SAR and QSAR in Environmental Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t716100694

Quantitative structure-activity relationship (QSAR) for insecticides: development of predictive *in vivo* insecticide activity models

P. K. Naik [a]; Sindhura [a]; T. Singh [a]; H. Singh [a]
[a] Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat, Himachal Pradesh, India

Online publication date: 12 November 2009

To cite this Article Naik, P. K., Sindhura, Singh, T. and Singh, H.(2009) 'Quantitative structure-activity relationship (QSAR) for insecticides: development of predictive *in vivo* insecticide activity models', SAR and QSAR in Environmental Research, 20: 5, 551 — 566
To link to this Article: DOI: 10.1080/10629360903278735
URL: http://dx.doi.org/10.1080/10629360903278735

SAR and QSAR in Environmental Research
Vol. 20, Nos. 5–6, July–September 2009, 551–566

Taylor & Francis
Taylor & Francis Group

Quantitative structure–activity relationship (QSAR) for insecticides: development of predictive *in vivo* insecticide activity models

P.K. Naik*, Sindhura, T. Singh and H. Singh

Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat, Solan 173215, Himachal Pradesh, India

(*Received 11 June 2009; in final form 19 August 2009*)

Quantitative structure–activity relationship (QSAR) analyses were performed independently on data sets belonging to two groups of insecticides, namely the organophosphates and carbamates. Several types of descriptors including topological, spatial, thermodynamic, information content, lead likeness and E-state indices were used to derive quantitative relationships between insecticide activities and structural properties of chemicals. A systematic search approach based on missing value, zero value, simple correlation and multi-collinearity tests as well as the use of a genetic algorithm allowed the optimal selection of the descriptors used to generate the models. The QSAR models developed for both organophosphate and carbamate groups revealed good predictability with r^2 values of 0.949 and 0.838 as well as q^2_{cv} values of 0.890 and 0.765, respectively. In addition, a linear correlation was observed between the predicted and experimental LD_{50} values for the test set data with r^2 of 0.871 and 0.788 for both the organophosphate and carbamate groups, indicating that the prediction accuracy of the QSAR models was acceptable. The models were also tested successfully from external validation criteria. QSAR models developed in this study should help further design of novel potent insecticides.

Keywords: insecticides; organophosphates; carbamates; QSAR; insecticide activity

1. Introduction

Insecticides are agents of chemical or biological origin that control insects. Control may result from killing the insect or otherwise preventing it from engaging in behaviours deemed destructive. Organophosphates (OPs) and carbamates exert their neurotoxic effects inhibiting actylcholinesterase (AchE, EC 1.1.1.7), a critical enzyme involved in nerve impulse transmission [1]. Chronic toxicity resulting from OP exposure ranges from cholinesterase inhibition in plasma, erythrocytes and brain tissue to the appearance of clinical signs of long-term damage to the central nervous system as well as the peripheral nervous system [2,3]. Although cholinesterase inhibition by carbamates is somewhat reversible, organophosphate poisoning is not reversible. This means that the insecticide does not release the bound cholinesterase [4]. Although the modes of action of inhibition

*Corresponding author. Email: pknaik73@rediffmail.com

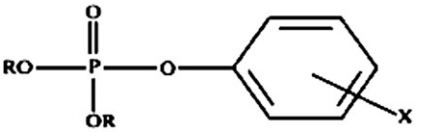
ISSN 1062–936X print/ISSN 1029–046X online

DOI: 10.1080/10629360903278735
http://www.informaworld.com

of acetylcholinesterase (AChE) by carbamate and organophosphorus esters are similar, there are distinct differences in the reactions between the enzyme and the two classes of compounds. First, while appropriate chemical reactivity is essential for high anticholinesterase activity for an organophosphorus ester, a good fit of the carbamate on the enzyme active site is essential for high anticholinesterase activity by a carbamate ester. Second, spontaneous regeneration of the carbamylated enzyme to active or original enzyme is relatively fast compared to spontaneous regeneration of a phosphorylated enzyme. For example, the half-life for recovery of *N*-methylcarbamylated AChE is approximately 30 min, while that for an organophosphorus ester ranges from several hours to days, depending upon the nature of the groups attached to the phosphorus atom. In some cases, AChE that is inhibited by certain types of organophosphorus esters is irreversibly phosphorylated and spontaneous regeneration does not occur [5]. Consequently, decrease of sensitivity of AchE to inhibition of insecticides has been implicated in insecticide resistance in many insects. Molecular studies indicated that the decrease of AchE sensitivity to inhibition is due to mutation(s) of the AchE gene. Such mutations show structural modifications of the enzyme, which often result qualitatively in modified enzyme property, including sensitivity of the enzyme to inhibition by insecticides. With the restrictions on the use of most of the persistent organochlorine insecticides imposed in the 1970s, the less persistent but highly effective organophosphate agents have become the insecticides of first choice, and have been the most widespread pesticides used worldwide. Hence, organophosphate and carbamate insecticides are under increasing regulatory pressure. All efforts are focused on developing insecticides with new, presumably safer, modes of action. A rational approach for developing new insecticides awaits development of a global mechanism of action model including high-throughput screening and/or a predictive quantitative structure–activity relationship (QSAR) model. With the advent of parallel synthesis methods and technology, we might also expect the number of insecticides based on scaffold structure of organophosphates and carbamates to be tested to achieve dramatic growth. One method of orchestrating these strategies is to make use of QSAR models for the rapid prediction and virtual pre-screening of insecticide activity.

A QSAR equation is a mathematical equation that correlates the biological activity to a wide variety of physical or chemical parameters. There are many examples available in the literature in which QSAR models have been used successfully for the screening of compounds for biological activity [6–9]. Traditional QSAR studies have been used since the early 1970s to predict activities of untested molecules. The pre-requisite of developing QSAR equations is the availability of a wide range of molecular structures and their complementary activities. For insecticides with the availability of molecular structures and their complementary activities from various laboratories it has been assumed that the most important criterion for a systematic study of QSARs has been satisfied. QSAR studies have been done for the organophosphates and carbamates [10] using only free-energy-related physiochemical substituent parameters such as π, σ and others. Accordingly, in this QSAR study for the organophosphates and carbamates we have applied E-state, electronic, structural, topological, quantum mechanics and physicochemical based descriptors, which can be calculated without structural alignments. Further, the behaviour of QSAR models was examined with a variety of statistical parameters in line with what has been used by Deswal and Roy [11] for the development of thrombin inhibitors.

Table 1. Organophosphate derivatives used in this work with their insecticide activity (log LD_{50}, mol L^{-1}) to housefly (*Musca nebulo*).



Analogue	*R*	*X*	*log LD_{50}*	*Analogue*	*R*	*X*	*log LD_{50}*
1	CH_3	H	2.75	19	C_2H_5	3-CN	5.0
2	CH_3	3-CH_3	2.0	20	C_2H_5	4-CN	5.1
3	CH_3	4-CH_3	1.99	21	C_2H_5	3-NO_2	5.1
4	CH_3	4-OCH_3	2.0	22	C_2H_5	4-NO_2	5.2
5	CH_3	3-Cl	2.1	23	C_2H_5	2,4-Cl	4.3
6	CH_3	4-Cl	2.6	24	C_2H_5	2,5-Cl	4.1
7	CH_3	3-Br	4.0	25	C_4H_9	H	2.5
8	CH_3	4-Br	3.53	26	C_4H_9	3-CH_3	2.0
9	CH_3	3-CN	4.99	27	C_4H_9	4-CH_3	2.1
10	CH_3	4-CN	4.84	28	C_4H_9	4-OCH_3	2.1
11	CH_3	3-NO_2	4.9	29	C_4H_9	3-Cl	2.8
12	CH_3	4-NO_2	5.1	30	C_4H_9	4-Cl	2.5
13	C_2H_5	H	3.2	31	C_4H_9	4-Br	2.95
14	C_2H_5	4-CH_3	3.0	32	C_4H_9	3-CN	4.0
15	C_2H_5	3-Cl	3.8	33	C_4H_9	4-CN	4.01
16	C_2H_5	4-Cl	3.72	34	C_4H_9	3-NO_2	4.21
17	C_2H_5	3-Br	4.11	35	C_4H_9	4-NO_2	4.38
18	C_2H_5	4-Br	4.06				

2. Materials and methods

2.1 *Data set*

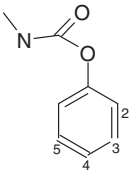
A total of 84 insecticide analogues were used in the study and were taken from various sources belonging to two different groups of modifications as mentioned in Tables 1 and 2.

Sublib-I, commonly known as organophosphate insecticides, consists of 35 compounds (Table 1). Structural modifications are mainly introduced at varying radicals at positions X and R in the scaffold structure (Table 1). The acute toxicity data (LD_{50}, mol L^{-1}) of these compounds to housefly (*Musca nebulo* L.) were taken from Hansch et al. [12], except for the compound 24 from Gandhe and Purnanand [13]. All chemicals are analogues to methyl and ethyl paraoxons (compounds 12 and 22), which are capable of inhibiting AChE directly [14]. The selected chemicals have significant differences in structure for the substituents X at *meta* and *para* positions ranging from electron-donating group (–CH_3) to electron-withdrawing group (–NO_2), while the alkyl group R varies from methyl to butyl.

Sublib-II includes 49 compounds commonly called carbamates (Table 2) with their insecticide activities to housefly (*Musca nebulo*) being taken from different sources [15–17].

All these insecticides were built from the scaffolds by different ring modifications and substitutions of functional groups as mentioned in Tables 1 and 2. We used ISIS Draw 2.3 software for sketching structures and converting them to their 3D

Table 2. Carbamate derivatives used in this work with their insecticide activity (log LD_{50}, mol fly^{-1}) to housefly (*Musca nebulo*).



Analogue	*Substituent*	*log LD_{50}*	*Analogue*	*Substituent*	*log LD_{50}*
1	H	−8.033	26	3-OMe	−8.12
2	2-Me	−7.77	27	3-OEt	−8.23
3	2-Et	−8.024	28	3-O-i-Pr	−8.033
4	2-i-Pr	−8.45	29	3-O-n-Bu	−7.77
5	2-n-Pr	−8.35	30	3-CN	−7.24
6	2-F	−7.97	31	3-NO_2	−7.0
7	2-Cl	−8.39	32	4-Me	−8.07
8	2-Br	−8.43	33	4-Et	−7.95
9	2-I	−8.46	34	4-i-Pr	−7.0
10	2-OMe	−7.86	35	4-n-Pr	−7.0
11	2-OEt	−8.4	36	4-t-Bu	−7.0
12	2-O-i-Pr	−8.81	37	4-F	−8.02
13	2-O-s-Bu	−7.97	38	4-Cl	−7.0
14	2-O-Allyl	−7.43	39	4-Br	−7.0
15	2-CN	−7.88	40	4-I	−7.0
16	2-NO_2	−7.95	41	4-OMe	−7.26
17	3-Me	−8.31	42	4-OEt	−7.0
18	3-Et	−8.17	43	4-CN	−7.0
19	3-i-Pr	−7.45	44	4-NO_2	−7.0
20	3-n-Pr	−7.06	45	3,4-Me_2	−7.85
21	3-t-Bu	−7.99	46	2-O-i-Pr-5-Me	−8.52
22	3-F	−8.13	47	2-O-i-Bu-3-Me	−8.33
23	3-Cl	−7.86	48	2-O-i-Pr-5-n-Pr	−8.74
24	3-Br	−8.31	49	2-O-i-Pr-5-s-Bu	−8.26
25	3-I	−8.38			

representations by using the ChemSketch 3D viewer of ACDLABS 8.0. LigPrep (Schrödinger L. L. C. 2007) was used for final preparation of ligands. LigPrep is a utility of the Schrödinger software suite that combines tools for generating 3D structures from 1D (Smiles) and 2D (SDF) representations, searching for tautomers and steric isomers and performing a geometry minimization of ligands. The ligands were energy minimized using a macromodel module of Schrödinger with default parameters and applying molecular mechanics force fields (MMFFs). A truncated Newton conjugate gradient (TNCG) minimization method was used with 500 iterations and a convergence threshold of 0.05 kJ mol^{-1}.

2.2 *Descriptor calculation*

E-state indices [18], M log P [19], superpendentic index [20], structural [21], symmetrical, topological, lead likeness [22], electronic Wang–Ford atomic charge and extended Huckel partial charge [23,24], bulk, moments, orbital energies, molecular connectivity indices [25],

Table 3. List of descriptors used in this study.

Type	*Descriptors*
E-state indices	Electro-topological-state indices
Electronic	Partial positive surface area, partial negative surface area, relative positive charge, relative negative charge, relative positive charged surface area, relative negative charged surface area, weighted positive charged partial surface area, weighted negative charged partial surface area, fractional negative charged partial surface area, fractional positive charged partial surface area, Huckel molecular orbital indices, highest occupied molecular orbital, lowest unoccupied molecular orbital, free valence value, nucleophilic superdelocalizability, free radical superdelocalizability, heat of formation, dipole moments, energy of the highest occupied orbital, energy of the lowest unoccupied orbital, electronegativity, hardness.
Information content	Information of atomic composition index, superpendentivity index.
Spatial	Radius of gyration, Jurs descriptors, shadow indices, area, density, length-to-breadth ratios.
Structural	Topological symmetry, geometrical symmetry, combined symmetry, conformational flexibility indices, molecular distance edge descriptors, moment of inertia indices, geometric moment indices, number of single bonds, number of aromatic bonds.
Thermodynamic	Average energy, bond strain energy, angle strain energy, non-bonded strain energy, torsional strain energy, total strain energy of molecule.
Lead likeness	log P (Meylan, Howard), log S, log P (Moriguchi, Hirono).
Topological	Wiener index, Kier and Hall molecular connectivity indices, path count and length descriptors, topological polar surface area (TPSA), Balaban indices.

gravitational indices [26], hydrophobicity, steric and thermodynamic factors and other topological descriptors were calculated using the ADME Model Builder software package (version 4.5). The superpendentic index was computed from the pendent matrix. These descriptors help differentiate the molecules mostly according to their size, degree of branching, flexibility and overall shape. Some of the descriptors included in the study are listed and described in Table 3.

2.3 *Regression analysis*

The total number of descriptors calculated initially was 372. A systematic search in the order of missing value test, zero test, correlation coefficient, multi-collinearity and genetic algorithm was performed to determine significant descriptors using the ADME Model Builder (version 4.5) software package (Fujitsu Inc.). Any parameter which was not calculated (missing value) for any number of the compounds in the data set was rejected in the first step. Some of the descriptors were rejected because they contained a zero value for all the compounds (zero tests). In order to minimize the effect of collinearity and to avoid redundancy, a correlation matrix was developed with a cutoff value of 0.6 and the variables physically removed from the analysis which show exact linear dependences between subsets of the variables and multi-collinearity (high multiple correlations between subsets of the variables). From the remaining descriptors, the set of descriptors that would give the statistically best QSAR models was selected by using a genetic function approach implemented in the ADME Model Builder (version 4.5) software package. The genetic

algorithm (GA) starts with the creation of a population of randomly generated parameter sets. The usage probability of a given parameter from the active set is 0.5 in any of the initial population sets. The sets are then compared according to their objective functions. The parameters set used for the GA includes: mutation 0.1, crossover 0.9, population 300, number of generations 1000 and r^2 floor limit 50% and the objective function was r^2/N_par. The form of the objective function favours sets that have r^2 as high as possible, while minimizing the number of parameters used as descriptors. The higher the score, the higher the probability that a given set will be used for the creation of the next generation of sets. Creation of a consecutive generation involves crossovers between set contents, as well as mutations. The algorithm runs until the desired number of generations is reached. Equations were developed between the observed activity and the descriptors. The best equation was taken based on the statistical parameters such as squared regression coefficient (r^2), adjusted regression coefficient (r^2_{adj}), regression coefficient cross validation and *F*-test values.

2.4 *Validation test*

The predictive capability of the QSAR equation was determined using the leave-one-out cross-validation method. The cross-validation regression coefficient (q^2_{cv}) was calculated by the following equation:

$$q^2_{cv} = 1 - \frac{\text{PRESS}}{\text{TOTAL}} = 1 - \frac{\sum_{i-1}^{n} (y_{exp} - y_{pred})^2}{\sum_{i-1}^{n} (y_{exp} - \bar{y})^2}$$

where y_{pred}, y_{exp} and $\bar{y}$ are the predicted, experimental and mean values of experimental activity, respectively. Also, the accuracy of the prediction of the QSAR equation was validated by *F*-value, r^2 and r^2_{adj}. A large *F* indicates that the model fit is not a chance occurrence. It has been shown that a high value of statistical characteristics is not necessary for the proof of a highly predictive model [27,28]. Hence, in order to evaluate the predictive ability of our QSAR model, we used the method described by Golbraikh and Tropsha [27] and Roy and Roy [28]. The values of the correlation coefficient of predicted and actual activities and the correlation coefficient for regressions through the origin (predicted vs. actual activities and vice versa) were calculated using the regression analysis Tool-pak option of Excel and other parameters were calculated as reported by the above authors [27,28]. The determination coefficient of prediction, q^2_{test}, was calculated using the following equation [28]:

$$q^2_{test} = 1 - \frac{\sum \left(Y_{\text{pred}_{\text{Test}}} - Y_{\text{Test}}\right)^2}{\sum \left(Y_{\text{Test}} - \bar{Y}_{\text{Traning}}\right)^2}$$

where $Y_{\text{pred}_{\text{Test}}}$ and Y_{Test} are the predicted value based on the QSAR equation (model response) and the experimental activity values, respectively, of the external test set compounds. Y_{Training} is the mean activity value of the training set compounds. Further evaluation of the predictive ability of the QSAR model for the external test set compounds was done by determining the value of rm^2 by the following equation [28]:

$$rm^2 = r^2\left(1 - \left|\sqrt{r^2 - r_0^2}\right|\right)$$

where r^2 is the square correlation coefficient between experimental and predicted values and r_0^2 is the squared correlation coefficient between experimental and predicted values without intercept for the external test set compounds. The values of k and k', slopes of the regression line of the predicted activity vs. actual activity and vice versa, were calculated using the following equations [29]:

$$k = \frac{\sum y_i \tilde{y}_i}{\sum \tilde{y}_i^2} \quad \text{and} \quad k' = \frac{\sum y_i \tilde{y}_i}{\sum y_i^2}$$

where $\tilde{y}_i$ and y_i are the predicted and experimental activities, respectively.

To check the intercorrelation of descriptors, variance inflation factor (VIF) analysis was performed. The VIF value is calculated from $1/(1-r^2)$, where r^2 is the multiple correlation coefficient of one descriptor's effect regressed on the remaining molecular descriptors. If the VIF value is larger than 10, information of descriptors can be hidden by correlation of descriptors [29,30].

3. Results and discussion

3.1 *QSAR model for organophosphate derivatives*

The 35 active compounds with their acute toxicity to housefly were randomly divided into a training set of 25 compounds and a test set of 10 compounds. The various molecular descriptors (372 in total) as described in Table 3 were calculated initially. By applying a missing value test, a zero test and a correlation test with a cutoff value of 0.6 we have discarded the most likely parameters that resulted in 107 parameters. Further additional parameters were discarded by applying the GA and finally seven parameters were selected for the development of the QSAR equation. Taking a brute-force approach, we increased the number of parameters in the QSAR equation one by one and evaluated the effect of addition of a new term on the statistical quality of the model. As the squared correlation coefficient, r^2, can be easily increased by the number of terms in the QSAR equation, we took the cross-validation correlation coefficient, q_{cv}^2, as the limiting factor for a number of descriptors to be used in the final model. It was observed that the q^2 value increased until the number of descriptors in the equation reached seven, as shown in Table 4. With further addition of parameters to the equation with seven descriptors, there was a decrease in the q_{cv}^2 value of the model. So, the numbers of descriptors were restricted to seven in the final QSAR model. The best significant relationship for the insecticide activity has been deduced to be

$$\begin{aligned} \log \mathrm{LD}_{50} = {} & 0.94 - 0.044\ \text{average energy} + 0.10\ \mathrm{TPSA} + 4.14\ \mathrm{PCHGMHT} \\ & - 4.71\ \mathrm{S3C} + 4.02\ \mathrm{V6PC} + 3.53\ \mathrm{SHDW6} + 0.13\ \mathrm{ELOW1}, \\ & n = 25,\ r^2 = 0.949,\ r_{\mathrm{adj}}^2 = 0.928,\ \mathrm{PRESS} = 3.344, \\ & s = 0.303,\ q_{\mathrm{cv}}^2 = 0.890,\ F = 45.11, \end{aligned} \tag{1}$$

where n is the number of compounds in the training set, r^2 is the squared correlation coefficient, s is the estimated standard deviation about the regression line, r_{adj}^2 is the square of the adjusted correlation coefficient for the degree of freedom, F-test is the measure of variance which compares two models differing by one or more variables to see if the more complex model is more reliable than the less complex one (the model is

Table 4. Statistical assessment of QSAR equations for prediction of insecticide activity to housefly (*Musca nebulo*) with varying numbers of descriptors for organophosphate analogues.

No. of descriptor	*Equation*	r^2	*PRESS*	q^2
1	$\log LD_{50} = -0.64 + 0.108$ AVERAGE	0.232	26.184	0.145
2	$\log LD_{50} = 2.91 - 0.087$ AVERAGE $+ 0.066$ TPSA	0.564	15.506	0.494
3	$\log LD_{50} = 8.84 - 0.204$ AVERAGE $+ 0.094$ TPSA $+ 26.5$ PCHGMHT	0.728	11.936	0.610
4	$\log LD_{50} = 6.88 - 0.233$ AVERAGE $+ 0.094$ TPSA $+ 21.3$ PCHGMHT $+ 1.83$ S3C	0.752	11.094	0.638
5	$\log LD_{50} = 5.49 - 0.144$ AVERAGE $+ 0.098$ TPSA $+ 13.9$ PCHGMHT $- 2.22$ S3C $+ 3.15$ V6PC	0.873	6.755	0.779
6	$\log LD_{50} = 1.65 - 0.122$ AVERAGE $+ 0.094$ TPSA $+ 8.13$ PCHGMHT $- 2.17$ S3C $+ 3.35$ V6PC $+ 4.78$ SHDW6	0.901	5.467	0.822
7	**$\log LD_{50} = 0.94 - 0.0436$ AVERAGE $+ 0.100$ TPSA $+ 4.14$ PCHGMHT $- 4.71$ S3C $+ 4.02$ V6PC $+ 3.53$ SHDW6 $+ 0.130$ ELOW1**	**0.949**	**3.344**	**0.890**
8	$\log LD_{50} = 1.62 - 0.103$ AVERAGE $+ 0.100$ TPSA $+ 3.58$ PCHGMHT $- 3.07$ S3C $+ 3.58$ V6PC $+ 2.64$ SHDW6 $+ 0.117$ ELOW1 $+ 0.027$ DIP	0.969	5.207	0.830

Table 5. Observed and predicted insecticide activities to housefly (*Musca nebulo*) of organophosphate derivatives (training set).

	Insecticide activity (log LD_{50})				*Insecticide activity (log LD_{50})*		
Compound no.	*Observed*	*Predicted*	*Residual*	*Compound no.*	*Observed*	*Predicted*	*Residual*
1	2.75	2.47	0.28	19	5.00	5.26	0.26
2	2.00	2.30	0.30	20	5.10	5.19	0.09
4	2.00	2.39	0.39	22	5.20	5.23	0.03
5	2.10	2.65	0.55	23	4.30	4.13	0.17
6	2.60	2.92	0.32	24	4.10	4.46	0.36
7	4.00	3.68	0.32	26	2.00	2.22	0.22
9	4.99	4.95	0.04	28	2.10	2.01	0.09
10	4.84	4.90	0.06	29	2.80	3.27	0.47
12	5.10	4.80	0.30	30	2.50	2.89	0.39
13	3.20	2.91	0.29	32	4.00	4.45	0.45
15	3.80	3.53	0.27	33	4.01	4.70	0.69
16	3.72	3.32	0.40	35	4.38	4.11	0.27
18	4.06	3.97	0.09				

supposed to be good if the *F*-test is above a threshold value) and q^2_{cv} is the square of the correlation coefficient of the cross validation using the leave-one-out cross-validation technique. The QSAR model developed in this study is statistically ($r^2 = 0.949$, $q^2_{cv} = 0.890$, *F*-test $= 45.11$) best fitted and consequently was used for prediction of insecticide activities (LD_{50}) of training and test sets of molecules as reported in Tables 5 and 6. The relationships between predicted (both training and test) activities and the corresponding

Table 6. Observed and predicted insecticide activities to housefly (*Musca nebulo*) of organophosphate derivatives (test set).

	Insecticide activity (log LD_{50})				*Insecticide activity (log LD_{50})*		
Compound no.	*Observed*	*Predicted*	*Residual*	*Compound no.*	*Observed*	*Predicted*	*Residual*
3	1.99	2.04	0.05	21	5.10	5.37	0.27
8	3.53	3.57	0.04	25	2.50	2.99	0.49
11	4.90	4.98	0.08	27	2.10	2.56	0.46
14	3.00	2.46	0.54	31	2.95	3.77	0.82
17	4.11	4.25	0.14	34	4.21	5.16	0.95

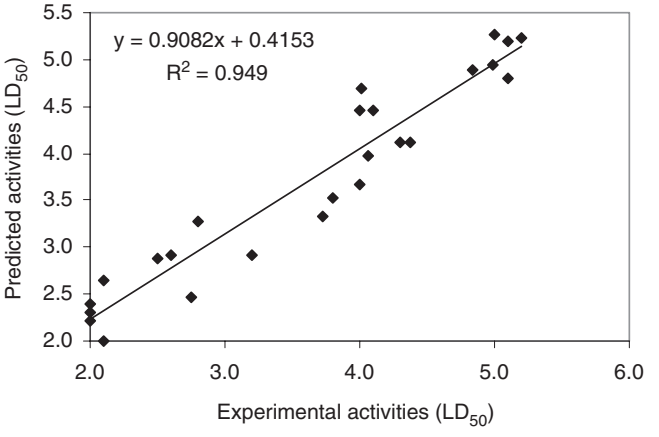


Figure 1. Relationship between predicted and experimental activities of organophosphate analogues for the training set compounds.

experimental activities are shown in Figures 1 and 2. The r^2 and q^2_{cv} values of 0.949 and 0.890, respectively, of the model corroborate with the criteria for a QSAR model to be highly predictive [27].

The intercorrelation of the descriptors used in the QSAR model (1) was very low (below 0.6), which is in conformity to the study that, for a statistically significant model, it is necessary that the descriptors involved in the equation should not be intercorrelated with each other. To further check the intercorrelation of descriptors, variance inflation factor (VIF) analysis was performed. In this model, the VIF values of these descriptors are 3.46 (average energy), 3.33 (TPSA), 1.92 (PCHGMHT), 2.32 (S3C), 2.27 (V6PC), 1.30 (SHDW6) and 2.17 (ELOW1), which are less than the threshold value of 10 [29,30].

Satisfied with the robustness of the QSAR model developed using the training set, we have applied the QSAR model to an external data set of organophosphate analogues constituting the test set. As the experimental values of LD_{50} for these compounds are already available, this set of molecules provides an excellent data set for testing the prediction power of the QSAR model for new ligands. Table 6 presents the predicted LD_{50} values of the test set based on Equation (1). The overall root mean square error (RMSE) between the experimental and predicted LD_{50} values was 0.265, which revealed

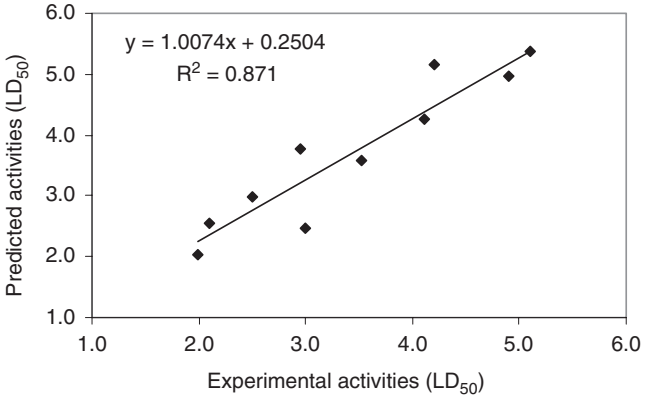


Figure 2. Relationship between predicted and experimental activities of organophosphate analogues for the test set compounds.

good predictability. The estimated correlation coefficients between experimental and predicted LD_{50} values with intercept (r^2) and without intercept (r_0^2) were 0.871 and 0.867, respectively. The value of $(r^2 - r_0^2)/r^2 = (0.871 - 0.867)/0.871 = 0.005$, which is less than 0.1 (stipulated value) [27]. Also, the values of k and k' were 0.920 and 1.074, which are well within the specified ranges of 0.85 and 1.15 [27]. The values of $q^2_{\text{test}} = 0.788$ and $rm^2 = 0.816$ were found to be in the acceptable ranges [28], thereby indicating the good external predictability of the QSAR model.

3.2 *QSAR model for carbamate derivatives*

The 49 carbamate derivatives considered as potentially toxic compounds to housefly were randomly divided into a training set of 39 compounds and a test set of 10 compounds. A set of 372 molecular descriptors initially was calculated using ADME Model Builder 4.5 and finally only six molecular descriptors were selected by applying systematic searches discussed above. Different QSAR equations were developed by applying a brute-force approach as shown in Table 7. By considering only five molecular descriptors, the QSAR Equation (2) with best significant relationship for the insecticide activity was developed:

$$\begin{aligned} \log \text{LD}_{50} = {} & -1.82 - 3.04\ \text{SYMM1} + 0.048\ \text{LOGP} - 3.93\ \text{PCHGPC} \\ & - 0.077\ \text{WNSA1\_AM1} + 0.00013\ \text{HF}, \\ & n = 39,\ r^2 = 0.838,\ r^2_{\text{adj}} = 0.812,\ \text{PRESS} = 2.652, \\ & s = 0.243,\ q^2_{\text{cv}} = 0.765,\ F = 32.07. \end{aligned} \tag{2}$$

The QSAR model developed is statistically ($r^2 = 0.838$, $q^2_{\text{cv}} = 0.765$, F-test $= 27.44$) best fitted and consequently used for computing insecticide activities of the training set (Table 8) as well as prediction of insecticide activities of the test set (Table 9) compounds. Figures 3 and 4 show the quality of fit between the predicted (both training and test sets) activities with their corresponding experimental activities. The VIF values of the descriptors used in the final equation are 1.23 (SYMM1), 1.69 (LOGP), 1.33 (PCHGPC), 1.39 (WNSA1_AM1) and 2.56 (HF), which revealed very low intercorrelation. The overall

Table 7. Statistical assessment of QSAR equations for prediction of insecticide activity to housefly (*Musca nebulo*) with varying numbers of descriptors for carbamate analogues.

No. of descriptor	*Equation*	r^2	*PRESS*	q^2
1	$\log LD_{50} = -4.67 - 3.58$ SYMM1	0.547	5.601	0.504
2	$\log LD_{50} = -3.91 - 3.37$ SYMM1 $- 3.62$ PCHGPC	0.714	3.728	0.670
3	$\log LD_{50} = -3.46 - 3.50$ SYMM1 $- 0.123$ LOGP $- 3.75$ PCHGPC	0.729	3.957	0.649
4	$\log LD_{50} = -1.77 - 3.04$ SYMM1 $+ 0.043$ LOGP $- 3.95$ PCHGPC $- 0.077$ WNSA1_AM1	0.838	2.414	0.786
5	**$\log LD_{50} = -1.82 - 3.04$ SYMM1 $+ 0.048$ LOGP $- 3.93$ PCHGPC $- 0.077$ WNSA1_AM1 $+ 0.00013$ HF**	**0.838**	**2.652**	**0.765**
6	$\log LD_{50} = -1.88 - 2.97$ SYMM1 $+ 0.057$ LOGP $- 4.10$ PCHGPC $- 0.077$ WNSA1_AM1 $- 0.0005$ HF $+ 0.012$ DIP	0.840	2.855	0.747

Table 8. Observed and predicted insecticide activities to housefly (*Musca nebulo*) of carbamate derivatives (training set).

Compound no.	*Insecticide activity (log LD_{50})*			*Compound no.*	*Insecticide activity (log LD_{50})*		
	Observed	*Predicted*	*Residual*		*Observed*	*Predicted*	*Residual*
2	−7.77	−7.57	0.20	28	−8.03	−7.82	0.21
3	−8.02	−7.66	0.36	29	−7.77	−7.64	0.13
5	−8.35	−8.00	0.35	30	−7.24	−7.29	0.05
6	−7.97	−7.97	0.00	31	−7.00	−7.42	0.42
8	−8.43	−8.15	0.28	32	−8.07	−7.57	0.50
9	−8.46	−7.97	0.49	33	−7.95	−7.63	0.32
10	−7.86	−7.76	0.10	34	−7.00	−7.29	0.29
11	−8.40	−8.10	0.30	35	−7.00	−7.42	0.42
14	−7.43	−7.87	0.44	37	−8.02	−8.02	0.00
15	−7.88	−8.04	0.16	38	−7.00	−7.37	0.37
16	−7.95	−7.76	0.19	40	−7.00	−7.06	0.06
17	−8.31	−7.94	0.37	41	−7.26	−7.04	0.22
18	−8.17	−7.96	0.21	42	−7.00	−7.41	0.41
20	−7.06	−7.26	0.20	43	−7.00	−7.11	0.11
21	−7.99	−8.00	0.01	44	−7.00	−7.33	0.33
22	−8.13	−7.73	0.40	46	−8.52	−8.36	0.16
23	−7.86	−7.75	0.11	47	−8.33	−8.41	0.08
24	−8.31	−7.85	0.46	48	−8.74	−8.44	0.30
25	−8.38	−8.03	0.35	49	−8.26	−8.34	0.08
26	−8.12	−7.63	0.49				

Table 9. Observed and predicted insecticide activities to housefly (*Musca nebulo*) of carbamate derivatives (test set).

Compound no.	*Insecticide activity (log LD_{50})*			*Compound no.*	*Insecticide activity (log LD_{50})*		
	Observed	*Predicted*	*Residual*		*Observed*	*Predicted*	*Residual*
1	−8.03	−7.75	0.29	36	−7.00	−7.19	0.19
7	−8.39	−8.04	0.35	39	−7.00	−7.23	0.23
13	−7.97	−7.97	0.00	45	−7.85	−8.18	0.33
19	−7.45	−7.84	0.39	4	−8.45	−8.09	0.36
27	−8.23	−7.90	0.33	12	−8.81	−8.49	0.32

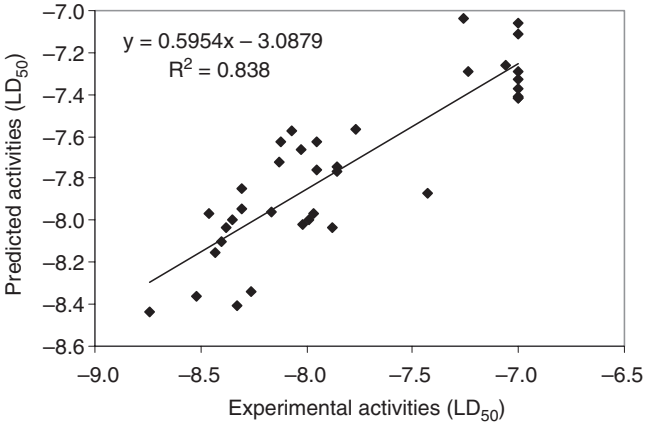


Figure 3. Relationship between predicted and experimental activities of carbamate analogues for the training set compounds.

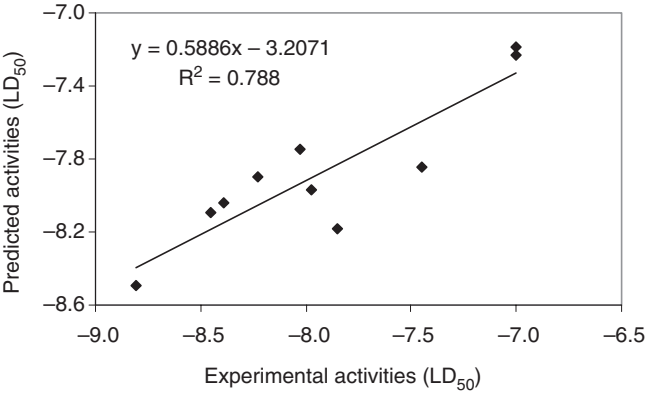


Figure 4. Relationship between predicted and experimental activities of carbamate analogues for the test set compounds.

Table 10. Descriptor set used in the development of QSAR equations for the two groups of compounds (organophosphates and carbamates).

Descriptor type	*Explanation*
Average energy	The average binding energy of the molecule calculated from all functional groups.
TPSA	Topological polar surface area.
PCHGMHT	Mean partial charge on heteroatoms.
PCHGPC	It is the most positive partial charge on C atom.
S3C	Third-order cluster molecular connectivity indices which treat all the bond types identically. This descriptor contains information about the size and the degree of branching in a molecule.
V6PC	Sixth-order path-cluster molecular connectivity valence. This descriptor contains information about the size and the degree of branching in a molecule.
SHDW6	Projection area of the molecules in *YZ* plane.
ELOW1	Difference between minimum and maximum E-state values of the molecules. The E-state is meant to encode information regarding intermolecular interactions.
M Log P	Octanol/water partition coefficient of the molecule based on the algorithm by Moriguchi et al. [31].
SYMM1	It includes information regarding topological symmetry of the molecule based on 2D structure.
WNSA1-AM1	Weighted negative charged partial surface area descriptor.
HF	Heat of formation descriptor.

RMSE between the experimental and predicted LD_{50} values for the external test set compounds is 0.259, which revealed good predictability of the QSAR equation. The estimated squared correlation coefficients between experimental and predicted LD_{50} values with intercept (r^2) and without intercept (r_0^2) are 0.788 and 0.731 respectively for the test set compounds. The value of $(r^2 - r_0^2)/r^2 = (0.788 - 0.731)/0.788 = 0.072$, which is less than 0.1 (stipulated value) [27]. Also, the values of k and k' were 1.007 and 0.991, which are well within the specified ranges of 0.85 and 1.15. All these values satisfied the criteria for a QSAR model to be highly predictive [27]. Also, the values of $q^2_{\text{test}} = 0.737$ and $rm^2 = 0.6$ were found to be in the acceptable ranges [28], thereby indicating the good external predictability of the QSAR model.

The definition of the molecular descriptors used in QSARs has been included in Table 10. Based on the developed QSAR models, it has been observed that the most important parameters that contribute to the potential insecticide activity are TPSA, PCHGMHT, PCHGPC and WNSA1-AM1. These descriptors depend upon the solvent-accessible surface area and partial charged surface area of the molecule. This is well supported if we examine compounds 25–31 (organophosphate derivatives), where substitution of non-polar groups results in significant increase of activity. Further substitution of polar groups among compounds 7–12, 15–24 and 32–35 (organophosphate derivatives) results in significant loss of activity. The same trend is also seen among the carbamate derivatives. Considering the molecules 46–49 (carbamate derivatives), the introduction of non-polar groups results in significant increase of activity, whereas substitution of polar groups in compounds 30, 31 and 38–44 (carbamate derivatives) results in significant loss of activity. M log P calculates the octanol/water partition

coefficient of the molecule based on the algorithm by Moriguchi et al. [31]. It is the most popular and traditional descriptor. It explains one of the principal characteristics of any preparation, the lipophilicity. The higher its value, the more probable the transfer of the preparation from the aqueous medium into the biological membrane. This property is critical for medicinal preparations that are administered orally and must be absorbed through the GI tract. A log P value less than 0.5 will be absorbed appropriately. For all the compounds, QSAR predictions produce exactly the same trend for insecticide activity, even though the exact magnitudes of these values do not match very well to experimental values. Coupled with the good predictive ability of the QSAR models developed in this study, we believe that these models would perform well as rapid screening tools to uncover new and more potent insecticides based on organophosphate and carbamate derivatizations.

4. Conclusion

In this study, we used a systematic way of variable selection in the order of missing value test → zero test → simple correlation test → multi-collinearity test → genetic algorithm to obtain QSAR models for 84 insecticides. Using a combination of topological, electro-topological-state index and electronic and thermodynamic descriptors of chemical structures, we have built several robust QSAR models with high values of q^2_{cv} (for training sets) and determination coefficient of prediction q^2_{test} (for test sets). The high predictive ability of the models allows virtual screening of chemical databases or virtual libraries determined by either synthetic feasibility or commercial availability of starting materials to prioritize the synthesis of most promising candidates. Therefore, these models should facilitate the rational design of novel derivatives, guide the design of focused libraries based on the skeleton of organophosphates and carbamates and facilitate the search for related structures with similar biological activity from large databases.

Acknowledgement

We wish to thank Scube Scientific Software, New Delhi for supplying ADME Model Builder for a trial period.

References

[1] K.Y. Zhu and J.R. Gao, *Increased activity associated with reduced sensitivity of acetylcholinesterase in organophosphate resistant greenbug,* Schizaphis graminum *(Homoptera: Aphididae)*, Pestic. Sci. 55 (1999), pp. 11–17.

[2] H.D. Durham and D.J. Ecobichon, *An assessment of the neurotoxic potential of fenitrothion in the hen*, Toxicology 41 (1986), pp. 319–332.

[3] D.J. Ecobichon, J.E. Davies, J. Doull, M. Ehrich, R. Joy, D. McMillan, R. MacPhail, L.W. Reiter, W. Slikker, and H. Tilson, *Neurotoxic effects of pesticides*, in *The Effects of Pesticides on Human Health*, Vol. 18, S.R. Baker and C.F. Wilkinson, eds., Princeton Scientific, Princeton, NJ, 1990, pp. 131–199.

[4] A.E. Brown, *Mode of Action of Insecticides and Related Pest Control Chemicals for Production Agriculture, Ornamentals, and Turf*, Pesticide Information Leaflet No. 43 (2005), pp. 1–13.

[5] T.R. Fukuto, *Mechanism of action of organophosphorus and carbamate insecticides*, Environ. Health Perspect. 87 (1990), pp. 245–254.
[6] L.M. Shi, Y. Fan, T.G. Myers, and J.N. Paul, *Mining the NCI anticancer drug discovery databases: Genetic function approximation for the QSAR study of anticancer ellipticine analogues*, J. Chem. Inf. Comput. Sci. 38 (1998), pp. 189–199.
[7] S. Oloff, R.B. Mailman, and A. Trospha, *Application of validated QSAR models of D_1 dopaminergic antagonists for database mining*, J. Med. Chem. 48 (2005), pp. 7322–7332.
[8] Y. Meneses-Marcel, Y. Marrero-Ponce, A. Machado-Tugores, D.M. Monterro-Torres, J.A. Pereira, J.J. Escario, C. Nogel-Ruiz, V.J. Ochao, A.R. Aran, R.N. Martinez-Fernadez, and S. Garcia, *A linear discrimination analysis based virtual screening of trichomonacidal lead-like compounds: Outcomes of in silico studies supported by experimental results*, Bioorg. Med. Chem. Lett. 15 (2005), pp. 3838–3843.
[9] L. Santana, H. Uriarte, H. Gonzalez-Diaz, R. Zagotto, E. Soto-Otero, and J. Mendez-Alvarez, *A QSAR model for in silico screening of MAO-A inhibitors. Prediction, synthesis, and biological assay of novel coumarins*, J. Med. Chem. 49 (2006), pp. 1149–1156.
[10] K. Kamoshita, I. Ohno, T. Fujita, T. Nishioka, and M. Nakajima, *Quantitative–activity relationships of phenyl N-methylcarbamates against house fly and its acetylcholinesterase*, Pestic. Biochem. Physiol. 11 (1979), pp. 83–103.
[11] S. Deswal and N. Roy, *Quantitative structure activity relationship studies of aryl heterocycle-based thrombin inhibitors*, J. Med. Chem. 41 (2006), pp. 1339–1346.
[12] C. Hansch, A. Kurup, and R. Garg, *Chem-bioinformatics and QSAR: A review of QSAR lacking positive hydrophobic terms*, Chem. Rev. 101 (2001), pp. 619–672.
[13] B.R. Gandhe and R.P. Purnanand, *Use of gas chromatographic retention indices for quantitative structure activity relationship studies of dialkyl phenyl phosphates*, Pestic. Sci. 29 (1990), pp. 379–385.
[14] C.N. Pope, *Organophosphorus pesticides: Do they all have the same mechanism of toxicity?* J. Toxicol. Environ. Health Part B: Crit. Rev. 2 (1999), pp. 161–181.
[15] R. Heiss, E. Boecker, and H. Jung, *Phenyl N-methylcarbamates*, Belg. Patent 615364 (1962), pp. 7–11.
[16] T.R. Fukuto, R.L. Metcalf, M.Y. Winton, and P.A. Roberts, *The synergism of substituted phenyl N-methylcarbamates by piperonyl butoxide*, J. Econ. Entomol. 55 (1962), p. 341.
[17] R.L. Metcalf, C. Fuertes-Polo, and T.R. Fukuto, *Insecticidal activity of multi-substituted chloro and methyl phenyl N-methylcarbamates*, J. Econ. Entomol. 56 (1963), p. 862.
[18] C. de Gregorio, L.B. Kier, and L.H. Hall, *QSAR modeling with the electrotopological state indices: Corticosteroids*, J. Comput. Aid. Mol. Des. 12 (1998), pp. 557–561.
[19] W.M. Meylan and P.H. Howard, *Atom/fragment contribution method for estimating octanol/water partition coefficients*, J. Pharm. Sci. 84 (1995), pp. 83–92.
[20] S. Gupta, M. Singh, and A.K. Madan, *Superpendentic index: A novel topological descriptor for prediction of biological activity*, J. Chem. Inf. Comput. Sci. 39 (1999), pp. 272–277.
[21] S.S. Liu, C.Z. Cao, and Z.L. Li, *Approach to estimation and prediction for normal boiling point (NBP) of alkanes based on a novel molecular distance edge (MDE) vector* λ, J. Chem. Inf. Comput. Sci. 38 (1998), pp. 387–394.
[22] A. Lipinski, F. Lombardo, B. Dominy, and P. Feeney, *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings*, Adv. Drug Del. Rev. 46 (2001), pp. 3–26.
[23] G.M. Eliopoulos and R.C. Moellering Jr, *Antimicrobial combinations*, in *Antibiotics in Laboratory Medicine*, 4th ed., V. Lorian, ed., Lippincott Williams & Wilkins, Philadelphia, USA, 1996, pp. 330–396.
[24] N.P. Brenwald, M.J. Gill, and R. Wise, *Prevalence of a putative efflux mechanism among fluoroquinolone-resistant clinical isolates of* Streptococcus pneumoniae, Antimicrob. Agents Chemother. 42 (1998), pp. 2032–2035.

[25] L.B. Kier and L.H. Hall, *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, 1976.
[26] A.R. Katritzky, M. Lan, V.S. Lobanov, and M. Karelson, *Correlation of boiling points with molecular structure. I. A training set of 298 diverse organics and a test set of 9 simple inorganics*, J. Phys. Chem. 100 (1996), pp. 10400–10407.
[27] A. Golbraikh and A. Tropsha, *Beware of q^2*, J. Mol. Graph. Model. 20 (2002), pp. 269–276.
[28] P.P. Roy and K. Roy, *On some aspects of variable selection for partial least squares regression models*, QSAR Comb. Sci. 27 (2008), pp. 302–313.
[29] M. Jaiswal, P.V. Khadikar, A. Scozzafava, and C.T. Supuran, *Carbonic anhydrase inhibitors: The first QSAR study on inhibition of tumor-associated isoenzyme IX with aromatic and heterocyclic sulfonamides*, Bioorg. Med. Chem. Lett. 14 (2004), pp. 3283–3290.
[30] S. Shapiro and B. Guggenheim, *Inhibition of oral bacteria by phenolic compounds. Part 1. QSAR analysis using molecular connectivity*, Quant. Struct.–Act. Relat. 17 (1998), pp. 327–337.
[31] I. Moriguchi, S. Hirono, Q. Liu, I. Nakagome, and Y. Matsushita, *Simple method of calculating octanol/water partition coefficient*, Chem. Pharm. Bull. 40 (1992), pp. 127–130.