Available online at www.sciencedirect.com

SCIENCE DIRECT

Bioorganic & Medicinal Chemistry Letters 16 (2006) 1941–1943

Bioorganic & Medicinal Chemistry Letters

QSAR models of quail dietary toxicity based on the graph of atomic orbitals

Andrey A. Toropov* and Emilio Benfenati

Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche 'Mario Negri', Via Eritrea 62, 20157 Milan, Italy

Received 25 November 2005; revised 21 December 2005; accepted 21 December 2005
Available online 25 January 2006

Abstract—Graphs of atomic orbitals (GAOs) have been used to represent molecular structures. We describe rules to convert the labelled hydrogen-filled graphs (LHFGs) into GAOs. The GAO is one possible way of taking account of the structure of atoms (i.e., atomic orbitals, such as $1s^1$, $2p^2$ and $3d^{10}$) for QSPR/QSAR analyses. Optimization of correlation weights of local invariants (OCWLI) of the LHFGs and the GAOs was used to obtain a method of quail dietary toxicity modelling. Statistical characteristics of the models based on the OCWLI of GAO are better than those based on the OCWLI of the LHFGs.

Prediction of the physicochemical and biological parameters of a substance from an analysis of its molecular structure and the similar data from experimentally investigated substances is an attractive alternative to obtaining such information by direct experiments. Searching algorithms for this prediction is an important part of theoretical chemistry. Usually this can be done by quantitative structure-property/activity relationships (QSPR/QSAR). The present study compared QSAR models of quail dietary toxicity based on labelled hydrogen-filled graphs (LHFGs)[1–5] with models of the endpoint based on graphs of atomic orbitals (GAOs)[6–8]. To take account of the structure of atoms in QSPR/QSAR analysis (i.e., presence of the $1s^1$, $2s^2, \ldots, 3p^6 \ldots$ orbitals), we used GAO. As a tool for building QSAR models we used optimization of correlation weights of local graph invariants.[1–8]

We used pesticide data for the quail toxicity dietary exposure, taking LC_{50}-96 h which is the dose/concentration that kills 50% of the animals in 96 h. The decimal logarithm $\log(1/C)$ was used, where C is the concentration in mmol/L. Toxicity data from the U.S. Environment Protection Agency Office of Pesticide Programs (EPA-OPP) were kindly provided by Dr. B. Montague. Toxicity values were collected and compared with the values from the BBA database,[9] as described in Ref. 10, in order to use the most reliable data in case of multiple values and to eliminate compounds when there was disagreement between values; if the toxicity values varied by more than a factor of 4, we eliminated the compounds. We obtained a data set of 110 pesticides. The set was sorted on the basis of toxicity values, and randomly split into a training ($n = 91$) and a test set ($n = 19$), extracting one molecule out of six for the test set.

The QSARs under consideration are based on descriptors calculated as

$$
\begin{aligned}
{}^{0}\mathrm{X}_{\mathrm{cw}} &= {}^{0}\mathrm{X}_{\mathrm{cw}}(v, {}^{X}\mathrm{EC}) \\
&= \left\{ \left[\prod_{k=1}^{n} \mathrm{CW}(v_k) \times \prod_{k=1}^{n} \mathrm{CW}({}^{X}\mathrm{EC}_k) \right] - 1 \right\} \times 100,
\end{aligned}
\tag{1}
$$

where $\mathrm{CW}(v_k)$ is the correlation weight related to the molecular graph for a given vertex (atom, a, for the LHFG or atomic orbital, ao, for the GAO); $\mathrm{CW}({}^{X}\mathrm{EC}_k)$ is the correlation weight related to the presence in the molecular graph for a given numerical value of the Morgan extended connectivity of Xth order ($X = 0$, 1, 2 and 3);[1,2] n is the number of vertices in the molecular graph.

Keywords: QSAR; Quail dietary toxicity; Flexible descriptor; Graph of atomic orbitals.

* Corresponding author at present address: Sergeli 8-A, Home 4, Room 6, 100085 Tashkent, Uzbekistan; e-mail: aatoropov@yahoo.com

0960-894X/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2005.12.085

The LHFG can be converted into GAO as follows:[6,7]

1. Each vertex of the LHFG is replaced by the group of atomic orbitals.
2. Elements of the adjacency matrix of the GAO are defined as

$$a_{ij} = \begin{cases} 1, \text{ if } i\text{th and } j\text{th vertices of GAO fall in group} \\ \quad \text{of different atoms from LHFG and these atoms} \\ \quad \text{have joint edge in the LHFG} \\ 0, \text{otherwise.} \end{cases}$$

We used the Monte Carlo optimization procedure to calculate numerical values of the correlation weights, which produce as large as possible correlation coefficients between the ${}^{0}\mathrm{X}_{\mathrm{CW}}$ and quail dietary toxicity on the training set. The least-squares method gave a model of

$$\log(1/C) = C_0 + C_1 \times {}^{0}\mathrm{X}_{\mathrm{CW}}(v_k, {}^{X}\mathrm{EC}_k). \qquad (2)$$

The predictive ability of the Eq. 2 was validated with compounds of the test set.

The most important indicators of quality of the QSAR model are the statistical characteristics on the external test set and, in fact, the correct results on the test set give an indication of the predictive ability of the model. The statistical characteristics of QSAR models obtained with different versions of the Eq. 1 are presented in Table 1. These versions vary for the used descriptors, LHFG or GAO, and the different connectivity order.

Taking into account the complexity of modelling toxic endpoints in general and quail dietary toxicity, in particular, reproducibility of the statistical quality of the QSARs under consideration shows adequate value.

The best models were obtained with the correlation weighting of the first-order extended connectivity (${}^{1}\mathrm{EC}$) in the GAO. In spite of the very good statistical quality of the QSAR model based on the correlation weighting of the third order extended connectivity on the training set, this model was unsatisfactory, since the statistical quality on the test set was poor. Thus, the ${}^{0}\mathrm{X}_{\mathrm{CW}}$ (ao_k, ${}^{1}\mathrm{EC}_k$)-based model of 78 optimized parameters is the best because its statistical characteristics are the best for the external test set.

The model obtained in first ${}^{0}\mathrm{X}_{\mathrm{CW}}$ (ao_k, ${}^{1}\mathrm{EC}_k$)-based OCWLI probe is the following:

$$\begin{aligned} \log(1/C) &= -1.269 + 2.793 \times {}^{0}\mathrm{X}_{\mathrm{CW}}(\mathrm{ao}_k, {}^{1}\mathrm{EC}_k) \\ &n = 91, r^2 = 0.7819, r^2(\text{pred}) = 0.7818, \\ &s = 0.346, F = 319 \text{ (training set)} \\ &n = 19, r^2 = 0.6534, r^2(\text{pred}) = 0.6443, \\ &s = 0.474, F = 32 \text{ (test set)}. \end{aligned} \qquad (3)$$

Some (CW-1) values are constantly positive or negative, while others change in different models. Positive (CW-1) values are related to greater toxicity [see Eq. 3] and vice versa for negative (CW-1) values. The results on the set of pesticides never used to build up the model are good, considering that the model is global, thus not specific for a particular chemical class. The results on the test set are

Table 1. Statistical characteristics of the QSAR obtained with OCWLI for different versions of the descriptors calculated by Eq. 1: r is correlation coefficient, s is standard error of estimation and F is Fisher F-ratio

Descriptor	Number of optimized parameters	Probe	Training set $n = 91$			Test set $n = 19$		
			r^2	s	F	r^2	s	F
${}^{0}\mathrm{X}_{\mathrm{CW}}$ (a_k, ${}^{0}\mathrm{EC}_k$)	13	1	0.5128	0.517	94	0.2719	0.688	6
		2	0.5051	0.521	91	0.3182	0.663	8
		3	0.5315	0.507	101	0.3018	0.671	7
${}^{0}\mathrm{X}_{\mathrm{CW}}$ (a_k, ${}^{1}\mathrm{EC}_k$)	24	1	0.6250	0.454	148	0.2127	0.735	5
		2	0.6338	0.448	154	0.1872	0.759	4
		3	0.6531	0.436	168	0.2296	0.720	5
${}^{0}\mathrm{X}_{\mathrm{CW}}$ (a_k, ${}^{2}\mathrm{EC}_k$)	50	1	0.7712	0.354	300	0.4069	0.631	12
		2	0.8123	0.321	385	0.4507	0.592	14
		3	0.7827	0.345	321	0.4053	0.623	12
${}^{0}\mathrm{X}_{\mathrm{CW}}$ (a_k, ${}^{3}\mathrm{EC}_k$)[a]	116	1	0.9250	0.201	1098	0.1977	1.026	4
		2	0.9190	0.209	1010	0.2195	0.900	4
		3	0.9184	0.210	1002	0.2142	0.907	4
${}^{0}\mathrm{X}_{\mathrm{CW}}$ (ao_k, ${}^{0}\mathrm{EC}_k$)	33	1	0.6109	0.462	140	0.2996	0.706	7
		2	0.5722	0.485	119	0.2870	0.701	7
		3	0.6187	0.458	144	0.3073	0.699	8
${}^{0}\mathrm{X}_{\mathrm{CW}}$ (ao_k, ${}^{1}\mathrm{EC}_k$)	78	1	0.7819	0.346	319	0.6534	0.474	32
		2	0.7919	0.338	339	0.6474	0.487	31
		3	0.7804	0.347	316	0.6515	0.476	32

[a] For the ${}^{0}\mathrm{X}_{\mathrm{CW}}$ (a_k, ${}^{3}\mathrm{EC}_k$)-based OCWLI for difenoconazole there are ${}^{3}\mathrm{EC}_k$values which are absent in the training set, because for this OCWLI there are18 compounds in the test set.

worse than on the training set, but in any case the error is within 1 log unit. We should remember that experimental data can typically vary by a factor of 4, at least—often more. The pesticides we used include chlorinated compounds, thiazines, organophosphates, carbamides, etc. These compounds typically have several functional groups and are quite active on living organisms, since pesticides are intended to have adverse efforts on a certain target. These pesticides produce their effect through a series of biological mechanisms, not a single one, and this increases the complexity of the modelling tasks. We know little about their toxicity towards birds and, in fact, many more studies have been done on aquatic toxicity.[11–17] Considering the lack of background knowledge on the toxic mechanisms involved in bird toxicity, the use of chemical descriptors that are not directly related to their bio-mechanism is acceptable. The chemical descriptors we used have the great advantage of ease of calculation and speed: needing no optimization of the chemical structure in the three-dimensional space. Another advantage of our approach is that it is reproducible, while chemical descriptors using three-dimensional optimization are more variable and can change with the operator.[13]

Since we found no QSAR models on quail toxicity in the literature, we compared the statistical characteristics on toxicity towards *Daphnia magna*[14–17] and those of the quail toxicity model calculated with Eq. 3. This comparison indicates that the model gave a reasonably good result.

There are only a few QSAR studies on birds, and the present study used quite a large set of compounds with widely varying chemical structures including complex compounds, such as pesticides, which exert their toxicity through different biochemical processes.

The GAO-based approach gave better results than the LHFG-based one for the toxic endpoint under consideration.

Groups of atomic orbitals for atoms which take place in pesticides under consideration, list of compounds in training and test; example of calculation ${}^{0}X_{CW}$ (ao_k, ${}^{1}EC_k$) are available as Supplementary materials.

Acknowledgment

We acknowledge the financial contribution of the European Community project DEMETRA.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmcl.2005.12.085.

References and notes

1. Toropov, A. A.; Toropova, A. P. *THEOCHEM* **2002**, *578*, 129.
2. Toropov, A. A.; Schultz, T. W. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 560.
3. Toropov, A.-A.; Roy, K. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 179.
4. Toropov, A. A.; Benfenati, E. *THEOCHEM* **2004**, *676*, 165.
5. Toropov, A. A.; Benfenati, E. *THEOCHEM* **2004**, *679*, 225.
6. Toropov, A. A.; Toropova, A. P. *THEOCHEM* **2003**, *637*, 1.
7. Toropov, A. A.; Toropova, A. P. *THEOCHEM* **2001**, *538*, 287.
8. Toropov, A. A.; Toropova, A. P.; Nesterova, A. I.; Nabiev, O. M. *Chem. Phys. Lett.* **2004**, *384*, 357.
9. <http://www.bba.de/>.
10. Roncaglioni, A.; Befenati, E.; Boriani, E.; Clook, M. *J. Environ. Sci. Health Part B* **2004**, *39*, 641.
11. Russom, C. L.; Bradbury, S. P.; Broderius, S. J.; Hammermeister, D. E.; Drummond, A. *Environ. Toxicol. Chem.* **1997**, *16*, 948.
12. Verhaar, H. J. M.; Van Leeuwen, C. J.; Hermens, J. L. M. *Chemosphere* **1992**, *25*, 471.
13. Benfenati, E.; Piclin, N.; Roncaglioni, A.; Varì, M. R. *SAR and QSAR Environ. Res.* **2001**, *12*, 593.
14. Liu, X.; Wang, B.; Huang, Zh.; Han, Sh.; Wang, L. *Chemosphere* **2003**, *50*, 403.
15. Faucon, J. C.; Bureau, R.; Faisant, J.; Briens, F.; Rault, F. *Chemosphere* **2001**, *44*, 407.
16. Tao, Sh.; Xi, X.; Xu, F.; Li, B.; Cao, J.; Dawson, R. *Environ. Pollut.* **2002**, *116*, 57.
17. Kaiser, K. L. E. *THEOCHEM* **2003**, *622*, 85.