Chemosphere 165 (2016) 434–441

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/

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