J Mol Model (2002) 8:168–172
DOI 10.1007/s00894-002-0074-0

ORIGINAL PAPER

Mohamed Zahouily · Abdallah Rhihil
Halima Bazoui · Saïd Sebti · Driss Zakarya

Structure-toxicity relationships study of a series of organophosphorus insecticides

Received: 3 December 2001 / Accepted: 4 February 2002 / Published online: 16 May 2002
© Springer-Verlag 2002

Abstract Structure-toxicity relationships were studied for a set of 47 insecticides by means of multiple linear regression (MLR) and artificial neural network (ANN). A model with three descriptors, including shape surface [$S(R_2)$], hydrogen-bonding acceptors [$HBA(R_2)$] and molar refraction [$MR(R_1)$], showed good statistics both in the regression ($r = 0.875$, $s = 0.417$ and $q^2 = 0.675$) and artificial neural network model with a configuration of [3-5-1] ($r = 0.966$, $s = 0.200$ and $q^2 = 0.647$). The statistics for the prediction on toxicity [log LD_{50} (lethal dose 50, oral, rat)] in the test set of 20 organophosphorus insecticides derivatives is ($r = 0.849$, $s = 0.435$) and ($r = 0.748$, $s = 0.576$) for MLR and ANN respectively. The model descriptors indicate the importance of molar refraction and shape contributions toward toxicity of organophosphorus insecticides derivatives used in this study. This information is pertinent to the further design of new insecticides.

Keywords Multiple linear regression · Artificial neural network · Organophosphorus · Insecticides · Acute LD_{50} · Descriptors

Introduction

Although the benefits of pesticides [1] are undeniable, attention has been focused in recent years on their impact on human health and environment. Pesticide is a generic term for a variety of chemical classes such as insecticides, herbicides, fungicides and nematicides.

Although pesticide laws require that both risks drive the processing terms of depth of analysis and allocation of federal resources, two questions relative to risks are appropriate: What is "acceptable risk"? How can we minimize the assessing the risk?

Computer simulation techniques potentially offer a further means to probe structure-toxicity relationships. Quantitative Structure-Activity relationships (QSAR) [2, 3] represent the most effective computational approaches in drug design. QSAR is largely used to predict activities and to define pesticides models [4, 5, 6].

In the present article, we have attempted to establish Structure-Toxicity relationships for organophosphorus insecticides derivatives by multiple linear regression (MLR) and artificial neural networks (ANN).

The objectives of our study were both to provide supplementary information concerning the behavior of these compounds and further define the criteria necessary for the rational design of a new generation of organophosphorus type insecticides.

Material and methods

Experimental data

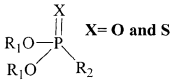
The organophosphorus insecticide derivatives were taken from different literature sources [7, 8, 9]. The log LD_{50} (lethal dose 50, acute, oral, rat) values were used as the dependent variable that represents LD_{50}. The toxicity (LD_{50}) was expressed in mols of toxicant per kilogram of body weight. When LD_{50} is given in an interval, the minimum value was used.

The chemical structures along with experimental toxicity (log LD_{50}) data of the compounds used in this study are shown in Table 1.

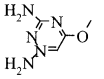
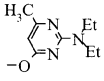
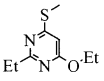
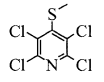
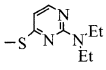
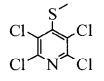
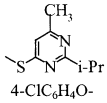
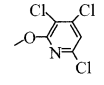
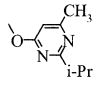
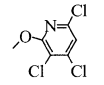
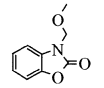
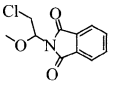
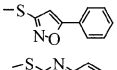
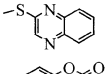
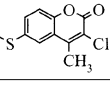
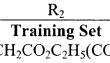
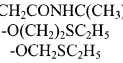
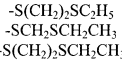
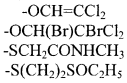
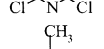
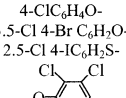
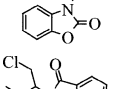
M. Zahouily (✉) · A. Rhihil · H. Bazoui · D. Zakarya
UFR Chimie Appliquée,
Laboratoire de Synthèse Organique
et Traitement de l'Information Chimique,
Département de Chimie, Faculté des Sciences et Techniques,
B.P. 146, Mohammedia 20650, Maroc
e-mail: zahouily@voila.fr
Tel.: +212-023-314705/08, Fax: +212-3-315353

A. Rhihil · H. Bazoui · S. Sebti
Laboratoire de Chimie Organique
Appliquée et Catalyse(LCOAC),
Faculté des Sciences Ben M'Sik B.P. 7955 Sidi Othmane,
Casablanca, Maroc

Table 1 Chemical structures compounds studied and experimental toxicity (log LD_{50}) values



X= O and S

N°	X	R_1	R_2	[a]log LD_{50}Exp.
			Training Set	
1	S	CH_3	$-OCHCH_2CO_2C_2H_5(CO_2C_2H_5)$	0.56
2	S	CH_3	$-OCH_2CONH(CH_2)_2OCH_3$	0.19
3	S	CH_3	$-OCH_2CON(CH_3)CHO$	0.15
4	S	CH_3	$-O(CH_2)_2SCH(CH_3)_2$	-0.24
5	S	C_2H_5	$-OCH_2CON(CH_3)CO_2C_2H_5$	-0.96
6	S	C_2H_5	$-OCH_2CONHC(CH_3)_2$	-1.55
7	S	C_2H_5	$-O(CH_2)_2SC_2H_5$	-2.02
8	S	C_2H_5	$-OCH_2SC_2H_5$	-2.11
9	S	CH_3	$-S(CH_2)_2SC_2H_5$	-0.11
10	S	C_2H_5	$-SCH_2SCH_2CH_3$	-1.85
11	S	C_2H_5	$-S(CH_2)_2SCH_2CH_3$	-2.14
12	O	CH_3	$-OCH=CCl_2$	-0.44
13	O	CH_3	$-OCH(Br)CBrCl_2$	0.05
14	S	CH_3	$-SCH_2CONHCH_3$	0.03
15	O	CH_3	$-S(CH_2)_2SOC_2H_5$	-0.58
16	S	CH_3		0.50
17	S	CH_3		0.85
18	S	CH_3		0.97
19	S	CH_3		0.47
20	S	C_2H_5		-0.36
21	S	C_2H_5		-0.41
22	S	C_2H_5		-0.45
23	S	CH_3	$4\text{-}ClC_6H_4O\text{-}$	-0.33
24	S	CH_3	$3.5\text{-}Cl\ 4\text{-}Br\ C_6H_2O\text{-}$	0.91
25	S	CH_3	$2.5\text{-}Cl\ 4\text{-}IC_6H_2S\text{-}$	0.55
26	S	C_2H_5		-0.56
27	S	CH_3	$3\text{-}CH_3\ 4\text{-}NO_2C_6H_3O\text{-}$	0.26
28	S	CH_3	$3\text{-}Cl\ 4\text{-}NO_2C_6H_3O\text{-}$	0.70
29	S	C_2H_5		-0.48
30	S	CH_3		0.47
31	S	C_2H_5		-0.43
32	S	C_2H_5		-0.90
33	S	C_2H_5		-0.45
34	S	C_2H_5		-0.63
35	S	C_2H_5		-0.56
36	S	C_2H_5		-0.84
37	S	C_2H_5		-1.06
38	S	CH_3	$2.4.5\text{-}ClC_6H_2S\text{-}$	0.59
39	S	CH_3	$3\text{-}Cl\ 4\text{-}NO_2C_6H_3S\text{-}$	0.47
40	S	CH_3	$2\text{-}Cl\ 4\text{-}NO_2C_6H_3S\text{-}$	0.13
41	S	CH_3	$3\text{-}CH_3\ 4\text{-}NO_2C_6H_3S\text{-}$	0.26
42	S	CH_3		-0.05
43	S	CH_3		-0.27
44	S	C_2H_5		-0.41
45	S	C_2H_5	$2.5\text{-}Cl\ 4\text{-}Br\ C_6H_2S\text{-}$	-0.88
46	S	CH_3		-0.11
47	S	C_2H_5		-0.90
			Prediction Set	
48	S	C_2H_5		-1.70
49	S	C_2H_5		-1.29
50	S	CH_3	$-OCH_2CONHC_2H_5$	0.15
51	S	CH_3	$-OCH_2CONHCH_3$	-0.03
52	S	CH_3		-0.32
53	S	C_2H_5	$-OCH_2Cl$	-1.52
54	S	CH_3		0.23
55	S	CH_3	$2\text{-}Cl\ 4\text{-}NO_2C_6H_3O\text{-}$	-0.90
56	S	CH_3	$-SCHCH_2CO_2C_2H_5(CO_2C_2H_5)$	0.62
57	O	CH_3	$-CH(OH)CCl_3$	0.24
58	S	CH_3		-0.33
59	O	C_2H_5		-1.48
60	S	C_2H_5		-1.09
61	S	CH_3		-0.20
62	S	C_2H_5		-0.73
63	S	C_2H_5		-0.39
64	S	CH_3		0.58
65	O	CH_3		1.04
66	O	CH_3	$-S(CH_2)_2SC_2H_5$	0.35
67	S	C_2H_5		-1.24

[a]log LD_{50} observed.

Structural descriptors

Several structural descriptors and physicochemical variables were used to characterize the organophosphorus insecticide derivatives under study. Those descriptors were calculated for the substituents R_1, R_2 and X.

These include the octanol/water partition coefficient (log P), [10] used as a descriptor of the hydrophobic molecular properties and electronegativity, [11] hydrogen-bonding donors (HBD), hydrogen-bonding acceptors (HBA) [12] and molar refraction [13].

The size and shape of the substituents were quantified by their van der Waals volume (V), molecular weight (MW), surface (S), [14] length (L), V/L [15] and topological descriptors [16].

60 parameters were calculated for each compound.

Statistical methods

Multiple linear regression (MLR)

Multiple linear regression was used to generate the linear models and it was performed with the Unistat statistical package running on a Pentium PC.

Because of the large number of descriptors considered, a stepwise multiple linear regression procedure based on the forward-selection and backward-elimination methods was used to select the powerful descriptors.

In order to avoid all difficulties in interpretation of the resulting models, pairs of variables with a correlation coefficient larger than 0.7 were classified as intercorrelated, and only one of these was included in the screened models. The quality of the model was proven by the correlation coefficient square r^2, the standard deviation s and the Fischer test value (F), when all parameters in the model were significant at the 95% confidence level. An analysis of the predictive ability was carried out in two ways. The predictive ability in the training set (47 compounds N°1 to 47) was carried out using leave-one-out cross-validation. For a reliable model, the squared predictive correlation coefficient q^2 [17] should be >0.60 [18]. In addition, 20 organophosphorus insecticide derivatives (N°48 to 67) were retained to test the actual prediction of the model (r^2 and s are considered).

Artificial neural network (ANN)

The ANN [19] was trained by the back-propagation (BP) of errors algorithm [20] had the following architecture:

- An input layer including pertinent descriptors of MLR
- A hidden layer for which the ratio of the number of data points in the training set and the number of connections controlled by the network, ρ, is critical to the predictive power of the neural net. The range $1.8<\rho<2.2$ [ρ = (number of data points in the training set)/(number of adjustable weights controlled by the network)] [21], was used as a guideline for an acceptable number of neurons in the hidden layer. It is claimed that, for $\rho << 1.0$, the network simply memorises the data, whereas for $\rho >> 3.0$, the network loses its ability to generalize.
- Output layer of one neuron, representing the toxicity (log LD_{50}). The input and output values were normalized.

After this step, the learning rate was varied from 0.01 to 0.9, and for each learning rate the momentum was examined from 0.1 to 0.9. The number of the neurons in the hidden layer with the use of optimized momentum and learning rate was determined.

Finally, to preclude training [22] we have studied the variation of the root mean-squares (RMS) error versus number of iteration and we have used two strategies for testing the validity of the selected ANN model.

Results and discussion

Multiple linear regression analysis

Multiple linear regression was performed on the compounds described in Table 1. We included all 47 organophosphorus insecticides derivatives (compounds N°1 to 47) of the training set for the model generation. After collecting the data, we submitted all parameters to the regression; many models were generated using this method. We obtained the best models without constant terms [Eq. (2)] because the constant term is not statistically significant. However, an ideal model [Eq. (2)] is one that has high r^2 and F values, low standard deviation, least numbers of independent variables, and high ability for prediction.

$$\begin{aligned}\log LD_{50} &= 0.132\,(\pm 0.030)\,S(R_2) \\ &\quad + 0.081\,(\pm 0.033)\,HBA(R_2) \\ &\quad - 0.274\,(\pm 0.027)\,MR(R_1) + 0.320\,(\pm 0.320) \\ n &= 47 \\ r^2 &= 0.740 \\ s &= 0.420 \\ F &= 40.16\end{aligned} \tag{1}$$

$$\begin{aligned}\log LD_{50} &= 0.156\,(\pm 0.019)\,S(R_2) \\ &\quad + 0.083\,(\pm 0.033)\,HBA(R_2) \\ &\quad - 0.260\,(\pm 0.023)\,MR(R_1) \\ n &= 47 \\ r^2 &= 0.766 \\ s &= 0.417 \\ F &= 47.94\end{aligned} \tag{2}$$

The statistical quality of Eq. (2) is fairly good and accounts for 77% of the variance in log LD_{50}. Low toxicity (high log LD_{50} values) is associated with high shape surface [$S(R_2)$] and hydrogen-bonding acceptors [$HBA(R_2)$] with decreased molar refractivity [$MR(R_1)$].

The plot of experimental log LD_{50} versus calculated log LD_{50} is given in Fig. 1a. Cross-correlation analysis

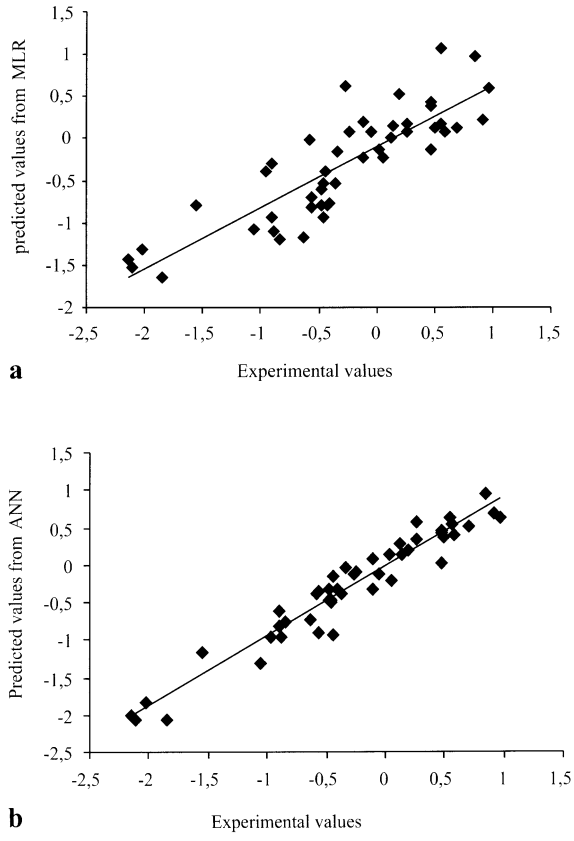


Fig. 1 Experimental and predicted values from MLR (**a**) and ANN (**b**) for the training sets.

Table 2 Correlation matrix

	S (R_2)	ABH (R_2)	MR (R_1)
S (R_2)	1		
ABH (R_2)	0.045	1	
MR (R_1)	0.039	0.229	1

showed that all pairwise correlations were ≤ 0.229 in this equation, also indicating a low collinearity (see Table 2). In the cross-validation phase, 47 subsets were created according to the leave-one-out method and the output of the removed compound was predicted for each subset. The cross-validation coefficient obtained was: $q^2 = 0.675$. The model obtained was considered to be good predictive one, according to Wold [18].

As a second strategy, the toxicity of 20 organophosphorus insecticides was predicted by using the best MLR model [Eq. (2)].

Results for the prediction in test set of 20 compounds were $r^2 = 0.721$ and s = 0.435. There were two compounds with a large estimation error for Eq. (2) (compounds N°59 and 64), and these were excluded from the standard deviation of predictions (s = 0.369). Hence, these results are in good agreement with those obtained for the training sets and reveal an good predictive quality for the MLR model.

As biological phenomena are considered to be non-linear by nature it therefore appears very interesting to study the present series of compounds with the ANN technique in order to discover possible non-linear relationships between toxicity (log LD_{50}) and the molecular descriptors that appeared pertinent for the linear model.

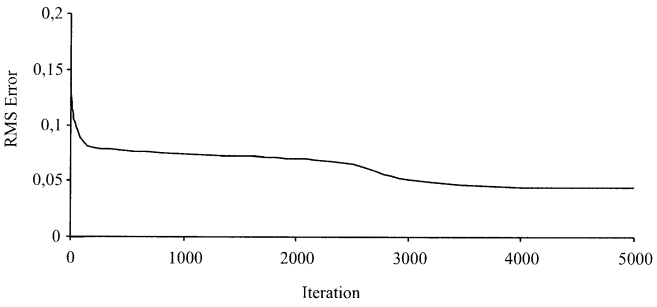


Fig. 2 Variation of RMS error versus number iteration.

Table 3 Variation of r^2 and s with number of hidden neurons

Hidden neurons	r^2(training)	s(training)	r^2(test)	s(test)
3	0.8964	0.2389	0.5445	0.5972
4	0.9093	0.2230	0.4925	0.6185
5	0.9333	0.2000	0.5597	0.5761
6	0.9170	0.2130	0.5521	0.5813
7	0.9181	0.2113	0.5287	0.5964
8	0.9155	0.2152	0.5152	0.6028

Artificial neural network analysis

The ANN was generated by using the pertinent descriptors appearing in the MLR model as input. A 3-5-1 neural network architecture was developed with the optimum learning rate and momentum 0.2 of and 0.9, respectively and with 5 000 iterations (the results of the ANN did not vary significantly between 4 800–5 000 iterations). The five hidden neurons were chosen to maintain ρ between 1.8 and 2.2. To verify this condition we have tried three to eight neurons in the hidden layer and found that five hidden neurons gives the best result for the training and test sets, as shown in Table 3.

The [3-5-1] neural network architecture shows that the standard deviation between calculated and observed toxicity was 0.200, which was found to be superior to that obtained using MLR (s = 0.417). In addition, the correlation coefficient square between observed and calculated values was 0.933. These results indicate the existence of non-linear relationships between toxicity and molecular descriptors that appeared pertinent for the linear model. The variation root mean-squares (RMS) error versus number of iteration is plotted in Fig. 2.

The plot in Fig. 1b indicates that there is a significant correlation between actual values and calculated values of logLD_{50}.

We used the same procedure as far the MLR analysis for testing the validity of the selected ANN model. The

Table 4 Evaluating the impact of each descriptor in ANN and MLR

Removed descriptor	C%	C%	r^2	s	r^2	s
	MLR [a]	ANN[a]	MLR[b]	MLR[b]	ANN[b]	ANN[b]
$S(R_2)$	34	30	0.3091	0.657	0.6708	0.444
$HBA(R_2)$	23	28	0.6906	0.439	0.7242	0.407
$RM(R_1)$	43	42	0.2116	0.810	0.3697	0.615

[a] the contribution (C%) of descriptor given by the second method described in the text.
[b] Given by the first method described in the text.

corresponding r^2 and s for the prediction in test set were 0.560 and 0.576 respectively. For the corresponding q^2 in cross-validation method is 0.647 [18].

Analysis of descriptors contribution in ANN and MLR models

To evaluate the influence of each descriptor on the calculated toxicity, we used two methods.

The first one consists of removing a descriptor and analyzing the statistical coefficient between observed and calculated using MLR and ANN. Comparison between these statistics and those calculated by MLR and ANN when no descriptor was removed gave an idea about the importance of the descriptor removed [23]. Indeed, when the descriptor [$RM(R_1)$] is remove, the model obtained is of lower quality (r^2 is only 0.212 and 0.370 for MLR and ANN respectively (Table 4).

The second method consists to use the relation established by Chastrette [24] [Eq. (3)] to calculate the contribution of each descriptor (Table 4).

$$C_i = 100.\Delta m_i / \sum_{i=1}^{3} \Delta m_i \qquad (3)$$

C_i: Contribution of descriptor i

Δm_i: The mean of deviation absolute values between the observed and estimated toxicity for all compounds.

These contributions allow the following classification: MR (R_1) >S (R_2) >HBA(R_2). These results confirm the large effect of the substituent R_1 on the toxicity see molecules: 9 (R_1= Me) – 11(R_1= Et) = -2.03 (Δ activity) and 19 (R_1= Me) – 21(R_1 = Et) = 0.88].

To ensure that the results obtained in MLR and ANN were not due to chance and lend credence to our results, we have run a scrambling experiment. The dependent variable log (LD_{50}) is randomly scrambled and then the same algorithms used in MLR and ANN run once again. The statistical results as the correlation coefficient square r^2 and the standard deviation s of its results are compared with the r^2 and s of the MLR and ANN models developed in this work. The r^2 values were 0.017 and 0.561 compared with 0.766 and 0.933 for the s values we have obtained 0.793 and 0.513 compared with 0.417 and 0.200 for the training set in MLR and ANN, respectively. This test confirms and clearly shows that the descriptors selected in this study describe very well toxicity studied.

Conclusion

Two important consequences emerge from the present report.

Firstly, taking into account the complex nature of modeled biological phenomena, on the one hand, and the large number of compounds analyzed, on the other hand, our results clearly indicate that the molar refraction is prime importance for the toxicity of the organophosphorus insecticides derivatives under study. In addition, the approach used for the contributions and classification of descriptors in MLR and ANN, may be of help in QSAR interpretations.

Secondly, this results revealed good stability of studied structure-toxicity relationships, and confirm the fact that toxicity depends, in a great part, on the structural features of the insecticide.

References

1. Young AL (1987) Pesticides minimising the risk. In: Ragsdale NN, Kuhr RJ (eds) ACS symposium series 336. American chemical society, Washington, pp 10–40
2. Devillers J, Karcher W (1990) Environmental chemistry and toxicology. In: Karcher W, Devillers J (eds) Kluwer academic press publishers, Dordrecht, pp 181–195
3. Bazoui H, Zahouily M, Sebti S, Boulajaaj S, Zakarya D (2002) Structure-toxicity relationships study of a series of organophosphorus insecticides. J Mol Mod DOI 10.1007/s00894-001-0054-9
4. Livingstone DJ (1989) Res Pest Sci 27:287–304
5. Nendza M (1991) Chemsphere 22:613–623
6. Vighi M, Garlanda MM, Calamari D (1991) Sci Total Environ 109/110:605–622
7. Büchel KH (1983) Chemistry of pesticides. John Wiley & Sons, pp 48–124
8. Meister RT, Fitzgerald GT, Zilenziger A (1987) Farm chemical handbook. Meister Publishing Co, pp 42–208
9. Thomson TW (1972) Agricultural chemicals book I. Insecticides, pp 160–266
10. Nys GG, Rekker RF (1974) Eur J Med Chem Ther 4:361–375
11. Pauling L (1960) The nature of chemical bond, 3rd edn. Cornell University Press, Ithaca NY, p 85
12. Yokohama T, Taft RW, Kamlet M J (1976) J Am Chem Soc 98:3233–3235
13. Weast RC (1988) Handbook of chemistry and physics, 1st edn. CRS, p E 318
14. Bondi A (1964) J Phys Chem 68:441–451
15. Zakarya D, Rayadh A, Samih M, Lakhlifi T (1994) Tetrahedron Lett 35:2345–2348
16. Randic M (1984) J Chem Inf Comput Sci 24:164 –175
17. Tetko IV, Villa AEP, Livingstone DJ (1996) J Chem Inf Comput Sci 36:794–803
18. Wold S (1991) Quant Struct Act Relat 10:191–193
19. Data pro Qnet 2000 for Windows V2 K build 721 neural network modelling. Vesta Services Inc, Winnetka, IL 60093, USA
20. Rumhelart DE, Hinton CE, Williams RJ (1986) Nature 323: 533–536
21. So S, Richards WG (1992) J Med Chem 35:3201–3207
22. Defernez M, Kemsley EK (1999) Analyst 124:1675–1681
23. Chastrette M, Zakarya D, Peyraud JF (1994) Eur J Med Chem 29:343–348
24. Cherquaoui D, Esseffar M, Villemin D, Cence JM, Chastrette M, Zakarya D (1998) New J Chem 22:839–843