Ecotoxicology and Environmental Safety 101 (2014) 184–190

Contents lists available at ScienceDirect

Ecotoxicology and Environmental Safety

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

Predictive modeling of chemical toxicity towards *Pseudokirchneriella subcapitata* using regression and classification based approaches

Subrata Pramanik, Kunal Roy*

Drug Theoretics and Cheminformatics Laboratory, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700032, India

ARTICLE INFO

Article history:
Received 21 November 2013
Received in revised form
28 December 2013
Accepted 30 December 2013
Available online 20 January 2014

Keywords:
Biodiversity
Toxicity
QSTR
LDA
Pseudokirchneriella subcapitata

ABSTRACT

Biodiversity nurturing may be a valuable pathway in controlling chemical stress on the ecosystem. In the present work, *in silico* studies have been performed to develop regression based quantitative structure toxicity relationship (QSTR) models using a data set containing 105 organic chemicals for the prediction of 48-h chemical toxicity towards *Pseudokirchneriella subcapitata*. Classification based linear discriminant analysis (LDA) was also performed to distinguish chemicals into toxic and nontoxic groups using the same data set. The developed models were found to possess good predictive quality in terms of internal, external and overall validation parameters. The regression based QSTR model suggests that second order molecular connectivity index (molecular size and lipophilicity), density (aromaticity), relative shape of molecules (cyclicity/aromaticity), and specific molecular fragments of the chemicals are important properties of chemicals to exert their toxicity on *P. subcapitata*. The classification based LDA QSTR model suggested that fused ring aromatic systems, secondary carbon atom fragments, second order valence molecular connectivity indices (molecular size and branching) and molecular weight are the distinguishing features to differentiate chemicals into toxic and nontoxic groups.

1. Introduction

Environmental management has aims to protect different living species from stresses arising from the chemicals released to the ecosystems (Cardinale et al., 2012; Hartung, 2009). Every species plays a momentous role in monitoring evolutionary diversification (Cardinale et al., 2012). The dynamics of all ecosystems are decided by intrinsic and extrinsic functions of individual species and their intimate interaction with non-living objects in the ecosystem such as bioaccumulation and excretion (Ahrens and Traas, 2007; Bell et al., 2005; Brose et al., 2004; Emerson and Kolm, 2005; Gravel et al., 2011). The number of species in an ecosystem and their traits are harmonized predictors of many ecological processes, such as rates of bioconservation, biomass sequestration, productivity, sustainable management of natural resources and biogeochemical cycle (Chapin et al., 2000; Hector and Bagchi, 2007; Kolter and Greenberg, 2006). In this consequence, algae communities provide valuable services to environmental management.

Over the last few decades, the environment has been much exposed to chemical industrializations, mostly through the increased use of agricultural fertilizers and pharmaceuticals, fossil fuel combustion, biomedical waste and petrochemicals (Planson et al., 2012; Rohr et al., 2008). Unrestricted release of chemicals into the environment contributes to the leading causes of pollution worldwide (Scherb and Voigt, 2011). A number of biomedical along with many ecological problems such as global warming, melting of ice caps, loss of biodiversity, abnormality in biogeochemical cycle are arising from the chemical revolution (Cardinale et al., 2012; Gonzalez et al., 2011; Raes et al., 2011; Scherb and Voigt, 2011). Therefore, the environment needs to be nurtured and conserved for ecological functions which may require natural biodiversity. Several studies suggest that conservation of biodiversity may be a valuable pathway in controlling the chemical stress or chemical toxicity of the ecosystem (Bell et al., 2005; Cardinale et al., 2012; Chapin et al., 2000; Hector and Bagchi, 2007). Understanding the chemical toxicity to different species is becoming a point of focus in environmental chemistry (Azarbad et al., 2013; Daouk et al., 2013; Garrigues, 2005). The effects of chemicals on individual species depend on the interaction of chemicals with cellular microenvironment (Hartung et al., 2012). The toxicity screening of a large number of chemicals and understanding their complex cellular interaction towards toxicity require defined *in-vitro*, *in-vivo* experiments which face some socioeconomic and bioethical complications such as time, cost, number of animals for experiment and difficulties in correlation/interpretation with human system (Ahrens and Traas, 2007; Garrigues, 2005). To assist the experimental work, reliable simulation/theoretical

* Corresponding author. Present address: Manchester Institute of Biotechnology, 131 Princess Street, Manchester M1 7DN, United Kingdom. Fax: +91 33 2837 1078.
E-mail addresses: kunalroy_in@yahoo.com, kroy@pharma.jdvu.ac.in (K. Roy).
URL: http://sites.google.com/site/kunalroyindia/ (K. Roy).

0147-6513/$ - see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ecoenv.2013.12.030

analysis should be performed to explain the chemical toxicity on different species. The quantitative structure toxicity/activity/property relationship (QSTR/QSAR/QSPR) methods can be applied to address the structural relationship of chemicals with their toxic potency and categorization of chemicals into toxic and nontoxic groups (Du et al., 2008; Hartung, 2009; Lee et al., 2013; Shi and Yang, 2013; Yuan et al., 2012). Therefore, the present study has been focused to explore the chemical attributes of pesticides, polycyclic aromatic hydrocarbons, nitriles, aldehydes along with other pollutants for their toxic manifestation towards *Pseudokirchneriella subcapitata* (Chen et al., 2009).

2. Materials and methods

2.1. The dataset selection

A number of QSAR models on the chemical toxicity of *P. subcapitata* can be found in the literature (Supplemental data Table S1). The toxicity values (EC_{50}) vary with the variation of experimental conditions. The EC_{50} value is the effective concentration of a chemical/drug that inhibits growth/kill 50% of species. The growth inhibition assay on *P. subcapitata* is usually performed for different time durations (24 h, 48 h, 72 h and 96 h). A review of previous QSAR studies on *P. subcapitata* toxicity (Supplemental data Table S1) shows that QSARs on 48-h toxicity data were previously carried out on small number of chemicals (references of the previous QSAR studies are listed in the Supplemental data at the end of Table S1). This prompted us to develop further QSAR models for 48-h data using a larger chemical set. QSAR models on 72-h data on toxicity susceptibility on *P. subcapitata* have been recently reported using non-polar and polar narcotic chemicals by Aruoja et al. (2014). However, it may be noted that QSAR models developed on data for toxicity of different durations should not be directly compared; according to the OECD guidelines for the QSAR model development (http://www.oecd.org/env/ehs/risk-assessment/37849783.pdf), the endpoint of a QSAR model should be definite one. In this perspective, the 48-h toxicity data of diverse chemicals (Chen et al., 2009) was used in the present study for the development of QSTR models which should be applicable on organic chemicals with structural diversity. The present data set contains 108 chemicals comprising benzenes, alkanes, phenols, anilines, aldehydes, nitriles, alcohols, ketones, pesticides, and polycyclic aromatic hydrocarbons (PAHs). The toxicity data (EC_{50}) was not reported for 2,4,6-trichlorophenol. In a preliminary regression analysis, it was observed that formaldehyde and acetaldehyde act as outliers (showing high residual values) in model development. A number of studies reported that formaldehyde forms polymer (paraformaldehyde) and acetaldehyde forms a number of polymers such as paraldehyde, metaldehyde and polyacetaldehyde (Furukawa and Saeguas, 1962; Ishida, 1981; Li et al., 2011). These two aldehydes were not used in model development. Therefore, three chemicals were excluded from this study and the remaining 105 chemicals have been used in the modeling. The toxicity of the chemicals was reported in terms of EC_{50} (mg/l) data against *P. subcapitata*. The EC_{50} values are the effective concentration of a chemical that inhibits 50% growth of these algae. The EC_{50} (mg/l) values were converted to negative logarithmic scale (pEC_{50} mM) for the model development (Supplemental data Table S2).

2.2. Software

Structures of all chemicals were drawn using Marvin Sketch 5.10.0 software (ChemAxon Ltd. http://www.chemaxon.com). The open source PaDEL-Descriptor software (http://padel.nus.edu.sg/software/padeldescriptor/) was used to calculate extended topochemical atom (ETA) indices while non-ETA descriptors were calculated using Cerius2 version 4.10 software (Cerius 2 Version 4.10. http://accelrys.com/products). The SPSS software (http://www.spss.com) was applied for *k*-means clustering analysis for data set division (training and test sets) and ROC analysis. Stepwise multiple linear regression regression (MLR) and partial least squares (PLS) were performed by MINITAB version 14.13 (http://www.minitab.com). The variable importance plot (VIP) and *Y*-randomization test for PLS regression based QSTR was carried out using SIMCA-P software (UMETRICS SIMCA-P 10.0, www.umetrics.com, Umea, Sweden). STATISTICA version 7.1 software (http://www.statsoft.com/) was used to perform linear discriminant analysis (LDA).

2.3. Descriptor calculation

A set of ETA and non-ETA (2D and 3D) descriptors was used as a pool of independent variables for model development. The non-ETA descriptors are topological (Balaban, Kappa shape indices, flexibility index, subgraph count indices, molecular connectivity indices, Wiener and Zagreb), thermodynamic (AlogP98 and MolRef), structural (MW, Rotlbonds, Hbond acceptor, Hbond donor and Chiral centers), spatial (RadOfGyration, Jurs descriptors, area, density, partial moment of inertia and molar volume), electronic (HOMO, LUMO, superdelocalizability and dipole moment), atom types (Atype), and electrotopological state indices (Supplemental data Table S3). The set of descriptors has been chosen based on their precise application, predictability and easy interpretability in terms of pEC_{50} determination. The conformer generation followed by energy minimization was done prior to 3D descriptor calculation. The multiple conformations of each molecule were generated using the optimal search as the conformational search method. Each conformer was subjected to energy minimization procedure using smart minimizer under open force field (OFF) to generate the lowest energy conformation for each structure. The Gasteiger method was used for charge calculation of molecules (Gasteiger and Marsili, 1980). Due to the importance of lipohilicity in aquatic toxicity modeling, experimental partition co-efficient ($\log K_{ow}$) (taken from the literature; Chen et al., 2009) was also tried as an additional descriptor.

2.4. Dataset splitting and model development

The data set (N_{total}=105) was divided into training and test sets based on the *k*-means clustering technique. Approximately 30% of chemicals were selected as the test set members (N_{test}=31) and the remaining 70% as the training set members ($N_{training}$=74) (Dougherty et al., 2002; Everitt et al., 2001; Johnson and Wichern, 2005). The splitting was done in such a way that each of the sets covers the total chemical space of the entire data set (Martin et al., 2012). The same division was used for both regression based QSTR and LDA studies. The models were generated using the structural information of chemicals from the training set, and the test set chemicals were employed to check model reliability or external predictive quality of models. The regression based QSTR models were developed using the partial least squares (PLS) method. The descriptors appeared in stepwise MLR with stepping criteria *F*-to-enter 4 and *F*-to-remove 3.9 (Darlington, 1990) were subjected to partial least squares (PLS) regression (Eriksson et al., 2001, 2002; Wold, 1995). PLS is a more robust regression method than MLR and it obviates the problem of intercorrelation in the latter approach. The linear discriminant analysis (LDA) has also been applied to identify the discriminatory features into higher and lower toxic chemicals (Fisher, 1936; Mitteroecker and Bookstein, 2011). The threshold value (pEC_{50}=0.936 mM) for LDA analysis was selected based on arithmetic mean of pEC_{50} values (Kar et al., 2012). Chemicals having the pEC_{50} value higher than or equal to 0.936 mM were considered as highly toxic to *P. subcapitata*. In the LDA studies, 60 descriptors out of 171 descriptors were selected as the pool of predictor variables based on molecular spectrum analysis.

2.5. Validation metrics for the regression based QSTR model

The robustness of the QSTR models was verified by using a number of statistical parameters. Three strategies were followed: (1) leave-one-out (LOO) internal validation or cross-validation for the training set compounds, (2) external validation using the test compounds and (3) overall validation using both training and test set compounds. The main objective of this QSTR modeling is the development of robust models which are able to make accurate and reliable predictions of toxicity of chemicals towards *P. subcapitata*. Therefore, mathematical equations developed from the training set were subsequently validated internally using training set chemicals as well as externally using the test set molecules for checking the predictive quality of the developed models. The overall validation strategies countercheck the reliability of the developed models for their possible application on a new set of data and assess confidence of such predictions.

The model fitness parameters $R^2$and R^2_a, internal validation metrics Q^2, $\overline{r^2_{m(LOO)scaled}}$ and $\Delta r^2_{m(LOO)scaled}$, external validation metrics R^2_{pred}, $\overline{r^2_{m(test)scaled}}$, $\Delta r^2_{m(test)scaled}$ and overall metrics $\overline{r^2_{m(overall)scaled}}$, $\Delta r^2_{m(overall)scaled}$ were reported in connection with validation for the developed models (Kubinyi et al., 1998; Roy et al., 2012, 2013) (the definitions of all fitness parameters have been provided in Supplemental data). Further, predictive qualities of the models were assessed based on Golbraikh and Tropsha′s approaches (Golbraikh and Tropsha, 2002). According to the acceptance criteria set forth by Golbraikh and Tropsha, a model must follow the following conditions:

(i) $Q^2 > 0.5$

(ii) $r^2 > 0.6$

(iii) $(r^2 - r_o^2)/r^2 < 0.1$ *or* $(r^2 - r_o'^2)/r^2 < 0.1$

(iv) $0.85 \le k \le 1.15$ *or* $0.85 \le k' \le 1.15$

Here, r^2 and ${r_o}^2$ are squared correlation coefficient values between the observed and predicted values (*Y* and *X* axes respectively) of the compounds with and without intercept respectively. An interchange of the axes gives the value ${r_o}'^2$ instead of ${r_o}^2$. The plot of observed values (*Y*-axis) against the predicted values (*X*-axis) of the test set compounds setting the intercept to zero gives the slope of the fitted line as the value of *k*. The interchange of axes gives the value of *k*′.

The developed models were also subjected to a randomization test (100 permutations) to check the possibility of chance correlationn. The pEC_{50} (*Y*) values