Contents lists available at ScienceDirect

Chemosphere

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

Prediction of acute toxicity of phenol derivatives using multiple linear regression approach for *Tetrahymena pyriformis* contaminant identification in a median-size database

Karel Dieguez-Santana [a, *, 1], Hai Pham-The [b], Pedro J. Villegas-Aguilar [c], Huong Le-Thi-Thu [d], Juan A. Castillo-Garit [e], Gerardo M. Casañola-Martin [a, b, f, **, 1]

[a] *Universidad Estatal Amazónica, Facultad de Ingeniería Ambiental, Paso Lateral Km 21/2 Via Napo, Puyo, Ecuador*
[b] *Hanoi University of Pharmacy, 13-15 Le Thanh Tong, Hoan Kiem, Hanoi, Viet Nam*
[c] *CUBEL Consultancy, 375, Baron Bliss Street, Benque Viejo del Carmen, Cayo District, Belize*
[d] *School of Medicine and Pharmacy, Vietnam National University, Hanoi (VNU) 144 Xuan Thuy, Cau Giay, Hanoi, Viet Nam*
[e] *Unidad de Toxicologia Experimental, Universidad de Ciencias Médicas Dr. Serafin Ruiz de Zárate Ruiz Santa Clara, 50200, Villa Clara, Cuba*
[f] *Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, Spain*

HIGHLIGHTS

- An enlarged data of 358 phenol derivatives against *T. pyriformis* overcoming previous datasets.
- A median-size database of nearly 8000 ChEMBl phenolic compounds was evaluated with the QSTR model.
- Some clues (SARs) for identification of ecotoxicological compounds with acute toxicity profiles.

ARTICLE INFO

Article history:
Received 26 July 2016
Received in revised form
10 September 2016
Accepted 12 September 2016
Available online 30 September 2016

Handling Editor: I. Cousins

Keywords:
QSTR
ChEMBL
Phenol
Tetrahymena pyriformis
Multiple linear regression
Dragon descriptor

ABSTRACT

In this article, the modeling of inhibitory grown activity against *Tetrahymena pyriformis* is described. The 0-2D Dragon descriptors based on structural aspects to gain some knowledge of factors influencing aquatic toxicity are mainly used. Besides, it is done by some enlarged data of phenol derivatives described for the first time and composed of 358 chemicals. It overcomes the previous datasets with about one hundred compounds. Moreover, the results of the model evaluation by the parameters in the training, prediction and validation give adequate results comparable with those of the previous works. The more influential descriptors included in the model are: X3A, MWC02, MWC10 and piPC03 with positive contributions to the dependent variable; and MWC09, piPC02 and TPC with negative contributions. In a next step, a median-size database of nearly 8000 phenolic compounds extracted from ChEMBL was evaluated with the quantitative-structure toxicity relationship (QSTR) model developed providing some clues (SARs) for identification of ecotoxicological compounds. The outcome of this report is very useful to screen chemical databases for finding the compounds responsible of aquatic contamination in the biomarker used in the current work.

1. Introduction

Phenol derivatives commonly exist in the environment. These compounds are used as components of dyes, polymers, pharmaceuticals and other organic substances. The presence of phenols in ecosystems is also related to the production and degradation of many pesticides, industrial waste generation and municipal wastewater. Some phenols are also formed during the natural processes (Michałowicz and Duda, 2007).

* Corresponding author.
** Corresponding author. Universidad Estatal Amazónica, Facultad de Ingeniería Ambiental, Paso Lateral Km 21/2 Via Napo, Puyo, Ecuador.
E-mail addresses: karel.dieguez.santana@gmail.com (K. Dieguez-Santana), gmaikelc@gmail.com, gerardo.casanola@uv.es (G.M. Casañola-Martin).

[1] Both authors contribute equally to the paper.

http://dx.doi.org/10.1016/j.chemosphere.2016.09.041
0045-6535/© 2016 Elsevier Ltd. All rights reserved.

In this sense, the phenolic compounds are considered as dangerous pollutants, which produces serious environmental problems by pollution of water streams because of their great water solubility and high toxicity (Mollaei et al., 2010). This type of chemicals can affect the microflora and fauna of the aquatic environment in a very low concentration of 5 mg/L and they are lethal to fish in 5–25 ppm concentration (dos Santos et al., 2009). Human exposure to these compounds causes critical damage to health and possible risks of carcinogenesis (Nuhoglu and Yalcin, 2005; El-Naas et al., 2009). Therefore, it is vital to protect the environment and prevent occupational poisoning by studying the aquatic toxicity of this family of phenols.

One of the toxicity tests used to determine the aquatic environmental impact is an assay based on the concentration of growth inhibition (IGC_{50}) to *Tetrahymena pyriformis* ciliated freshwater. It is considered appropriate for toxicological testing and safety assessment of chemical components (Cronin et al., 2002).

The experimental tests provide most reliable data on the effects of chemicals, but they involve much time consumption and extensive resources, which makes it difficult to research great numbers of potential toxic compounds. In recent years, the predictions from computer models have been widely used in modern toxicological research, as an important alternative for obtaining experimental evidence and play an important role in evaluating the toxicity of chemicals (Nicolau et al., 2004).

In this sense the QSTR (Quantitative Structure-Toxicity Relationship) models emerge as powerful tools in predictive ecotoxicology, and applied, as scientifically credible tools to predict the acute toxicity of chemicals when there are few empirical data. In the development of a QSAR-based ecotoxicity, integration of subjects (biology, chemistry, and statistics) has allowed the development of structure-toxicity relationships as a subdiscipline accepted in toxicology (McKinney et al., 2000).

Therefore, there is a constant need for development of reliable methods that allow the prediction of computational aquatic toxicity in chemicals. In previous investigations, several QSTR models based on multiple linear regressions (MLR) have been proposed by various research groups to predict the toxicity of phenolic compounds (Roy and Ghosh, 2004; Castillo-Garit et al., 2008; Bellifa and Mekelleche, in press; Ghamali et al., 2015; Singh et al., 2015). Following this aim in this work it is proposed a QSTR model for *Tetrahymena pyriformis* using a chemical wider database. For this, several internal and external validation criteria were applied to the QSTR model developed to ensure robustness, not casual correlation and predictive ability. Furthermore, the results of the QSTR-MLR were compared to those of the previous works to illustrate the advantages of iterative addition of new compounds into the data set which increase the applicability domain of the models by providing a great chemical space of prediction, and hence increasing the prediction potential of the QSTR model.

2. Material and methods

2.1. Experimental data and descriptor calculation

The general dataset used in this study is based on aquatic toxicity tests with *Tetrahymena pyriformis* as biomarker. This dataset is assembled using diverse families of phenol derivatives previously published by other researchers, (Cronin et al., 2001; Cronin and Schultz, 2001; Aptula et al., 2002; Cronin et al., 2002; Mekapati and Hansch, 2002; Seward et al., 2002; Netzeva et al., 2003; Ren, 2003; Schüürmann et al., 2003; Pasha et al., 2005, 2007; Melagraki et al., 2006). The final dataset has 358 compounds that include phenol and phenolic derivatives, and the SMILES notation for this dataset is giving in Table S1 of Supplementary Material.

For this purpose, seven classes of molecular descriptors of the Dragon program were calculated. They were selected based on its confirmed effectiveness and easy interpretability. These families of structural descriptors are extensively described in item **SI1** of the Supplementary Material. Finally, in our case, more than 447 structural descriptors were computed for the 358 phenol derivatives.

2.2. Design of training and prediction set

In our case, in order to design the training, validation and prediction series to guarantee structural and toxicity variability in these three series, it was carried out the two types of cluster analyses (k-MCA and k-NNCA) for the whole dataset of compounds (STATISTICA, 2007). The number of members in every cluster and the standard deviation of the variables in the cluster (kept as low as possible) were taken into account to have an acceptable statistical quality of data partition into clusters.

The database was split into training, validation and prediction series in order to perform the horizontal validation. Thus, a k-means cluster analysis (k-MCA) was carried out for the entire data set to design in a rational representative way, the training (learning) validation(calibration) and prediction series using the STATISTICA software 8.0. (STATISTICA, 2007).

Before carrying out the cluster processes, all the molecular descriptors were substituted by their standardized values which are computed as follows: Std. core = (raw_score − mean)/Std.deviation. The number and members in each cluster and the standard deviation to the variables in the cluster (as low as possible) were considered in order to guarantee acceptable statistical quality of data cluster. In addition, the standard deviation between and within cluster, the respective Fisher ratio and p-level of significance ($p < 0.05$) were examined. The selection of the training and prediction sets was executed by randomly taking compounds which belong to each chemical class (as determined by clusters). This procedure contributes to select in a usual way in the whole level of the linking distance (Y-axis), and compounds for the three subsets.

Finally, the training, validation and prediction sets were composed by 240, 78 and 40 compounds, respectively (the last two series representing around 33% of the complete database), respectively. Compounds, belonging to the calibration and prediction sets, were never used in the development of the regression functions and they were set aside to evaluate the predictability of obtained QSAR models.

2.3. MLR technique for model development

The modeling technique selected was the Multiple Linear Regression (MLR). In this case, the regression coefficients and statistical parameters were obtained by this regression-based approach. The software selected for the development of the QSTR model was the STATISTICA (STATISTICA, 2007). The considered tolerance parameter for minimum acceptable tolerance was the default value of 0.01. The forward stepwise procedure was the strategy for variable selection. The principle of parsimony (Occam's razor) was taken into account at time of model selection. Therefore, the model with the highest statistical signification, but having as few parameters (a_k) as possible was selected. The log (1/IGC50) (decimal logarithm of the inverse 50 percent growth inhibitory concentration) values were used as the dependent variable, where concentration is described as mmol/L.

A single MLR model was developed for phenolic compounds using the Statistic software (STATISTICA, 2007). The multiple linear regression model was built using a training set and validation process was done with the two external prediction sets. The stepwise regression was applied in the selection of the significant variables to be included in the QSTR-MLR model.

2.4. Model validation and applications

2.4.1. Parameter for model internal validation

The quality of the models was determined by examining the regression's statistical parameters and those of the cross-validation procedures (Wold and Erikson, 1995). Therefore, the following parameters were verified: the determination coefficient (R^2), Fisher's ratio (F) and the standard error in calculation (s).

On the other hand, the robustness of model refers to the stability of its parameters (predictor coefficients). Consequently, the stability of its predictions, when a perturbation (deletion of one or more chemicals) is applied to the training set, and the model is regenerated from the "perturbed" training set. Therefore, bootstrapping (BOOT) and Y-scrambling were the procedures used to the assessment of the internal validity of models obtained by multivariate regression methods. Specifically, the cross-validated determination coefficient calculated in BOOT (q^2_{BOOT}) strategies was used to evaluate the robustness and stability of the linear regression equations, together with the standard error in prediction (SDEP); the parameters a(R^2) and a(q^2) estimated in a Y-randomization experiment were also calculated to test the absence of chance correlation.

2.4.2. Model external validation

However, while the internal validation techniques described above can be used to establish model robustness, they do not directly assess model predictability. One of the main steps involved in a QSAR study consists in the statistical validation of the results to determine its reliability and significance, while providing an indication of how well the model can predict activity for new molecules. Several procedures are available for this task, and they were carried out for internal and external ones, in this report.

Validation external process is necessary to ensure the quality and predictive power of the QSAR models to predict the activity of compounds that were not used to the model development. In this study, the original data were divided into three series, TS, PS and ES by "*rational*" design according to the principles stood out in previous works (Golbraikh et al., 2003). The TS is used to build the QSAR models, and these discriminant functions (DFs) are used to predict the activities of compounds in the PS and ES. The predictabilityof a model is estimated by comparing the predicted and observed classes of a sufficiently large and representative test of compounds.

2.5. QSTR-MRL model applicability domain

The applicability domain (AD) of a QSAR model is the response and chemical structure space in which the model makes prediction with a given reliability (Netzeva et al., 2005). Through the leverage approach (Atkinson, 1985) (shown below), it is possible to verify whether a new chemical will lie within the structural model domain.

In order to visualize the AD of the QSTR model, the plot of standardized cross-validated residuals (R) versus leverage (Hat diagonal) values (h) (the Williams plot) can be used to an immediate and simple graphical detection of both the response outliers (i.e., compounds with standardized residuals greater than three standard deviation units, $>3s$) and structurally influential chemicals in mode ($h > h^*$) (Gramatica, 2007).

3. Results and discussion

3.1. Design of training validation and external subsets

As it was mentioned above, the assesment of any QSAR model depends on the quality of the selected data set, but one of the most critical aspects is to warrant enough molecular diversity for the construction of the training set. Firs, a hierarchical CA was performed of the entire dataset to demonstrate its structural diversity (Mc Farland and Gans, 1995.). The dendrogram (binary tree) given in Fig. S1 (see Supplementary Material) was done using the Euclidean distance (X-axis) and the complete linkage (Y-axis) as grouping algorithm, as a result of the k-NNCA developed for the complete dataset of 358 compounds. As it is shown in the same Fig. S1, there are a number of different subsets, which proves the molecular variability of the selected chemicals in the database. The horizontal line that go through all the dendrogram delimitated the most suitable number of clusters which are assembled for this dataset. At this time, eigth clusters were selected as the quantity that ensures the maximal variance between the groups and minimal variance inside the clusters.

As the difficulty in evaluating the output dendrogram, other kind of CAs is usually performed to verify the molecular variability in the data of compounds. Therefore, it was performed a k-MCA with the objective of spliting the whole group into three data sets (training, predicting and validation ones). The main idea of this procedure consists of making a partition of the chemicals into several statistically representative classes of compounds. This procedure ensures that any chemical class (as determined by the clusters derived from k-MCA) will be represented in both compounds' series. This procedure makes possible the distribution of the dataset of phenolic compounds into the eight clusters previously selected by the k-NNCA. These clusters have 47, 57, 34, 62, 54, 16, 61 and 27 compounds respectively, by bundling a total of 358 phenolic derivates. From these 358 compounds, 240 compounds were chosen as the training set (TS). The remaining subset with 118 compounds was used as the test set for validation of the models. The prediction set (PS) with 78 chemicals and the validation set (VS) with 40 phenols. These compounds were never used in the development of the QSAR models.

3.2. Development of QSTR model to predict growth inhibitory concentration

The MLR analysis was used to develop a QSTR model for the prediction of aquatic toxicity against *T. pyriformis*. The quality of the QSTR-MLR model was determined by examining the statistical parameters of the regression in cross-validation procedures and in the number of SP and VS datasets (See Table 1).

From this model, N is the size of the data set; R, the correlation

Table 1
Performance model QSTR-MLR.

N (TS/PS/ES)	R^2	Q^2_{train}	SDEP	q^2_{BOOT}	a(R^2)	a(q^2)	Q^2_{pred}	Q^2_{ext}
240/78/40	0.7404	0.7019	0.4550	0.6858	0.0320	−0.1030	0.6999	0.4585

N*: Number of compounds in the training set (TS) prediction (PS) and external (ES).*

Table 2
Symbols and definitions of the molecular descriptors in the QSTR-MLR model.

Descriptors	Definition	Descriptor family	Sign value
MWC09	molecular walk count of order 9	Walk and path counts	−
piPC02	molecular multiple path count of order 2	Walk and path counts	−
TPC.	total path count	Walk and path counts	−
nCconj	number of non-aromatic conjugated C(sp^2)	Functional group counts	−
nRCN	number of nitriles (aliphatic)	Functional group counts	−
nCXr =	number of X on ring C(sp^2)	Functional group counts	−
O-059	Al-O-Al	Atom-centred fragments	−
BLTD48	Verhaar Daphnia base-line toxicity from MLOGP (mmol/L)	Molecular properties	−
MWC02	molecular walk count of order 2	Walk and path counts	+
MWC10	molecular walk count of order 10	Walk and path counts	+
piPC03	molecular multiple path count of order 3	Walk and path counts	+
piPC08	molecular multiple path count of order 8	Walk and path counts	+
X3A,	average connectivity index of order 3	Connectivity indices	+
nR = Cs	number of aliphatic secondary C(sp^2)	Functional group counts	+

coefficient; R^2, the determination coefficient; F, the Fisher-ratio; *s*, the standard deviation of the regression; SDEP, the standard error in prediction; q^2_{BOOT}, the cross-validated determination coefficient calculated for bootstrapping experiment (this method generally gives the most accurate estimates of model performance in terms of "internal predictability") (Gramatica et al., 2007); coefficients a(R^2) and a(q^2) were estimated in the Y- randomization experiment; and R^2_{pred} is the square correlation coefficient for the external set and the parameters.

The model gives a squared regression coefficient (R^2) value of 0.7404, it explains more than 74% of the experimental variance aquatic toxicity, using *Tetrahymena pyriformis* (log 1/ICG50) and a standard deviation of 0.439 with F = 45.91 and P = 0.001. Other statistical parameters related with robustness and predictability are also suitable which demonstrate the goodness-of-fit of the obtained equation and they will be discussed in following subsections.s

The linear relationship between toxicity of the phenolic compounds (represented by value of log(1/IGC_{50})) and fourteen 0-2D Dragon descriptors is shown in Equation (1).

$$\begin{aligned}\mathrm{Log}(1/\mathrm{IGC}_{50}) &= -17.537 + 12.384\ast\mathrm{MWC02} - 16.530\ast\mathrm{MWC09}\\ &\quad + 12.794\ast\mathrm{MWC10} - 8.264\ast\mathrm{piPC02}\\ &\quad + 9.150\ast\mathrm{piPC03} + 0.070\ast\mathrm{piPC08} - 4.408\ast\mathrm{TPC}\\ &\quad + 24.910\ \mathrm{X3A} - 0.083\ \mathrm{nCconj} + 0.743\ast\mathrm{nR}\\ &= \mathrm{Cs} - 0.487\ast\mathrm{nRCN} - 0.624\ast\mathrm{nCXr}\\ &= -0.505\ast\mathrm{O} - 059 - 0.877\ \mathrm{BLTD4}\end{aligned} \tag{1}$$

$$N = 240 \quad R^2 = 0.74 \quad s = 0.439 \quad F = 45.91 \quad p < 0.0001$$

In Table 2, the definitions of the molecular descriptors included in the current study for the development of the QSTR-MLR model are provided.

There is a useful aspect that should be highlighted, the analysis of the factors that are likely to govern the log(1/IGC50) of the compounds by the descriptor interpretation in the regression model. Therefore, it is useful to link the characteristics of compounds for describing the relationship between the structure and the acute toxicity in the regression analysis. In this study, the descriptors used in the model are shown in Table 2 together with the description of the family. In the model, the descriptors: X3A, MWC02, MWC10, piPC03, piPC08, and nR = Cs have a positive relationship to the acute toxicity of phenolic compound (see last column of Table 2) and the first four ones have the highest contribution to the dependent variable given by the coefficient values.

In the case of MWC10 and MWC02, they have a positive relationship with the acute toxicity of phenols derivatives, belonging to the walk and path count which is a topological index, based on the counting of paths, molecular walks and self-returning walks in an H-depleted graph (Ruecker and Ruecker, 1993). The MWC10 and MWC02 are molecular walk count of order 10 and 2 respectively, which are related to molecular branching and size, as well as the molecular complexity of the graph (i.e. the larger size and more complex molecule, the larger MWC10 and MWC02 values.). However, it cannot be proved whether the phenolic compounds toxicity increased with the complexity of the molecule, as in the model the MWC09 descriptor has negative contribution to the inhibitory concentration growth; the same thing occurs to other multiple path count molecular descriptors for which the impact on the model is not clearly defined.

In Equation (1), other functional group count descriptors have negative influence on the toxicity: nCconj number of non-aromatic conjugated C (sp^2), nRCN number of nitriles (aliphatic), nCXr = number of X on ring C (sp^2), TPC total count of the Walk path and path counts descriptors, O-059 O-Al-O-Al Atom-centered fragments and descriptor of the molecular properties BLTD48-line basis Verhaar Daphnia toxicity from MLOGP (mmol/L), which the toxicity could decrease with the increase of these structures in phenolic compounds. At the same way, it should be highlighted that molecular walk count descriptor of order 9 (MWC09) and molecular multiple path count of order 2 (piPC02) have the highest negative coefficient as the highest contribution to the decrease of the inhibitory concentration growth.

3.3. Validation of the toxicity-based QSAR models

The main importance of the horizontal validation is to prove the predictability and the robustness of the model. In this report, a cross-validation was performed in both internal (procedures leave-one-out and bootstrapping) and external (using a test set) validation experiments only for the final models obtained with Dragon Descriptors 0-2D (Equation (1)).

The variance explained for the Dragon 0-2D descriptors in model (Equation (1)) to the BOOT procedure was higher than 68% [$q^2_{BOOT} = 0.6858$]; besides, the value of SDEP was 0.455. According to the criteria of several authors these results can be interpreted regarding the robustness and stability of the models (Belsey et al., 1980; Wold and Erikson, 1995). In addition, the Y-scrambling parameters for this model [a(R^2) = 0.032 and a(q^2) = −0.103] showed low values, indicating that there is not a significant difference in the

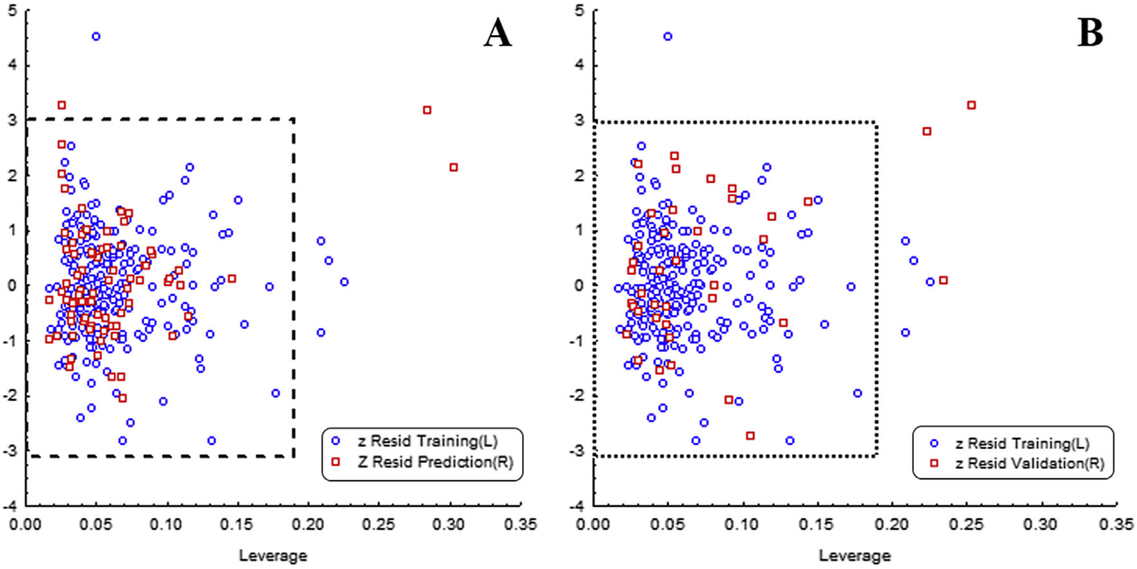


Fig. 1. Applicability domain of the QSTR MLR model. **A**. Training and prediction sets. **B**. Training and external validation sets.

Table 3
Comparison of MLR models for log (1/IGC50) *T. pyriformis* of phenol derivatives.

Index[a]	N[b]	n[c]	R[2d]	s[e]	Statistical method[f]	Ref[g]
Dragon descriptors 0-2D (current work)	**358**	**14**	**0.74**	**0.44**	**MLR**	Equation 1
Molconn-Z descriptors	250	6	0.69	0.49	MLR	(Jiang et al., 2011)
Molconn-Z descriptors	250	10	0.78	–	PLS	(Jiang et al., 2011)
SMILES-based optimal descriptors	250	1	0.77	0.41	MCOA	(Toropov et al., 2010)
Quantum topological molecular indices	17	5	0.99	–	PLS	(Hemmateenejad et al., 2010)
Quantum topological molecular indices	**17**	**5**	**0.99**	**-**	**GA-PLS**	(Hemmateenejad et al., 2010)
Simple molecular descriptors	221	6	0.98	–	RF	(Chen et al., 2012)
Z-matrices (Dragon MDs)	250	6	0.75	–	GA-MLR	(Habibi-Yangjeh and Danandeh-Jenagharad, 2009)
Z-matrices (Dragon MDs)	**250**	**6**	**0.93**	**-**	**GA-ANN**	(Habibi-Yangjeh and Danandeh-Jenagharad, 2009)
Dragon MDs, Quantum-Chemical	250	7	0.85	0.44	RM	(Duchowicz et al., 2008)
Molecular weight PharmaAlgorithms	207	3	0.84	0.33	PLS	(Zhao et al., 2009)
Physico-chemical	250	4	0.66	–	MLR	(Enoch et al., 2008)
Physico-chemical	250	4	0.71	–	ANN	(Enoch et al., 2008)
Quantum chemical	43	4	0.92	0.21	MLR	(He et al., 2012)
Quantum chemical	43	–	0.94	0.19	SVM	(He et al., 2012)
Quantum chemical	97	3	0.90	0.29	RA	(Roy et al., 2006)
Physicochemical and structural features.	250	10	0.87	–	PLS	(Devillers, 2004)
Physicochemical and structural features.	250	9	0.91	–	ANN	(Devillers, 2004)
Quantitative neighbourhoods of atoms (QNA)	200	12	0.69	0.49	SCR	(Lagunin et al., 2007)

[a] Index: Molecular descriptors for the described study.
[b] N: number de compounds.
[c] n: number of parameters in the model.
[d] R^2:determination coefficient.
[e] s: standard deviation of the regression.
[f] Statistical Method: Description of statistical Method, GA – genetic algorithm, PLS – Partial Least Squares Regression, MLR – Multiple Linear Regression, ANN – artificial neural network, RA – regression analysis, SVM – Support vector machine, MCOA – Monte Carlo optimization Algorithm, SCR – Self-consistent regression, RF – Random Forest, RM – Replacement Method.
[g] Reference (Author, Year).

quality of the original model and that one associated with models obtained with random responses. This suggests that the original models have no chance correlation.

Two sets composed of 78 and 40 compounds were used as prediction and external validation series to judge the predictability of the QSTR-MLR model. In this case, the equation showed a coefficient determination for this two sets of $Q^2_{pred} = 0.6999$ and $Q^2_{ext} = 0.4568$, respectively. This prediction value is suitable for the prediction set, but it should be carefully considered for external validation set of compound. Therefore, this model has be used carefully for the prediction of external chemical or should be retrained to include those compounds of the external set for the improvement of the predictions.

Similarly Figs. S2 and S3 (see Supplementary Information) showed the results of the observed vs predicted values for the training and prediction sets respectively. This is another way to assess the quality of QSTR-MLR model developed. Therefore, there is an adequate agreement between the observed and predicted values for the training and prediction sets.

3.4. Applicability domain of the QSTR-MLR model

Another main problem in chemometric and QSAR studies is the definition of the applicability domain (AD) of a classification or

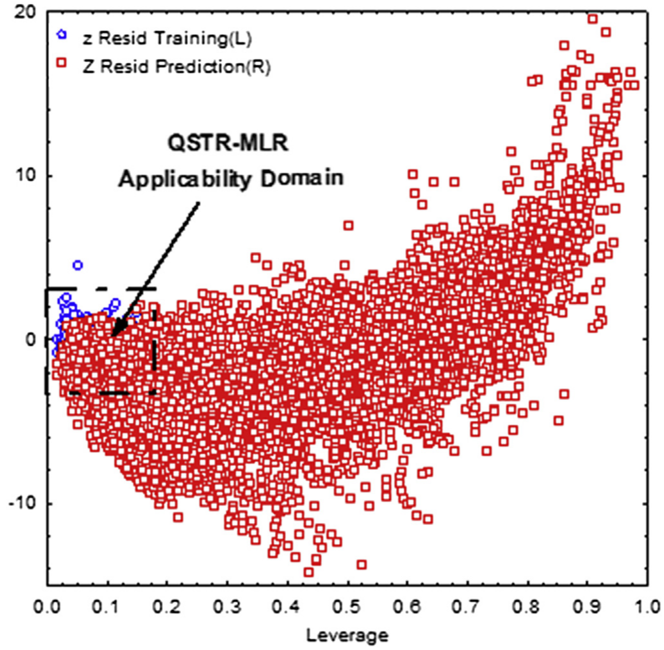


Fig. 2. Applicability domain of the ChEMBL phenols dataset.

regression model. Therefore, the next step of this report was to develop a study to access about chemical scope of the model. In this work, the leverage values (h) and standardized residuals (Std. Res. or σ) were calculated for the QSTR-MLR model to determinate the AD, using the William's plot approach as showed in Fig. 1. For the construction of this graphic, the leverage values (h) were used in the *x* axis and the standardized residual in the *y* axis. The limits of the AD were established leverage threshold $h^* = 0.1875$ and the ±3 standard deviation values. In Fig. 1, only five chemicals in training set are outside the AD, representing the 2.08% of the training data as an adequate considered value. For the case of the prediction and external validation, three and four compounds are outside this area respectively. Totally (including training, prediction and external validation) five compounds have standard deviation values outside the ±3 standard deviation values; therefore only these compounds could be considered as *outliers*. The remaining seven cases with leverage values greater than h^*, should be considered as influential compounds.

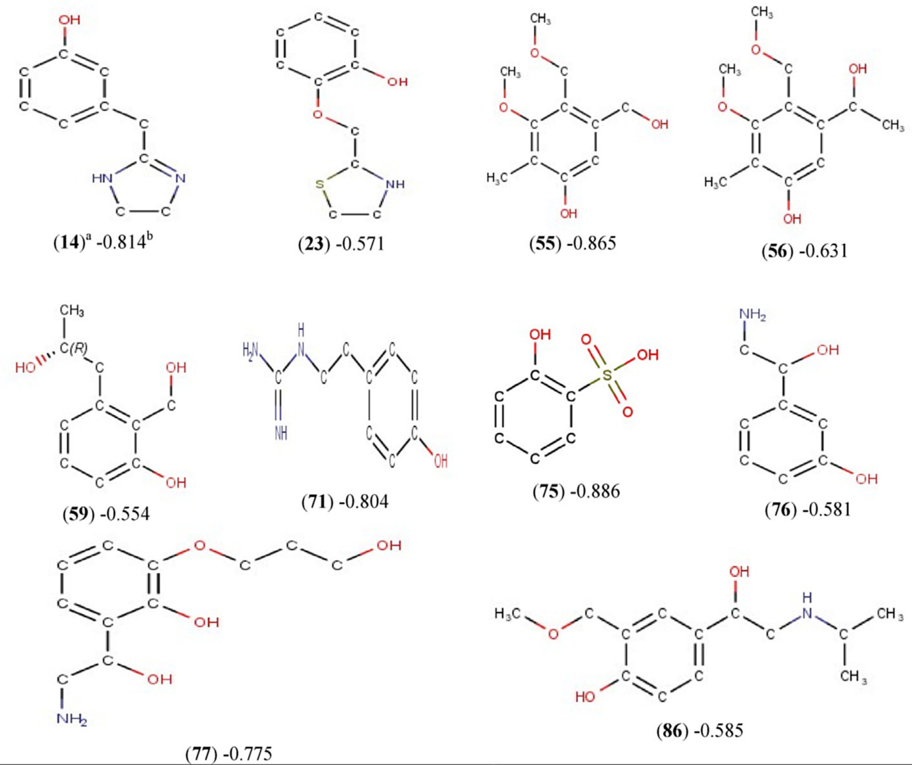
The figure above is useful for the recognition of influential compounds (leverages above threshold) and outliers in the QSTR-MLR model. This could be used for the preliminary evaluation of compounds for ensuring accurate predictions in the chemicals submitted to the ecotoxicological potential estimation.

3.5. Comparison with other approaches

The use of Dragon 0-2D descriptors were compared to other reports previously described in the literature for the prediction of aquatic toxicity of phenol derivatives against *T. Pyriformis*. Table 3 summarizes the statistics parameters of the QSTR-MLR model in the current study and those of other researchers also describing the growth inhibition of *Tetrahymena Pyriformis* of phenols.

An analysis of the results of Table 3 shows that this work has the

Table 4
Top ten molecules with the lowest log(1/IGC_{50}) value predicted by the QSTR-MLR model.


(14)[a] -0.814[b]

(23) -0.571

(55) -0.865

(56) -0.631

(59) -0.554

(71) -0.804

(75) -0.886

(76) -0.581

(77) -0.775

(86) -0.585

[a]Number of the compound in the 95 negative values dataset. [b]Predicted values of log(1/IGC_{50}) by the QSTR-MLR model.

greatest dataset (in number of compounds). Although, the model presented in this report has a large number of parameters (14) when compared with models that have twelve as number of parameters (Lagunin et al., 2007). It is the one that shows the better model performance with 0.74. It overcomes the previously reported with a $q^2 = 0.685$. Then, these results are compared to a model developed with Molconn-Z descriptors in a dataset of 250 compounds (Jiang et al., 2011) where a better goodness of fit is found, but with PLS as statistical technique is more powerful than MLR. Other models with the same data size (250 chemicals) have better performances, but in the majority of cases, more complex techniques or descriptors are used (Devillers, 2004; Duchowicz et al., 2008; Enoch et al., 2008; Habibi-Yangjeh and Danandeh-Jenagharad, 2009; Toropov et al., 2010). The same occurs for a model done with Random Forest, a machine learning technique, in a 221 phenol derivatives data (Chen et al., 2012) and another using PLS for 207 phenols (Zhao et al., 2009). The following reports (Roy et al., 2006; Hemmateenejad et al., 2010; He et al., 2012) have $q^2 > 0.90$, but small datasets below 97 compounds are used in all cases, representing a constrained chemical space at a time to perform virtual screening to detect possible aquatic contaminants for the biomarker of the present study.

3.6. Evaluation of ecotoxicological potential in a ChEMBL phenol dataset

The main idea of any developed model is its use. Taking into consideration this aim, a dataset composed of 7842 curated compounds downloaded from the ChEMBL dataset was evaluated by the QSTR-MLR model. Before the in-silico ecotoxicological activity prediction, the AD evaluation for this set of phenols was carried out. For this analysis all the compounds, with a leverage value above the leverage threshold were discarded, as well as those with z-residual outside the ±3s range. For doing this, only 600 compounds were inside the AD of the QSTR-model as can be shown in Fig. 2.

Later, the compounds were evaluated in the regression model and from this, 500 chemicals have predicted values of $\log(1/IGC_{50})$ above zero, indicating a low acute toxicity in *Tetrahymena pyriformis*. The remaining 100 compounds have $\log(1/IGC_{50})$ values below zero, i.e. high toxicities and hence they could have high aquatic contaminant profiles. An examination in this subset allows the detection of five stereoisomers, sharing the same structure that other compounds in this dataset of 100 chemicals. Therefore, one of the stereoisomer was discarded because they are only used 0-2D MDs that do not accout chiral or tridimensional features, and after that the dataset remain with 95 potent ecotoxicological phenols.

In a next step, the top ten compounds with the lowest predicted values of $\log(1/IGC_{50})$ were selected (see Table 4). In a closer inspection to the chemicals in the table above referred, it can be visualized that the majority of the structures share some common substructural features that could be associated with the high acute toxicity and some structure-toxicity relationships which could be detected.

For example, the presence of nitriles (compound 14) could be associated with the presence of RCN descriptor in the model. The same occurs for the O-059 descriptors related with Al-O-Al fragments that appears in compounds 55, 56, 77 and 86.

Finally, an interesting structure alert appears related to the introduction of a methyl group in compound 56 with regard to compound 55. In this case an increase of the value of $\log(1/IGC_{50})$ occurs, showing a decreasing of the acute toxicity of compound 56 in comparison to compound 55. Therefore, this structure-toxicity relationship could be used together with the other identified structural features for the design of safer compounds in aquatic toxicity environment in the biomarker *Tetrahymena pyriformis*.

4. Conclusions

In this study, the MLR technique was used to develop a linear QSTR model for prediction of phenols toxicity to *Tetrahymena pyriformis*. Chemical descriptors derived from molecular structures were calculated with Dragon software. The obtained QSTR-MLR model was statistically significant, robust and with positive values of $R^2 = 0.74$ and $q^2 = 0.69$ in the training, and an adequate R^2 predictive value of 0.70, indicating the capability of predicting the aquatic toxicity of phenol derivatives in the impairment of the population growth of *T. pyriformis*.

In addition, the outcomes of the present report showed similar performance to those of previous studies with smaller datasets. In this sense, it should be highlighted the increase of the data size in the present report, as well as, the applicability domain of the model developed. An example of this applicability is provide by the evaluation of a ChEMBL phenol dataset with the detection of 95 potential aquatic contaminants and the examination of some structural alerts related to the molecular descriptors included in the QSTR-MLR which is developed in this report. Finally, these type of predictive QSTR models is an alternative to the replacement of *in-vivo* or *in-vitro* assays.

Conflict of interest

The authors confirm that this article content has no conflicts of interest.

Acknowledgements

Le-Thi-Thu, H. gratefully acknowledges support from the National Vietnam National University, Hanoi. Casañola-Martin, G.M. and Dieguez-Santana, K. thanks to internal project from the Universidad Estatal Amazonica.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.chemosphere.2016.09.041.

References

Aptula, A.O., Netzeva, T.I., Valkova, I.V., Cronin, M.T.D., Schultz, T.W., Kühne, R., Schüürmann, G., 2002. Multivariate discrimination between modes of toxic action of phenols. Quant. Structure-Activity Relat. 21, 12–22.

Atkinson, A.C., 1985. Plots, Transformations, and Regression: an Introduction to Graphical Methods of Diagnostic Regression Analysis. Clarendon Press.

Bellifa, K., Mekelleche, S.M., 2012. QSAR study of the toxicity of nitrobenzenes to Tetrahymena pyriformis using quantum chemical descriptors. Arabian J. Chem. http://dx.doi.org/10.1016/j.arabjc.2012.04.031 (in press).

Belsey, D.A., Kuh Page, E., Welsch, R.E., 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley, New York.

Castillo-Garit, J.A., Marrero-Ponce, Y., Escobar, J., Torrens, F., Rotondo, R., 2008. A novel approach to predict aquatic toxicity from molecular structure. Chemosphere 73, 415–427.

Chen, J., Tang, Y.Y., Fang, B., Guo, C., 2012. In silico prediction of toxic action mechanisms of phenols for imbalanced data with random forest learner. J. Mol. Graph. Model. 35, 21–27.

Cronin, M.T.D., Aptula, A.O., Duffy, J.C., Netzeva, T.I., Rowe, P.H., Valkova, I.V., Wayne Schultz, T., 2002. Comparative assessment of methods to develop QSARs for the prediction of the toxicity of phenols to Tetrahymena pyriformis. Chemosphere 49, 1201–1221.

Cronin, M.T.D., Manga, N., Seward, J.R., Sinks, G.D., Schultz, T.W., 2001. Parametrization of electrophilicity for the prediction of the toxicity of aromatic compounds. Chem. Res. Toxicol. 14, 1498–1505.

Cronin, M.T.D., Schultz, T.W., 2001. Development of quantitative Structure–Activity relationships for the toxicity of aromatic compounds to Tetrahymena pyriformis: comparative assessment of the methodologies. Chem. Res. Toxicol. 14, 1284–1295.

Devillers, J., 2004. Linear versus nonlinear QSAR modeling of the toxicity of phenol derivatives to Tetrahymena pyriformis. SAR QSAR Environ. Res. 15, 237–249.

dos Santos, V.L., Monteiro, A.D.S., Braga, D.T., Santoro, M.M., 2009. Phenol degradation by Aureobasidium pullulans FE13 isolated from industrial effluents. J. Hazard Mater. 161, 1413–1420.

Duchowicz, P.R., Mercader, A.G., Fernández, F.M., Castro, E.A., 2008. Prediction of aqueous toxicity for heterogeneous phenol derivatives by QSAR. Chemom. Intelligent Laboratory Syst. 90, 97–107.

El-Naas, M.H., Al-Muhtaseb, S.A., Makhlouf, S., 2009. Biodegradation of phenol by Pseudomonas putida immobilized in polyvinyl alcohol (PVA) gel. J. Hazard Mater. 164, 720–725.

Enoch, S.J., Cronin, M.T.D., Schultz, T.W., Madden, J.C., 2008. An evaluation of global QSAR models for the prediction of the toxicity of phenols to Tetrahymena pyriformis. Chemosphere 71, 1225–1232.

Ghamali, M., Chtita, S., Adad, A., Hmamouchi, R., Bouachrine, M., Lakhlifi, T., 2015. Combining DFT and QSAR results for predicting the cytotoxicity of a series of orthoalkyl substituted 4-X-phenols. J. Mater. Environ. Sci. 6, 280–288.

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. Computer-Aided Mol. Des. 17, 241–253.

Gramatica, P., 2007. Principles of QSAR models validation: internal and external. QSAR Comb. Sci. 26, 694–701.

Gramatica, P., Giani, E., Papa, E., 2007. Statistical external validation and consensus modeling: a QSPR case study for Koc prediction. J. Mol. Graph. Model. 25, 755–766.

Habibi-Yangjeh, A., Danandeh-Jenagharad, M., 2009. Application of a genetic algorithm and an artificial neural network for global prediction of the toxicity of phenols to Tetrahymena pyriformis. Monatsh. fur Chem. 140, 1279–1288.

He, G., Feng, L., Chen, H., 2012. A QSAR study of the acute toxicity of halogenated phenols. Procedia Eng. 204–209.

Hemmateenejad, B., Mehdipour, A.R., Miri, R., Shamsipur, M., 2010. Comparative qsar studies on toxicity of phenol derivatives using quantum topological molecular similarity indices. Chem. Biol. Drug Des. 75, 521–531.

Jiang, D.X., Li, Y., Li, J., Wang, G.X., 2011. Prediction of the aquatic toxicity of phenols to tetrahymena pyriformis from molecular descriptors. Int. J. Environ. Res. 5, 923–938.

Lagunin, A.A., Zakharov, A.V., Filimonov, D.A., Poroikov, V.V., 2007. A new approach to QSAR modelling of acute toxicity. SAR QSAR Environ. Res. 18, 285–298.

Mc Farland, J.W., Gans, D.J., 1995. Cluster significance analysis. In: Waterbeemd, H. (Ed.), Chemometric Methods in Molecular Design. VCH Publishers, Winheim, pp. 295–307.

McKinney, J.D., Richard, A., Waller, C., Newman, M.C., Gerberick, F., 2000. The practice of structure activity relationships (SAR) in toxicology. Toxicol. Sci. 56, 8–17.

Mekapati, S.B., Hansch, C., 2002. On the parametrization of the toxicity of organic chemicals to Tetrahymena pyriformis. The problem of establishing a uniform activity. J. Chem. Inf. Comput. Sci. 42, 956–961.

Melagraki, G., Afantitis, A., Sarimveis, H., Igglessi-Markopoulou, O., Alexandridis, A., 2006. A novel RBF neural network training methodology to predict toxicity to Vibrio Fischeri. Mol. Divers. 10, 213–221.

Michałowicz, J., Duda, W., 2007. Phenols – sources and toxicity. Pol. J. Environ. Stud. 16, 347–362.

Mollaei, M., Abdollahpour, S., Atashgahi, S., Abbasi, H., Masoomi, F., Rad, I., Lotfi, A.S., Zahiri, H.S., Vali, H., Noghabi, K.A., 2010. Enhanced phenol degradation by Pseudomonas sp. SA01: gaining insight into the novel single and hybrid immobilizations. J. Hazard Mater. 175, 284–292.

Netzeva, T.I., Aptula, A.O., Chaudary, S.H., Duffy, J.C., Wayne Schultz, T., Schüürmann, G., Cronin, M.T.D., 2003. Structure-activity relationships for the toxicity of substituted poly-hydroxylated benzenes to Tetrahymena pyriformis: influence of free radical formation. QSAR Comb. Sci. 22, 575–582.

Netzeva, T.I., Worth, A.P., Aldenberg, T., Benigni, R., Cronin, M.T., Gramatica, P., Jaworska, J.S., Kahn, S., Klopman, G., Marchant, C.A., 2005. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. Altern. Laboratory Animals 33, 1–19.

Nicolau, A., Mota, M., Lima, N., 2004. Effect of different toxic compounds on ATP content and acid phosphatase activity in axenic cultures of Tetrahymena pyriformis. Ecotoxicol. Environ. Saf. 57, 129–135.

Nuhoglu, A., Yalcin, B., 2005. Modelling of phenol removal in a batch reactor. Process Biochem. 40, 1233–1239.

Pasha, F.A., Srivastava, H.K., Singh, P.P., 2005. Comparative QSAR study of phenol derivatives with the help of density functional theory. Bioorg. Med. Chem. 13, 6823–6829.

Pasha, F.A., Srivastava, H.K., Srivastava, A., Singh, P.P., 2007. QSTR study of small organic molecules against Tetrahymena pyriformis. QSAR Comb. Sci. 26, 69–84.

Ren, S., 2003. Ecotoxicity prediction using mechanism- and non-mechanism-based QSARs: a preliminary study. Chemosphere 53, 1053–1065.

Roy, D.R., Parthasarathi, R., Subramanian, V., Chattaraj, P.K., 2006. An electrophilicity based analysis of toxicity of aromatic compounds towards Tetrahymena pyriformis. QSAR Comb. Sci. 25, 114–122.

Roy, K., Ghosh, G., 2004. QSTR with extended topochemical Atom indices. 3. Toxicity of nitrobenzenes to Tetrahymena pyriformis. QSAR Comb. Sci. 23, 99–108.

Ruecker, G., Ruecker, C., 1993. Counts of all walks as atomic and molecular descriptors. J. Chem. Inf. Comput. Sci. 33, 683–695.

Schüürmann, G., Aptula, A.O., Kühne, R., Ebert, R.-U., 2003. Stepwise discrimination between four modes of toxic action of phenols in the Tetrahymena pyriformis assay. Chem. Res. Toxicol. 16, 974–987.

Seward, J.R., Hamblen, E., Wayne Schultz, T., 2002. Regression comparisons of tetrahymena pyriformis and poecilia reticulata toxicity. Chemosphere 47, 93–101.

Singh, K.P., Gupta, S., Basant, N., 2015. QSTR modeling for predicting aquatic toxicity of pharmacological active compounds in multiple test species for regulatory purpose. Chemosphere 120, 680–689.

STATISTICA, 2007. In: Tulsa, O. (Ed.), Data Analysis Software System. StatSoft, Inc.

Toropov, A.A., Toropova, A.P., Benfenati, E., Manganaro, A., 2010. QSAR modelling of the toxicity to Tetrahymena pyriformis by balance of correlations. Mol. Divers. 14, 821–827.

Wold, S., Erikson, L., 1995. Chemometric methods in molecular design. In: Waterbeemd, H.van.de. (Ed.), Chemometric Methods in Molecular Design. VCH Publishers, Weinheim, Ger., pp. 309–318

Zhao, Y.H., Yuan, X., Su, L.M., Qin, W.C., Abraham, M.H., 2009. Classification of toxicity of phenols to Tetrahymena pyriformis and subsequent derivation of QSARs from hydrophobic, ionization and electronic parameters. Chemosphere 75, 866–871.