Science of the Total Environment 610–611 (2018) 937–943

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

The quantitative structure–insecticidal activity relationships from plant derived compounds against chikungunya and zika *Aedes aegypti* (Diptera:Culicidae) vector

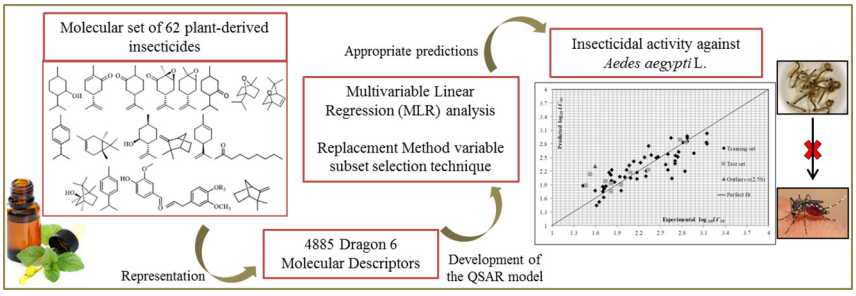
Laura M. Saavedra [a,*], Gustavo P. Romanelli [b,c], Ciro E. Rozo [d], Pablo R. Duchowicz [a,*]

[a] Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), CONICET, UNLP, Diag. 113 y 64, C.C. 16, Sucursal 4, 1900 La Plata, Argentina
[b] Centro de Investigación y Desarrollo en Ciencias Aplicadas "Dr. J.J. Ronco" (CINDECA), Departamento de Química, Facultad de Ciencias Exactas, CONICET, UNLP, Calle 47 No. 257, B1900AJK La Plata, Argentina
[c] Cátedra de Química Orgánica, Centro de Investigación en Sanidad Vegetal (CISaV), Facultad de Ciencias Agrarias y Forestales, Universidad Nacional de La Plata, Calles 60 y 119 s/n, B1904AAN La Plata, Argentina
[d] Grupo de Investigaciones Ambientales para el Desarrollo Sostenible (GIADS), Universidad Santo Tomas, Seccional Bucaramanga, Carrera 18 No. 9-27. 680011 Bucaramanga, Colombia

HIGHLIGHTS

- Prediction of the insecticidal activity for sixty-two plant derived compounds against chikungunya and zika *A. aegypti* vector.
- QSAR models are suggested for modelling the acute toxicity of bioactive molecules using 4885 Dragon 6 molecular descriptors.
- The Replacement Method based on Multivariable Linear Regression is a reliable feature selection method in QSAR.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 3 April 2017
Received in revised form 8 August 2017
Accepted 12 August 2017
Available online xxxx

Keywords:
Insecticidal activity
Chikungunya
Zika
Aedes aegypti
QSAR theory
MLR analysis

ABSTRACT

The insecticidal activity of a series of 62 plant derived molecules against the chikungunya, dengue and zika vector, the *Aedes aegypti* (Diptera:Culicidae) mosquito, is subjected to a Quantitative Structure–Activity Relationships (QSAR) analysis. The Replacement Method (RM) variable subset selection technique based on Multivariable Linear Regression (MLR) proves to be successful for exploring 4885 molecular descriptors calculated with Dragon 6. The predictive capability of the obtained models is confirmed through an external test set of compounds, Leave-One-Out (LOO) cross-validation and Y-Randomization. The present study constitutes a first necessary computational step for designing less toxic insecticides.

1. Introduction

The transmission of endemic diseases is mainly allocated to mosquitoes. There are reported alarming values of morbidity and mortality in tropical and subtropical regions,(Katritzky et al., 2008) where it is estimated that approximately 2.5 billion people live under the threat of

* Corresponding authors.
E-mail addresses: laurasaa0913@gmail.com (L.M. Saavedra), pabloducho@gmail.com (P.R. Duchowicz).

http://dx.doi.org/10.1016/j.scitotenv.2017.08.119
0048-9697/

various arbovirus types such as yellow fever, dengue and chikungunya fever.(Ocampoa et al., 2011) This last one has an important impact on public health in Brazil, Mexico, Colombia and Argentina, with 40,000 case reports of viral infection from 2013 to 2015.(World Health Organization (WHO), n.d.-a)

During the last years, mosquitoes have been responsible for the transmission of the zika virus (ZIKV) in Brazil and Colombia with 146,675 recognized cases.(World Health Organization (WHO), n.d.-b) The main infection reason is the proliferation of vectors such as *Aedes aegypti*, *Aedes leucocaelenus*, *Aedes albopictus* and *Aedes sabethes*. Owing to continuous climatic changes, these insects have increased their population, thus spreading to new territories where they have not been previously found.(Gillij et al., n.d.) Moreover, the effectiveness of pesticides used for diseases vector control has been increasingly affected by environmental conditions.(WHO, 2006; WHO, 2009)

The vectors propagation control method, proposed by the Pan American Health Organization (PAHO) during the fifties, has applied the organochlorine insecticide dichloro-diphenyl-trichloroethane (DDT) in twenty-one countries. Unfortunately, some years later, the vector population has become highly resistant to this pesticide, thus leading to its use ban. According to the National Pesticide Information Center (NPIC), DDT currently persists in the environment resulting in high toxicity to humans and animals as birds, fish and rats.(National Pesticide Information Center- NPIC; Oregon State University, 1999) In the eighties, Colombia has proposed the implementation of organophosphorus compounds (OPs) such as temephos, with a high insecticidal activity against *A. aegypti* in larval stage and also in others non-targeted animals.

Synthetic repellents have been developed for use as personal protection from bloodsucking insects. These substances have local action modes acting on the central nervous system (CNS) of the insect, causing deterrent effects that result in host avoidance. The widely used *N*,*N*-diethyl-toluamide (DEET) repellent,(Environment Protection Agency-EPA.; U. E. P. A., 1980) a compound of topical application, has some problems with his efficacy, limited protection time, irritation cases are reported, allergies and systemic intoxication in humans, and also resistance in arthropods, such as *Drosophila melanogaster* house fly are reported.(Bhattacharjee et al., 2005)

The need to find new environmentally friendly compounds having insecticidal properties leads to the use of natural products, particularly the essential oils (EOs) and plant extracts, which are complex mixtures of bioactive compounds possessing various biological properties.(Song et al., 2013; Ceferina et al., 2006) Secondary metabolites, such as terpenes, alkaloids and phenylpropanoids are abundant compounds in nature, present in fruits, leaves and flowers. They are involved in a wide range of applications and are found in cosmetics, therapeutic drugs and food additives. Furthermore, they have outstanding biological and organoleptic properties, such as *d*-limonene with inhibitory and insecticidal effects, pulgone with larvicidal property, menthol or linalool like effective insect repellents, and eugenol with antifungal activity.(Rice & Coats, 1994; Zhou et al., 2012; Carrasco et al., 2012)

The search for natural compounds with insecticidal activity against *A. aegypti* requires time, large budget and reagents for biological and clinical assays. In this sense, the possibility of employing a simple theoretical methodology for predicting biological, organoleptic or physicochemical properties of the compounds from knowing of its molecular structure represents a plausible tool in the rational design of novel bioactive molecules.

The Quantitative Structure–Activity Relationships (QSAR) theory is pioneer in the prediction of physicochemical properties or biological activities.(Hansch & Leo, 1995) Linear or non-linear mathematical models are established that include molecular descriptors characterizing relevant structural aspects of the compounds.(Katritzky et al., 1995) In fact, it has been reported that the assignation of the physicochemical meaning of the molecular descriptors could be assessed by considering the chemical orthogonal space of chemical reactivity descriptors,(Putz et al., 2017; Putz & Dudas, 2013) while the interaction mechanism, which may not be clear from the combined influence of numerical descriptors in the linear correlation, may be also pursued through a computational-conceptual algorithm for better understanding the chemical-biological interaction.(Putz & Dudas, 2013) QSAR studies are able to reduce time and costs in experimental measurements.(Ibezim et al., 2012)

In the present QSAR study, we predict the insecticidal activity against *A. aegypti* from plant derived molecules with known experimental data (62 molecules). The larvicidal activity is expressed as the median lethal concentration (LC_{50}), a standard measure of the toxicity of compounds, which measures the concentration at which 50% of third-instar larvae show lethal effect after in a specified period of the testing solutions. We apply the Replacement Method (RM) variable subset selection approach applied in the linear regression analysis of 4885 Dragon descriptors.(Duchowicz et al., 2006) In the last years, the RM technique has been successful for selecting relevant structural information and for establishing linear QSAR models with high predictive capability.(Duchowicz et al., 2008)

2. Materials and methods

2.1. Experimental data

The experimental LC_{50} insecticidal activities of 62 natural or semi-synthetic compounds are collected from the literature.(Santos et al., 2010; Santos et al., 2011; Scotti et al., 2014; Barbosa et al., 2012) For modelling purposes, such values is converted into logarithmic scale ($LC_{50} = \log LC_{50}$). Fig. 1 displays the heterogeneous molecular structures analyzed, involving terpenes, phenylpropanoids, ketones and oxygenated compounds. The complete list of LC_{50} values of the molecular set studied here is included in Table 1S of the Supplementary material.

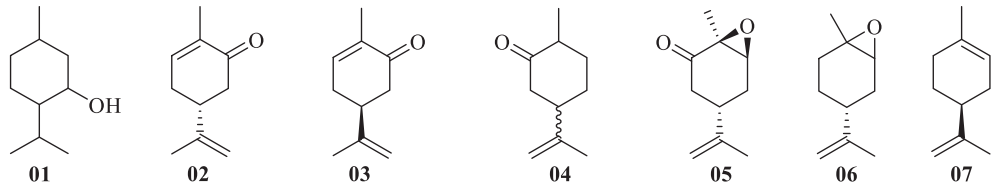
2.2. Calculation of molecular descriptors

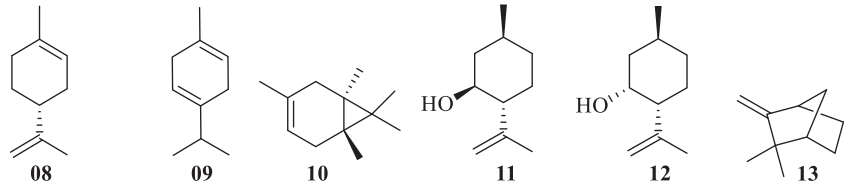
The initial conformations of the compounds are drawn in HyperChem for Windows.(HyperChem 7, n.d.) The structures are pre-optimized with the Molecular Mechanics Force Field (MM+), followed by the PM3 (Parametric Method-3) semi-empirical method to refine the structures using the Polak-Ribiere algorithm and a gradient norm limit of 0.01 kcal mol^{-1} A^{-1}.

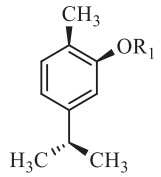
Afterward, the molecular descriptors are calculated with the Dragon 6 software.(Talete srl, n.d.) The descriptors set contains 4885 variables and includes several types characterizing the multidimensional aspects of the chemical structure: constitutional, topological, geometrical, charge, GETAWAY (geometry, topology and atoms-weighted assembly), WHIM (weighted holistic invariant molecular descriptors), 3D-MoRSE (3D molecular representation of structure based on electron diffraction), walk and path counts, 2D and 3D autocorrelations, connectivity indices, burden eigenvalues, ETA indices, edge adjacency indices, radial distribution function, Randic molecular profiles, functional groups counts and atom-centred fragments. For the descriptor set, we exclude descriptors with constant or near-constant values, and those with at least one missing value. With this process a set containing 1738 linearly-independent descriptors is achieved.

2.3. Molecular descriptor selection in MLR

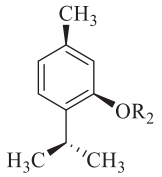
The Multivariable Linear Regression (MLR) technique has proven to be of multidisciplinary use and valuable applicability for establishing predictive QSAR models.(Duchowicz et al., 2008) Linear models are general and clearly show the effect of including/excluding descriptors in the equation, therefore, it is possible to suggest cause/effect relationships through such simple parallelisms. The main advantage of developing linear regression models is that they pose fewer over-fitting (over-training) problems, because the MLR method does not require too



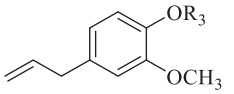




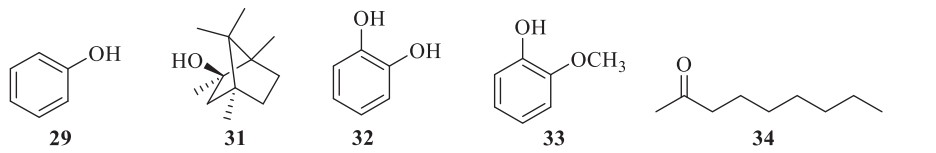
14-R_1=H
15-R_1=$COCH_3$
16-R_1=$COCH(Cl)_2$
17-R_1=$COC(Cl)_3$
18-R_1=$COCH_2CH_3$
46-R_1=COPh
47-R_1=CH_2COOH

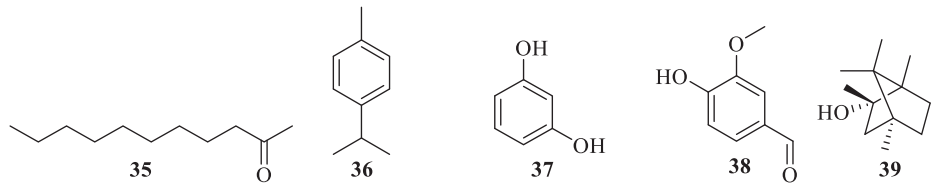


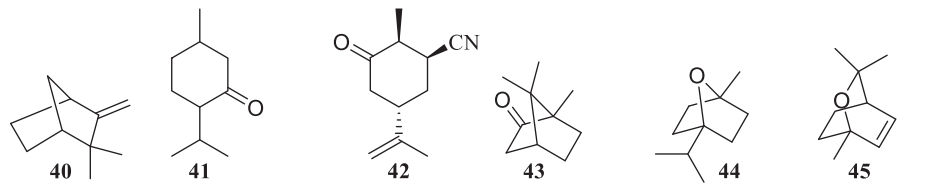
19-R_2=H
20-R_2=$COCH_3$
21-R_2=$COC(Cl)_3$
22-R_2=$COCH_2CH_3$
50-R_2=$COCH_2Cl$
51-R_2=COPh
52-R_2=CH_2CH_3
53-R_2=CH_2COOH

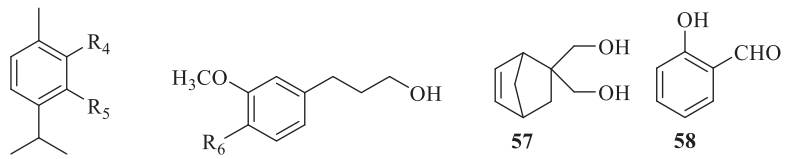


23-R_3=$COCH_3$ **27**-R_3=COPh
24-R_3=H **28**-R_3=CH_2CH_3
25-R_3=CH_3 **30**-R_3=$Si(CH_3)_2C(CH_3)_3$
26-R_3=$COCH_2CH_3$ **54**-R_3=CH_2COOH









48-R_4=OH, R_5=CHO
49-R_4=CHO, R_5=OH
55-R_6=OCOPh
56-R_6=OH

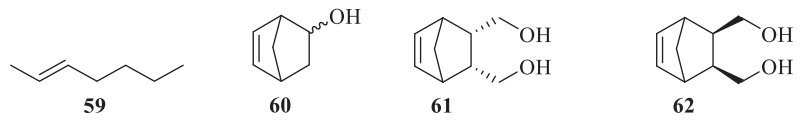


Fig. 1. Molecular structures of terpenes, phenylpropanoids, ketones and oxygenated compounds with insecticidal activity data against *Aedes aegypti*.

many optimized parameters during the model design (only a regression coefficient per descriptor). In this sense, we consider that the MLR technique is the best choice for developing predictive QSAR models, especially from molecules with few experimental data available, such as the present study.

An active research field in the QSAR theory focuses on finding new and more efficient tools for the selection of the best descriptors that explain a specific experimental activity. In this work, we choose to the Replacement Method (RM) procedure. The RM technique is an efficient optimization tool that generates MLR models on the training set by searching in a set having D descriptors for an optimal subset having $d << D$ ones with smallest standard deviation (S_{train}) or smallest root mean square error ($RMSE_{train}$).(Duchowicz et al., 2005) The quality of the results achieved with this technique is quite similar to that obtained by performing an exact (combinatorial) full search (FS) of molecular descriptors, although, of course, it requires much less computational work. The RM provides models with better statistical parameters than the one obtained with the forward stepwise regression procedure, and quite similar to the results found by the Genetic Algorithms approach. Table 2S includes a list of mathematical equations involved in the present study. The MATLAB-programmed algorithms used in our calculations are available on request.(Matlab 7.0, n.d.)

2.4. Internal and external validation in QSAR

We validate our QSAR models in order to determine their predictive power by predicting LC_{50} on compounds not considered during the calibration, and then comparing such predicted data with the real values. Therefore, the whole set of 62 compounds is partitioned into training (train) and test (test) sets. The training set is implemented for calibrating the model and obtaining optimized parameters; the test set includes "unknown" compounds not contemplated in the development procedure of the model and demonstrates the predictive capability.

It is known that randomly splitting the compounds into training and test sets does not lead to a rational selection, as both sets should have similar structure-activity relationships. For this purpose, the split of the dataset is carried out by means of the Balanced Subsets Method (BSM),(Rojas et al., 2015) based on k-means cluster analysis (k-MCA).(Xiao & Yu, 2012) The procedure involved in BSM ensures that the training set is representative of the test set.

We apply the popular theoretical validation criteria based on cross validation using the Leave-One-Out (LOO) method. The R^2_{LOO} (LOO explained variance) and S_{LOO} (LOO standard deviation) statistical parameters measure the stability of the QSAR model upon inclusion/exclusion of molecules. According to the specialized literature, R^2_{LOO} should be >0.5 for a valid model, although this is a necessary but not enough condition for determining its predictive capability.

Another validation parameter for determining the model's robustness is based on the Y-Randomization procedure.(Rücker et al., 2007) This technique consists on scrambling the logLC_{50} values in such a way that they do not correspond to the respective compounds. After calculating 30,000 cases of Y-Randomization, the obtained standard deviation (S^{rand}) has to be a poorer value than the one found by considering the true calibration (S). Therefore, when $S^{rand} > S$ it is expected that the QSAR is not fortuitous and does not result from happenstance, assuring a real structure-activity relationship.(Duchowicz et al., 2012)

Finally, another important validation criteria used here is the one proposed by Golbraikh et al.(Golbraikh et al., 2003) where some model's parameters should accomplish specific conditions for assuring predictive capability: $R^2_{test} > 0.6$, $0.85 \le k \le 1.15$ or $0.85 \le k' \le 1.15$, also $1 - R^2_0/R^2_{test} < 0.1$ or $1 - R'^2_0/R^2_{test} < 0.1$, and $R^2_m > 0.5$.

3. Results and discussion

Over the last years, there has been an increased number of research studies with the aim to find novel insecticidal agents obtained from vegetable materials; the natural products as extracts and the EOs have secondary metabolites (terpenes, phenylpropanoids, alkaloids, flavonoids) that exhibit important toxic effects against adult female mosquitoes *A. aegypti* (Diptera: Culicidae).(Santos et al., 2010; Scotti et al., 2014)

By means of the BSM, we split the dataset into a training set ($N_{train} = 52$) and a test set ($N_{test} = 10$, including compounds **3**, **8**, **13**, **17**, **19**, **20**, **22**, **39**, **46** and **62**). The cluster centroid locations in terms of descriptor values are provided as a 50×56 matrix in the c1.txt file (Supplementary material).

Afterwards, the best linear regressions are established with the RM approach, thus providing a way to explore 1738 different linearly-independent descriptors calculated with the Dragon 6 program. The RM minimizes the S_{train} parameter and selects the best "representative" $d = 1 - 6$ descriptors. The model selection is also evaluated based on the coefficient of determination (R^2) and the maximum R^2 value between descriptor pairs in the model ($R^2_{ij\max}$).

It is appreciated from Table 1 that S_{train} improves when d increases, and that for the case when $d = 6$ the S_{test} parameter is significantly higher in relation to such value for the rest of the models. The descriptor meanings are provided in Table 3S. Therefore, we conclude that a five-variables QSAR equation has acceptable statistical parameters for both the training and test sets:

$$\log LC_{50} = 8.45 - 0.45(\pm 0.1)J.DZ(i) + 0.02(\pm 0.003)ATSC5s + 19.89(\pm 6)JGI7 - -1.65(\pm 0.2)SpMax2.Bh(m) + 0.07(\pm 0.01)H.052 \quad (1)$$

$$N_{train} = 52,\ d = 5,\ R^2_{train} = 0.69,\ S_{train} = 0.28,\ N_{train}/d = 10,\ F = 20,\ R^2_{ij\,\max} = 0.26,\ o(2.5S) = 1$$

$$R^2_{LOO} = 0.60,\ S_{LOO} = 0.32,\ S^{rand} = 0.36$$

$$N_{test} = 10,\ R^2_{test} = 0.78,\ S_{test} = 0.39$$

where F is the Fisher parameter, $o(2.5S)$ indicates the number of outlier compounds having a residual (difference between experimental and calculated insecticidal activity) greater than 2.5-times S_{train}, and the N_{train}/d ratio indicates that the model satisfies with the rule of thumb.

The QSAR model's predictive capability is defined with the external validation test set, for which the percentage of explained variance is 78% and $S_{test} = 0.39$. Moreover, $R^2_{LOO} = 0.60$ and $S_{LOO} = 0.32$, indicating that the equation does not deteriorate so much with the removal of compounds (R^2_{LOO} should be higher than 0.50 for a validated model). The Y-Randomization procedure leads to a valid structure-activity relationship with $S_{train} < S^{rand}(0.36)$. We also check that Eq. 1 accomplishes the validation criteria suggested by Golbraikh et al.(Golbraikh et al., 2003) in order to assure predictive capability: $k = 0.95$, $k' = 1.04$, $R^2_0 = 0.76$, $1 - R^2_0/R^2_{test} = 0.014$, $R'^2_0 = 0.61$, $1 - R'^2_0/R^2_{test} = 0.21$ and $R^2_m = 0.69$.

Now it is possible to examine the role of the molecular descriptors involved in the QSAR model, by giving a brief description for them. Five descriptors do not depend on molecular conformation: a 2D

Table 1
The best QSAR models of different size established on 62 insecticidal compounds. The selected model appears in bold.

d	R^2_{train}	S_{train}	R^2_{test}	S_{test}	$R^2_{ij\max}$	Molecular descriptors
1	0.30	0.41	0.39	0.43	0.00	*CATS2D_05_LL*
2	0.52	0.34	0.80	0.33	0.04	*X5v*, *BLTF96*
3	0.60	0.31	0.74	0.40	0.08	*X4v*, *MATS5e*, *BLTF96*
4	0.64	0.30	0.74	0.39	0.13	*X4sol*, *SpMAD_B(m)*, *GASTS7i*, *BLTF96*
5	**0.69**	**0.28**	**0.78**	**0.39**	**0.26**	***J.Dz(i)*, *ATSC5s*, *JGI7*, *SpMax2.Bh(m)*, *H.052***
6	0.74	0.26	0.76	0.55	0.40	*JGI3*, *Chi1_EA(bo)*, *Eig08_EA(dm)*, *CATS2D_08_DA*, *F10[C−C]*, *BLTF96*

matrix-based descriptors, *J.DZ(i)*: the Balaban-like index from Barysz matrix weighted by ionization potential; two 2D autocorrelations, *ATSC5s*: the Centred Broto-Moreau autocorrelation of lag 5 weighted by I-state, and *JGI7*: the mean topological charge index of order 7; a Burden eigenvalues, *SpMax2.Bh(m)*: the largest eigenvalue n. 2 of Burden matrix weighted by the atomic mass; and finally, an Atom-centred fragment, *H.052*: H attached to C0(sp3) with 1× attached to next C. The model's correlation matrix is provided in Table 4S, revealing the absence of high inter-correlations between the 5 variables. Furthermore, the values of the model's molecular descriptors are included in Table 5S.

Fig. 2A plot the 62 predicted log LC_{50} insecticidal activities as a function of the experimental values for the compounds of the training and test sets. The dispersion plot of residuals in Fig. 2B tends to obey a random pattern around the zero line, suggesting that the assumption of the MLR technique is fulfilled. It is found an outlier for Eq. 1, the compound **55** has an irregular behavior with respect to the other compounds. After checking its structure, molecular descriptor values and experimental activity, we conclude that Eq. 1 fails to predict this molecule.

The predictions of the insecticidal property by Eq. 1 demonstrate that some molecules have high values for the acute toxicity against *Diptera* order insects (Table 6S). The compounds **7** and **8** have unsaturated cyclic hydrocarbons with endo and exo double bonds, exhibiting toxic effects against the *Aedes aegypti* mosquito in concentrations lower than 50 $\mu g/mL$ (log<1.7). Molecules **55** and **19** and its derivatives **21** and **50** also exhibit a relevant insecticidal activity; these compounds have similar structures with an aromatic ring bonded to an activator ester group with hydroxyl, chlorine or benzene substituent. Finally, we observe that molecule **59** with the aliphatic structure ($CO(CH_2)_6$) has an important acute toxicity value, quite similar to **55**, but does not have the same structure.

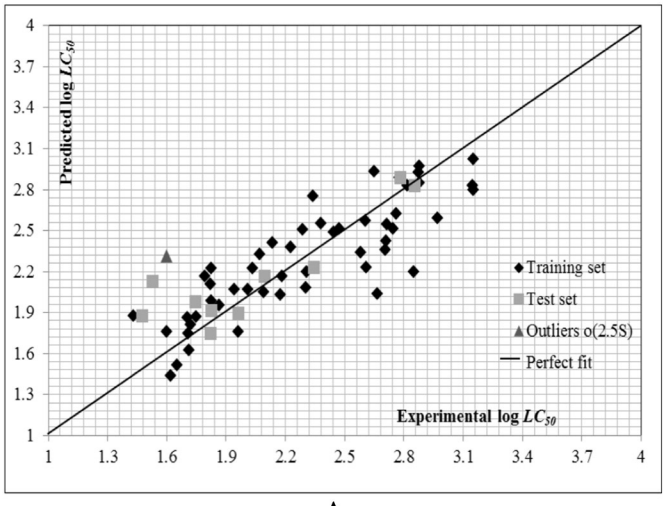


Fig. 2. A. Predicted and experimental log LC_{50} values according to Eq. 1. **B.** Dispersion plot of residuals.

As a next step of the present study and with the main purpose of improving the LC_{50} predictions provided by Eq. 1, we investigate the performance of QSAR when such models are established on plant derived molecules from specific chemical classes. We search for the best linear regressions on two different molecular sets. The first is 'set A', which includes 34 aromatic compounds (**14**–**30**, **32**, **33**, **36**–**38**, **46**–**56**, **58**) and the partition selected by means of BSM technique is $N_{train}=27$(**14**,**16**, **19**–**24**, **26**–**30**, **32**, **33**, **36**, **38**, **46**, **48**–**50**, **52**–**56**, **58**) and $N_{test}=7$(**15**, **17**, **18**, **25**, **37**, **47**, **51**). The second molecular set is 'set B', including 28 aliphatic, cyclic and bicyclic compounds (**1**–**13**, **31**, **34**, **35**, **39**, **40**–**45**, **57**, **59**–**62**) and the partition used is $N_{train}=22$ (**1**, **2**, **4**–**8**, **10**, **11**, **13**, **31**, **34**, **35**, **41**–**45**, **57**, **59**–**61**) and $N_{test}=6$ (**3**, **9**, **12**, **39**, **40**, **62**).

Table 7S includes the best 1–5 variables MLR models found in such pool of 1738 numerical descriptors through the RM technique. It is noted that the best QSAR for set A involves 4 descriptors marked in bold (Table 8S), while the best QSAR for set B has 2 descriptors (Table 9S). According to the results shown in the table, they do not ameliorate the predictive power of our first model in terms of the S_{test} parameter (the training set statistics is better but not the one for the test set). Therefore the LC_{50} predictions are not improved when the plant derived compounds are considered as belonging to specific chemical classes. Thus, the best quantitative structure–activity relationship established on this dataset of 62 insecticidal activities is given by Eq. 1.

The statistical quality of Eq. 1 is quite similar to other previous reported models by Scotti et al.,(Scotti et al., 2014) where 55 heterogeneous natural compounds are employed for calculating 128 molecular descriptors from 3D Molecular Interaction Fields (MIFs) and GRID Force Field via the VolSurf + commercial program. These authors apply Principal Component Analysis (PCA), Consensus PCA (CPCA) and Partial Least Squares Regression (PLS) methods. The results found reveal that the first two PCs account for over 60% of the data variance and the best model obtained in PLS includes $d=6$ descriptors with acceptable values for $R^2_{train}=0.71$, $R^2_{test}=0.68$ (14 compounds) and $R^2_{LOO}=0.67$.

Finally, we consider the proposal of a simpler model having simpler descriptors (simpler for interpretation - to be useful for designers or organic chemists). In this effort, we analyze the simpler descriptors out of 4885 Dragon variables, resulting in a set of 233 descriptors. The best calculated linear models are provided in Table 10S, while the following four-variables QSAR model is thus selected for predicting and better interpreting the acute toxicity:

$$\log LC_{50} = 3.38 + 0.07(\pm 0.02)TRS + 0.02(\pm 0.006)TIE - 0.3(\pm 0.06)Ui + 0.38(\pm 0.07)BLTF96 \quad (2)$$

$$N_{train}=52,\ d=4,\ R^2_{train}=0.60,\ S_{train}=0.32,\ N_{train}/d=13,\ F=17,\ R^2_{ij\,max}=0.21,\ o(2.5S)=2$$

$$R^2_{LOO}=0.50,\ S_{LOO}=0.36,\ S^{rand}=0.40$$

$$N_{test}=10,\ R^2_{test}=0.87,\ S_{test}=0.33$$

The simpler molecular descriptors appearing in Eq. 2 belong to three different classes: (i) two molecular property descriptors, *Ui*: unsaturation index, and *BLTF*96: Verhaar Fish base-line toxicity from MLOGP (mmol/l); one ring descriptor, *TRS*: total ring size, and a topological indices descriptor, *TIE*: *E*-state topological parameter. The

model's correlation matrix is provided in Table 11S, indicating the lack of high intercorrelations. The numerical values of such four descriptors is provided in Table 12S from the Supplementary material.

It is observed that the training quality of Eq. 2 does not improve the result obtained for our first model of Eq. 1 ($R^2_{train} = 0.60$, $S_{train} = 0.32$ compared to $R^2_{train} = 0.69$, $S_{train} = 0.28$), and it involves two outliers (instead of one) with a residual greater than 2.5-times S_{train}. Eq. 2 has a better predictive capability for the test set ($R^2_{test} = 0.87$, $S_{test} = 0.33$ compared to $R^2_{test} = 0.78$, $S_{test} = 0.39$), but this may result by chance as the training set quality behaves poorer in Eq. 2. The cross validation parameter $R^2_{LOO} = 0.50$ is also a poorer value than for Eq. 1 (R^2_{LOO} should be greater than 0.5 for a valid model). However, the Y-Randomization proof with $S_{train} < S^{rand}(0.40)$ and the external validation criteria suggested by Golbraikh et al., 2003, ($k = 0.93$, $k' = 1.07$, $R^2_0 = 0.86$, $1 - R^2_0/R^2_{test} = 0.005$, $R'^2_0 = 0.82$, $1 - R'^2_0/R^2_{test} = 0.05$ and $R^2_m = 0.81$) are checked in order to assure that a valid structure-activity relationship is achieved. Fig. 3A & B plot the predictions and residuals for this model, respectively.

In conclusion, the various reasons commented above allow us to select the five conformation-independent descriptors linear model (proposed in Eq. 1) as the best model found in the present QSAR study for predicting the insecticidal activities of plant-derived molecules against *A. aegypti* vector.

Fig. 3. A. Predicted and experimental logLC_{50} values for the model with $d = 4$ simpler descriptors. B. Dispersion plot of residuals.

4. Conclusion

In this work, we develop linear QSAR models from bioactive molecules that result appropriate for predicting the insecticidal activity against the *Aedes aegypti* mosquito, an important arbovirus vector that affects the public health in Latin America. The 62 plant-derived compounds are studied in three different molecular sets, which are selected according to their chemical classes and partitioned through the Balanced Subsets Method. The linear regression models explore 4885 Dragon 6 descriptors. The results obtained by means of the Multivariable Linear Regression technique coupled with the Replacement Method are successful, showing statistical parameters with suitable values that corroborate quality, veracity and robustness of the equations. Therefore, the best QSAR model found includes 5 non-conformational descriptors and achieves acceptable predictive capability, thus can be used for calculating the insecticidal activity in non-evaluated or non-synthetized compounds. In this way, it is possible to obtain bioactive compounds using renewable feedstocks, which have a rapid environmental degradation, less toxic on non-target species and the ecosystems, and also are effective against mosquitoes.

Notes

The authors declare no competing financial interest

Funding

PRD acknowledges the financial support from the National Research Council of Argentina (CONICET) PIP11220130100311 project. PRD and GPR are members of the scientific researcher career of CONICET. CER acknowledges the contributions from "Patrimonio Autónomo Fondo Nacional de Financiamiento para la Ciencia, la Tecnología y la Innovación, Francisco José de Caldas" contract RC-0572-2012 Bio-Red –CENIVAM.

Supplementary data

Supplementary data associated with this article can be found in the online version, at doi:http://dx.doi.org/10.1016/j.scitotenv.2017.08.119

References

Barbosa, J.D.F., Silva, V.B., Alves, P.B., Gumina, G., Santos, R.L.C., Sousa, D.P., Cavalcantia, S.C.H., 2012. Structure–activity relationships of eugenol derivatives against *Aedes aegypti* (Diptera: Culicidae) larvae. Pest Manag. Sci. 68, 1478–1483.

Bhattacharjee, A.K., Dheranetra, W., Nichols, D.A., Gupta, R.K., 2005. 3D pharmacophore model for insect repellent activity and discovery of new repellent candidates. QSAR Comb. Sci. 24, 593–602.

Carrasco, H., Raimondi, M., Svetaz, L., Di Liberto, M., Rodriguez, M.V., Espinoza, L., Madrid, A., Zacchino, S., 2012. Antifungal activity of eugenol analogues. Influence of different substituents and studies on mechanism of action. Molecules 17, 1002–1024.

Ceferina, A., Zygadlo, J., Mougabure, G., Biurrun, F., Zerba, E., Picollo, M., 2006. Fumigant and repellent properties of essential oils and component compounds againts permethrin-resistant *Pediculus humanus capitis* (anoplura: pediculidae) from Argentina. J. Med. Entomol. 43, 889–895.

Duchowicz, P.R., Castro, E.A., Fernández, F.M., González, M.P., 2005. A new search algorithm of QSPR/QSAR theories: normal boiling points of some organic molecules. Chem. Phys. Lett. 412, 376–380.

Duchowicz, P.R., Castro, E.A., Fernández, F.M., 2006. Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR studies. MATCH Commun. Math. Comput. Chem. 55, 179–192.

Duchowicz, P.R., Talevi, A., Bruno-Blanch, L.E., Castro, E.A., 2008. New QSPR study for the prediction of aqueous solubility of drug-like compounds. Bioorg. Med. Chem. 16, 7944–7955.

Duchowicz, P.R., Comelli, N.C., Ortiz, E.V., Castro, E.A., 2012. QSAR study for carcinogenicity in a large set of organic compounds. Curr. Drug Saf. 7, 282–288.

Environment Protection Agency- EPA; U. E. P. A *N,N*-Diethyl-m-toluamide (DEET). 1980, 1, 12–32.

Gillij, Y. G.; Gleiser, R. M.; Zygadlo, J. A., Mosquito repellent activity of essential oils of aromatic plants growing in Argentina. Bioresour. Technol. 2008, 99. 2507–2515.

Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y.D., Lee, K.H., Tropsha, A., 2003. Rational selection of training and test sets for the development of validated QSAR models. J. Comput. Aided Mol. Des. 17, 241–253.

Hansch, C., Leo, A., 1995. Exploring QSAR. Fundamentals and Applications in Chemistry and Biology. American Chemical Society, Washington, D. C.
HyperChem 7 Hypercube. Inc. URL (http://www.hyper.com).
Ibezim, E., Duchowicz, P.R., Ortiz, E.V., Castro, E.A., 2012. QSAR on aryl-piperazine derivatives with activity on malaria. Chemom. Intell. Lab. Syst. 110, 81–88.
Katritzky, A.R., Lobanov, V.S., Karelson, M., 1995. QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. Chem. Soc. Rev. 24, 279–287.
Katritzky, A.R., Wang, Z., Slavov, S., Tsikolia, M., Dobchev, D., Akhmedov, N.G., Hall, C.D., Bernier, U.R., Clark, G.G., Linthicum, K.J., 2008. Synthesis and bioassay of improved mosquito repellents predicted from chemical structure. Proc. Natl. Acad. Sci. U. S. A. 105, 7359–7364.
Matlab 7.0. Masachussetts, USA: The MathWorks. Inc. URL (http://www.mathworks.com).
National Pesticide Information Center- NPIC; Oregon State University, 1999n. DDT (Technical Fact Sheet). URL. http://npic.orst.edu/factsheets/ddttech.pdf Consulting date 20 of March 2015.
Ocampoa, C.B., Salazar-Terrerosa, M.J., Minaa, N.J., McAllisterb, J., Brogdonc, W., 2011. Insecticide resistance status of *Aedes aegypti* in 10 localities in Colombia. Acta Trop. 118, 37–44.
Putz, M.V., Dudas, N.A., 2013. Variational principles for mechanistic quantitative structure–activity relationship (QSAR) studies: application on uracil derivatives' anti-HIV action. Struct. Chem. 24, 1873–1893.
Putz, M.V., Tudoran, M.A., Putz, A.M., 2017. Structure properties and chemical-bio/ecological of PAH interactions: from synthesis to cosmic spectral lines, nanochemistry, and lipophilicity-driven reactivity. Curr. Org. Chem. 17, 2845–2871.
Rice, P., Coats, J., 1994. Insecticidal properties of monoterpenoid derivatives to the house fly (Diptera: Muscidae) and red flour beetle (Coleoptera: Tenebrionidae). Pestic. Sci. 41, 195–202.
Rojas C.; Duchowicz, P. R.; Tripaldi, P.; Pis Diez, R., Quantitative structure–property relationship analysis for the retention index of fragrance-like compounds on a polar stationary phase. J. Chromatogr. A 2015, 1422, 277–288.
Rücker, C., Rücker, G., Meringer, M., 2007. Y-randomization and its variants in QSPR/QSAR. J. Chem. Inf. Model. 47, 2345–2357.
Santos, S.R.L., Silva, V.S., Melo, M.A., Barbosa, J.D.F., Santos, R.L.C., Sousa, D.P., Cavalcanti, S.C.H., 2010. Toxic effects on and structure-toxicity relationships of phenylpropanoids, terpenes and related compounds in *Aedes aegypti* larvae. Vector-Borne Zoonotic Dis. 10, 1049–1054.
Santos, S.R.L., Melo, M.A., Cardoso, A.V., Santos, R.L.C., Sousa, D.P., Cavalcanti, S.C.H., 2011. Structure–activity relationships of larvicidal monoterpenes and derivatives against *Aedes aegypti* Linn. Chemosphere 84, 150–153.
Scotti, L., Scotti, M.T., Silva, V.B., Santos, S.R.L., Cavalcanti, S.C.H., Mendonça Jr., F.J.B., 2014. Chemometric studies on potential larvicidal compounds against *Aedes aegypti*. Med. Chem. 10, 201–210.
Song, J., Wang, Z., Findlater, A., Han, Z., Jiang, Z., Chen, J., Zheng, W., Hyde, S., 2013. Terpenoid mosquito repellents: a combined DFT and QSAR study. Med. Chem. Lett. 23, 1245–1248.
Talete srl, d. Dragon (Software for Molecular Descriptor Calculation) Version 6.0–2014. URL. http://www.talete.mi.it.
WHO, 2006. Pesticides and their application for the control of vectors and pests of public health importance. 6th ed. World Health Organization, Department of Control of Neglected Tropical Diseases, Pesticides Evaluation Scheme, Geneva URL. http://whqlibdoc.who.int/hq/2006/WHO_CDS_NTD_WHOPES_GCDPP_2006.1_eng.pdf (Consulting date 5 of February 2017).
WHO, 2009. Temephos in Drinking-water: Use for Vector Control in Drinking-water Sources and Containers. World Health Organization, Background document for development of WHO Guidelines for Drinking-water Quality, Geneva URL. http://www.who.int/water_sanitation_health/dwq/chemicals/temephos.pdf (Consulting date 5 of February 2017).
World Health Organization (WHO). Number of Reported Cases of Chikungunya Fever in the Americas. by Country or Territory 2013–2015. Epidemiological Week/EW 4 (Updated as of 30 January 2015). URL (http://www.paho.org/chikungunya) (Consulting date: 05 of February 2015).
World Health Organization (WHO), d. Emergencies preparedness response. Zika virus infection — Brazil and Colombia. Epidemiological Week/EW 41 (Update as of 21 of October 2015). URL. http://www.who.int/csr/don/21-october-2015-zika/en/ (Consulting date: 23 of October 2015).
Xiao, Y., Yu, J., 2012. Partitive clustering (K-means family). WIREs Data Mining Knowl Discov. 2, 209–225.
Zhou, L.; Wang, J.; Wang, K.; Xu, J.; Zhao, J.; Shan, T.; Luo, C., Secondary Metabolites with Antinematodal Activity from Higher Plants. Studies in Natural Products Chemistry. Bioactive Natural Products. Atta-ur-Rahman F.R.S., Eds.; Elsevier Press: Oxford. 2012, p. 67–114.