log\left(1/\mathrm{EC}_{50}\right) = 0.018 M_\mathrm{r} - 2.442$$

This model is characterized by n = 13, R^2(adj.) = 0.723, Δ = 0.482, F = 32.3 and p_F = 1.4×10^{-4}. The t value for the M_r descriptor is 5.686. No outliers were observed. It is true for this full set of 13 structurally-similar molecules that the corresponding two-descriptor results obtained by Cronin *et al.*[1] are slightly better, when judged by the values of R^2(adj.) and Δ, but the present remarkably simple correlation is statistically more significant, according to the value of F.

We have also constructed several regression models using the 'enter' method, again starting from our standard pool of p-space and conventional molecular descriptors.[17] Our preferred model using this method is:

$$\log\left(1/\mathrm{EC}_{50}\right) = -5.027\acute{S}_0 - 6.494\acute{S}_{-2} + 44.735$$

which is characterized by n = 13, R^2(adj.) = 0.731, Δ = 0.475, F = 17.3 and p_F = 0.001. The t values for $\acute{S}_0$ and $\acute{S}_{-2}$ are −4.149 and −5.877, respectively. This two-descriptor model is better, as judged only by the values of R^2(adj.) and Δ than was the simple one based on relative molecular mass, but it is statistically less significant (according to the values of F and p_F). Clearly p-space descriptors provide little, if any, advantage, when predicting the toxicity of these nitrobenzenes.

Series B − Aromatic Compounds

Netzeva *et al.*[4] have determined experimentally the toxicity to Chlorella vulgaris of 65 aromatic compounds, including phenols, anilines, nitrobenzenes, benzaldehydes and other poly-substituted benzenes. Their MLR model based on log P and E_{LUMO} is characterized by an R^2(adj.) value of 0.839, a standard error of 0.429 and an F value of 161 for the full set of 65 molecules. They also used partial least squares analysis to develop a more sophisticated four-descriptor model by means of stepwise elimination of variables from a set of 102 calculated descriptors; they reported R^2(adj.) = 0.86 and a root mean square error (RMSE) of 0.40.

We consider here a set of 50 aromatic compounds, which we call Series B. It includes all of the molecules from Series A plus the 14 anilines that were considered in Ref. 5 plus a selection of other aromatic species,[4] including phenols, benzaldehydes and other poly-substituted benzenes. All of the experimental log(1/EC$_{50}$) data were taken from Ref. 4; there are no inconsistencies for the nitrobenzenes and anilines with the corresponding data in Ref. 1 and in Ref. 5, respectively, but some of the molecules were named differently. Our Series B consists of the 49 molecules listed in Table 1, plus 4-iodophenol, for which the experimental value of log(1/EC$_{50}$) is 0.16.[4]

In building our various regression models for Series B, we started from the same pool of p-space and conventional molecular descriptors as we considered for Series A.[17] Nonetheless, we found again that our preferred model using the 'stepwise' method for the full set of 50 molecules is again based only on the relative molecular mass:

$$\log\left(1/\mathrm{EC}_{50}\right) = 0.017M_{\mathrm{r}} - 2.548$$

This remarkably simple model is characterized by n = 50, R^2(adj.) = 0.683, Δ = 0.535, F = 107 and p_F = 8.6×10^{-14}. The t value for M_r is 10.332. There are no outliers. It is useful to note that the coefficients in this one-descriptor model are very similar to those in the corresponding one-descriptor model for Series A. Direct comparison with the models of Netzeva *et al.*[4] would be inappropriate, on account of the different sets of molecules and the different numbers of descriptors, but it is nonetheless clear that our one-descriptor model is statistically significant. At the suggestion of a referee, we also examined a one-descriptor model based only on $\log P$. For this purpose we used the values of $\log K_{\mathrm{ow}}$ that were tabulated in Ref. 4. The resulting model, which is characterized by an R^2(adj.) value of 0.519, a standard error of 0.660, an F value of 54 and p_F = 2.2×10^{-9} for the full set of 50 molecules, is clearly somewhat inferior to the one that is based only on the relative molecular mass.

Our preferred MLR model using the 'enter' method, again starting from our standard pool of p-space and conventional molecular descriptors,[17] turns out to be:

$$\begin{aligned}&\log\left(1/\mathrm{EC50}\right) = \\ &35.220 + 0.301S_0 - 8.462\acute{S}_0 - 0.139S_{-2} \\ &-3.109\acute{S}_{-2} - 0.305n_{\mathrm{H}} + 1.088n_{\mathrm{Cl}} - 2.723n_{\mathrm{I}}\end{aligned}$$

This seven-descriptor MLR model is characterized by n = 50, R^2(adj.) = 0.905, Δ = 0.293, F = 67.8 and p_F = 7.8×10^{-21}. The t values for S_0, $\acute{S}_0$, S_{-2}, $\acute{S}_{-2}$, n_{H}, n_{Cl} and n_{I} are 5.543, −4.518, −6.874, −2.417, −6.077, 6.248 and −4.189, respectively. Except for a decrease in the value of F, which is still comfortably large, this model is clearly performing much better than is the one based only on the relative molecular mass. Excluding six molecules (2,4-dinitroaniline, 2,6-dichlorobenzaldehyde, 4-methoxyphenol, 1,2-dinitrobenzene, 1-chloro-4-nitrobenzene and 2,6-dichlorobenzaldehyde) which have absolute standardized residuals greater than 2 does of course enhance the statistical quality of the correlation: n = 44, R^2(adj.) = 0.963, Δ = 0.179, F = 161 and p_F = 3.2×10^{-25}. The presence of such outliers could indicate

Table 1. Observed values of acute algal toxicity for 49 molecules of Series B compared to the predictions using the six-descriptor regression model. These toxicity values to Chlorella vulgaris are based on millimolar concentrations, *i.e.* concentrations expressed in millimoles per litre. The 13 nitrobenzenes that are also members of Series A have been identified with the symbols (A)

Molecule		$\log(1/EC_{50})$ (observed)	$\log(1/EC_{50})$ (predicted)	Molecule		$\log(1/EC_{50})$ (observed)	$\log(1/EC_{50})$ (predicted)
phenol		−1.46	−1.47	3,5-dichloroaniline		0.24	0.41
aniline		−1.34	−1.43	2,4,6-trimethylnitrobenzene	(A)	0.25	0.25
2-fluorophenol		−1.08	−1.06	2,6-dichloroaniline		0.26	0.37
2-fluoroaniline		−1.05	−1.22	1,2-dichlorobenzene		0.37	0.64
3-cresol		−1.01	−0.94	1,3-dinitrobenzene	(A)	0.38	0.47
4-methoxyphenol		−0.97	−0.53	2,4-dinitrophenol		0.40	0.62
2-hydroxyaniline		−0.91	−1.37	1,4-dinitrobenzene	(A)	0.41	0.45
2-methoxyphenol		−0.88	−0.49	3-nitrobenzaldehyde		0.45	0.31
2,6-dimethylaniline		−0.87	−0.77	2,6-dichloro-4-nitroaniline		0.64	0.86
benzaldehyde		−0.81	−0.72	2,4-dinitrotoluene	(A)	0.70	0.54
2-cresol		−0.81	−0.92	6-chloro-2,4-dinitroaniline		0.80	0.75
2-hydroxybenzaldehyde		−0.80	−0.58	2,6-dibromo-4-nitrophenol		0.81	0.91
nitrobenzene	(A)	−0.78	−0.61	2,5-dichloronitrobenzene	(A)	0.97	1.29
4-cresol		−0.66	−0.92	2,4,6-trichloroaniline		1.11	1.01
3,4-dimethylphenol		−0.65	−0.57	2-chloro-6-nitrotoluene	(A)	1.17	0.50
3-nitrotoluene	(A)	−0.50	−0.25	4-chloro-2,6-dinitroaniline		1.19	0.79
4-chlorophenol		−0.42	−0.27	1,2-dinitrobenzene	(A)	1.23	0.46
2,4-dinitroaniline		−0.36	0.29	1-chloro-4-nitrobenzene	(A)	1.25	0.31
4-bromophenol		−0.35	−0.34	2,3,5,6-tetrachloroaniline		1.48	1.51
4-bromoaniline		−0.33	−0.41	2,4-dichloro-6-nitrophenol		1.50	1.33
3-chloroaniline		−0.31	−0.44	2,6-dichlorobenzaldehyde		1.50	0.95
3,5-dinitroaniline		0.03	0.23	2,4,5-trichloronitrobenzene	(A)	1.88	1.97
2-chlorobenzaldehyde		0.06	0.31	2,4,6-trichloro-1,3-dinitrobenzene	(A)	1.89	2.05
4-ethylbenzaldehyde		0.16	0.16	2,3,5,6-tetrachloronitrobenzene	(A)	2.34	2.30
2-isopropylphenol		0.17	−0.05				

that some of these molecules act by a different mechanism than do the others.

It is clear that *p*-space descriptors are much more useful when dealing with the greater molecular diversity present in Series B than was the case for Series A. Nonetheless, as an internal test of the seven-descriptor model, we used regression parameters for the 37 molecules that are not included in Series A to model $\log(1/EC_{50})$ for the 13 nitrobenzenes of Series A. The results are relatively good, with $R^2 = 0.818$ and RMSE = 0.280.

Subsequent analysis of the seven molecular descriptors that have been retained in the preferred model for the full set of 50 molecules revealed an oversight in the original selection of Series B. It turns out that just one iodine-containing molecule had been included. As such, the appearance in this MLR model of the value of n_I is entirely artificial, given that the coefficient that

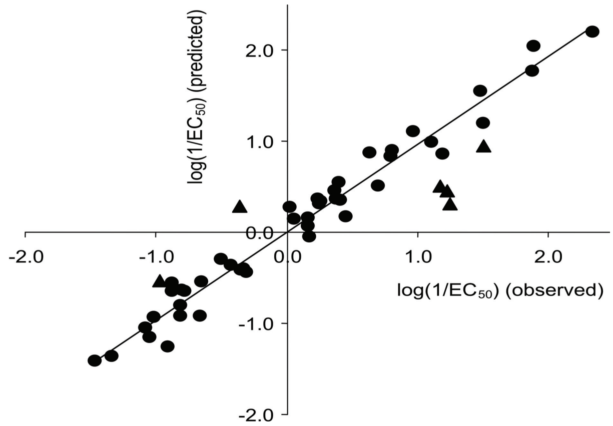

Figure 1. Predicted and observed values of algal toxicity for 49 molecules of Series B. The predictions of toxicity to Chlorella vulgaris were made using the six-descriptor MLR model and all of the experimental values of $\log(1/EC_{50})$ were taken from Ref. 4. Six outliers are represented as triangles.

multiplies this trivial 'feature count' descriptor must have been solely determined by the requirement to match exactly the experimental log(1/EC_{50}) value for 4-iodophenol. Realistically, we should ignore the presence of this molecule in Series B and drop from the seven-descriptor model the term that involves n_I. For the remaining 49 molecules, the observed values of algal toxicity expressed as log(1/EC_{50}) are compared numerically in Table 1 to the predictions from our six-descriptor regression model (without redetermining the coefficients from the MLR 'fit' for all 50 molecules). Additionally, the correlation between the predicted and observed values of 15-minute algal toxicity is shown in Figure 1.

CONCLUSIONS

We have examined the toxicity to Chlorella vulgaris, expressed as log(1/EC_{50}), for two sets of aromatic compounds. For the first set, Series A, which consists of 13 mono- and di-substituted nitrobenzenes, we found that one- or two-descriptor models provide useful correlations. We have also developed a seven-descriptor regression model for the log(1/EC_{50}) values of a set of 50 aromatic compounds, but subsequently reduced it to a six-descriptor model for 49 aromatic systems, after exclusion of the single iodine-containing compound. Overall, we have demonstrated that a combination of a small number of trivial 'feature count' classical descriptors (numbers of atoms of a particular type) and less traditional quantities (entropy-like momentum-space descriptors) have potential benefits for useful QSAR models which incorporate only a small number of parameters. On the other hand, an obvious drawback of using such momentum-space descriptors is that unlike (say) values of log P or E_{LUMO} they cannot currently be understood in terms of specific structural features or the mechanisms of action.

REFERENCES

1. M. T. D Cronin, J. C. Dearden, J. C. Duffy, R. Edwards, N. Manga, A. P. Worth, and A. D. P Worgan, *SAR QSAR Environ. Res.* **13** (2002) 167−176.
2. A. D. P Worgan, J. C Dearden, R. Edwards, T. I. Netzeva, and M. T. D Cronin, *QSAR Comb. Sci.* **22** (2003) 204−209.
3. M. T. D Cronin, T. I. Netzeva, J. C. Dearden, R. Edwards, and A. D. P. Worgan, *Chem. Res. Toxicol.* **17** (2004) 545−554.
4. T. I. Netzeva, J. C. Dearden, R. Edwards, A. D. P Worgan, and M. T. D. Cronin, *J. Chem. Inf. Comput. Sci.* **44** (2004) 258−265.
5. T. I. Netzeva, J. C. Dearden, R. Edwards, A. D. P. Worgan, and M. T. D. Cronin, *Bull. Environ. Contam. Toxicol.* **73** (2004) 385−391.
6. For example, P. T. Measures, K. A. Mort, N. L. Allan, and D. L. Cooper, *J. Comput. Aided Mol. Des.* **9** (1995) 331−340.
7. E. F. McCoy and M. J. Sykes, *Chem. Phys. Lett.* **313** (1999) 707−712.
8. E. F. McCoy and M. J. Sykes, *J. Chem. Inf. Comput. Sci.* **43** (2003) 545−553.
9. J.H. Al-Fahemi, D. L. Cooper, and N. L. Allan, *J. Mol. Struct. (THEOCHEM)* **727** (2005) 57−61.
10. J. H. Al-Fahemi, D. L. Cooper, and N. L. Allan, *Chem. Phys. Lett.* **416** (2005) 376−380.
11. J. H. Al-Fahemi, D. L. Cooper, and N. L. Allan, *J. Mol. Graphics. Modell.* **26** (2007) 607−612.
12. We used analytic gradients (ANALYT), a gradient tolerance (GRATOL) of 0.0015 and, in the SCF calculations, a tight convergence criterion (SCFCRT) of 1×10^{-10}.
13. For example, N. L. Allan and D. L. Cooper, *Top. Curr. Chem.* **173** (1995) 85−111.
14. *SPSS Base 10.0 Applications Guide*, SPSS Inc., Chicago IL, USA, 1999.
15. J. G. Topliss and R. P. Edwards, *J. Med. Chem.* **22** (1979) 1238−1244.
16. For example, A. R. Katritzky, V. S. Lobanov, and M. Karelson, *Chem. Soc. Rev.* **24** (1995) 279−287.
17. J. H. A. Al-Fahemi *Momentum-space descriptors for QSPR and QSAR studies*, PhD Thesis, Liverpool University, UK, 2006.

SAŽETAK

Predviđanje toksičnosti Chlorella vulgaris i korištenje impuls-prostor deskriptora§

Jabir H. A. Al-Fahemi,[a] David L. Cooper[a] i Neil L. Allan[b]

[a]*Department of Chemistry, University of Liverpool, Liverpool L69 7ZD, UK*
[b]*School of Chemistry, University of Bristol, Cantock's Close, Bristol BS8 1TS, UK*

Istraživana je toksičnost alge Chlorella vulgaris, izražena kao log(1/EC$_{50}$), dva seta aromatskih spojeva. Za prvi set, koji se sastoji od 13 mono- i di-supstituiranih nitrobenzena, nađeno je da jedno- ili dvo-deskriptorski modeli daju korisne korelacije. Također razvijen je šest-deskriptorski regresijski model za log(1/EC$_{50}$) vrijednosti za veći set koji se sastoji od 49 aromatskih spojeva. Pronađeno je da kombinacija malog broja trivijalnih prebrojivih (engl. *feature count*) klasičnih deskriptora (broja atoma određenog tipa) i manje tradicionalnih vrijednosti (entropijski impuls-prostor (engl. *momentum-space*) deskriptori) imaju potencijalne pozitivne korisne strane za upotrebljavane QSAR modele.

§ impuls-prostor deskriptori (engl. *momentum-space descriptors*)