Computers in Biology and Medicine ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Three dimensional quantitative structure–toxicity relationship modeling and prediction of acute toxicity for organic contaminants to algae

Xiangqin Jin [a], Minghao Jin [b], Lianxi Sheng [a,*]

[a] State Environmental Protection Key Laboratory of Wetland Ecology and Vegetation Restoration, School of Environment, Northeast Normal University, Changchun 130117, PR China

[b] Department of Mathematics, Heilongjiang Institute of Technology, Harbin 150050, PR China

ARTICLE INFO

Article history:
Received 7 February 2014
Accepted 23 May 2014

Keywords:
Organic contaminant
Acute toxicity
Three dimensional quantitative structure–toxicity relationship
Machine learning
Alga

ABSTRACT

Although numerous chemicals have been identified to have significant toxicological effect on aquatic organisms, there is still lack of a reliable, high-throughput approach to evaluate, screen and monitor the presence of organic contaminants in aquatic system. In the current study, we proposed a synthetic pipeline to automatically model and predict the acute toxicity of chemicals to algae. In the procedure, a new alignment-free three dimensional (3D) structure characterization method was described and, with this method, several 3D-quantitative structure–toxicity relationship (3D-QSTR) models were developed, from which two were found to exhibit strong internal fitting ability and high external predictive power. The best model was established by Gaussian process (GP), which was further employed to perform extrapolation on a random compound library consisting of 1014 virtually generated substituted benzenes. It was found that (i) substitution number can only exert slight influence on chemical's toxicity, but low-substituted benzenes seem to have higher toxicity than those of high-substituted entities, and (ii) benzenes substituted by nitro group and halogens exhibit high acute toxicity as compared to other substituents such as methyl and carboxyl groups. Subsequently, several promising candidates suggested by computational prediction were assayed by using a standard algal growth inhibition test. Consequently, four substituted benzenes, namely 2,3-dinitrophenol, 2-chloro-4-nitroaniline, 1,2,3-trinitrobenzene and 3-bromophenol, were determined to have high acute toxicity to *Scenedesmus obliquus*, with their EC_{50} values of 2.5 ± 0.8, 10.5 ± 2.1, 1.4 ± 0.2 and 42.7 ± 5.4 μmol/L, respectively.

1. Introduction

Algae are a polyphyletic and paraphyletic group of organisms that occur in most habitats and usually found in damp places or bodies of aquatic environments [1]. Diverse compounds have been reported to exhibit acute toxicity on algae [2]. However, it is too time-consuming and expensive to systematically conduct experimental assay on all existing and potential chemicals for their toxicity. Therefore, an important aspect of modern toxicology research is the prediction of toxicity of compounds from their molecular structure using quantitative structure–toxicity relationship (QSTR) approach [3,4]. Recently, a number of QSTR studies have been successfully performed to investigate the structural basis and molecular properties underlying compound toxicity to algae [5–9]. For example, Lu et al. developed a series of simple linear correlations between the molecular structure and acute toxicity of aromatic compounds using quantum chemical parameters and physicochemical features [10–12]; these correlation models were later extended to the joint toxicity of substituted phenols and anilines to algae [13–15]. Despite these successful cases, above works can only be regarded as preliminary QSTR studies on acute toxicity to algae owing to their small data size and simple regression method. In this respect, Lessigiarska et al. organized a large compound set to investigate fish, algae and Daphnia toxicity [16]. Recently, Gramatica and co-workers performed a case study of toxicity of (benzo)triazoles to the algae *Pseudokirchneriella subcapitata* and suggested that a satisfactory QSTR model can only be built from elaborate selection of structural descriptor, regression method and data set [17].

In the current study, we first collected hundreds of organic contaminants and environmental pollutants and their acute toxicity to a variety of algae to define a distinct data set. Second, a new

* Corresponding author. Tel.: +86 431 85099797; fax: +86 431 85099797.
E-mail addresses: lianxi_sheng@126.com, shenglx@nenu.edu.cn (L. Sheng).

http://dx.doi.org/10.1016/j.compbiomed.2014.05.009
0010-4825/

three-dimensional QSTR (3D-QSTR) methodology was developed to characterize the structural and nonbonded profile of organic molecules. Third, seven machine learning methods, including three linear and four nonlinear, were employed to correlate the generated descriptors with compound toxicity based on the collected data set. Forth, the acute toxicities of several promising candidates were evaluated *in vitro* using a standard algae inhibition test. In this way, we expect to obtain an optimized, general-purpose 3D-QSTR model that is able to carry out high-throughput screening of vast chemical space to determine potential toxicity for aquatic organisms.

2. Materials and methods

2.1. Data set

A total of 873 organic contaminants with known acute toxicities to algae were compiled from a number of literatures [10–29] and publicly available sources. These compounds include diverse industrial chemicals and environmental pollutants such as nitroaromatics, phenols, anilines, triazoles and substituted benzenes, and their toxicity was experimentally determined as 50% effective inhibition concentration (EC_{50}) for a variety of algae species such as *Scenedesmus obliquus*, *Chlorella pyrenoidosa* and *P. subcapitata* in 48 or 72 h. Here, the EC_{50} values were converted to negative logarithm form (pEC_{50}) for subsequent analysis. The 873 chemicals as well as their acute toxicity data are tabulated in Supporting Information Table S1.

In a high-profile article Golbraikh and Tropsha pointed out that internal validation appears to be the necessary but not the sufficient condition for a statistical regression model to have high predictive power, and they emphasized that the external validation is the only way to establish reliable model [30]. Therefore, we randomly split the data set into two parts with the ratio of ~3:1; the large part consisted of 655 compounds was used as internal training set to develop QSTR models, and the small part contained 218 samples and was served as external test set to blindly validate the developed models.

2.2. Molecular structure characterization

Previously, quantum chemical descriptors (*e.g.* the energies of the highest and lowest occupied molecular orbital) and physicochemical descriptors (*e.g.* dipole moment and hydrophobicity) have been widely used to perform QSTR modeling of compound toxicity to algae [10,11]. The drawback of these 2D descriptors is that they are incapable of capturing molecular 3D structure information. Currently, the most widely applied 3D-QSAR methodology is the comparative molecular field analysis (CoMFA) [31]. However, CoMFA can only be applied to the cases where all investigated molecules share a common sketch to facilitate alignment manipulation. To solve this issue, we herein used a new 3D molecular structure characterization method that is free of the alignment manipulation step.

Organic molecules are commonly composed of atoms H, C, N, P, O, S, F, Cl, Br and I; they can be categorized into five classes in terms of their chemical attribute, that is, class I: H, class II: C, class III: N and P, class IV: O and S, and class V: F, Cl, Br and I. All the 15 possible combinations among the five classes of atoms are represented in Table 1. For example, the combination of classes III with IV includes four atom pairs N–O, N–S, P–O and P–S. Subsequently, three nonbonded interaction potentials that dominate biomolecular recognition, *i.e.* electrostatic, steric and hydrophobic, were calculated for all atom pairs in an organic molecule, and the resulting values were assigned into the 15 combinations.

Table 1
The classification of atoms in organic molecules and all the 15 possible combinations between different classes of atom.

Atomic class	Atoms	I	II	III	IV	V
I	H	1–1	1–2	1–3	1–4	1–5
II	C		2–2	2–3	2–4	2–5
III	N, P			3–3	3–4	3–5
IV	O, S				4–4	4–5
V	F, Cl, Br, I					5–5

In this way, the nonbonded interaction profile of an organic molecule can be characterized using at most 45 nonbonded descriptors; the first, secondary and third 15 descriptors represent, respectively, electrostatic, steric and hydrophobic potentials for the 15 combinations of five atomic classes in an organic molecule.

The electrostatic, steric and hydrophobic interaction potentials between two atoms i and j in a molecule can be calculated using, respectively, Coulomb's law $E_{ij}^{E}=(q_iq_j/d_{ij})$, Lennard–Jones equation $E_{ij}^{S}=\varepsilon_{ij}[(D_{ij}/d_{ij})^{12}-2(D_{ij}/d_{ij})^{6}]$ and empirical hydrophobic potential $E_{ij}^{H}=-(S_i\rho_i+S_j\rho_j)e^{-d_{ij}}$ [32], where the q_i is partial charge of atom i, d_{ij} is spatial distance between atoms i and j, $\varepsilon_{ij}=(\varepsilon_{ii}\varepsilon_{jj})^{1/2}$ is the potential well of atom pair i–j [33], $D_{ij}=(C_hD_{ii}+C_hD_{jj})/2$ indicates the van der Waals contact distance of atom pair i–j [34], and S_i and ρ_i are the solvent accessible surface area and hydrophobicity of atom i [35,36].

Here, chlorobenzene was used as an example to illustrate the process of generating molecular descriptors. As shown in Fig. 1, the 2D molecular structure of chlorobenzene was automatically converted to 3D conformation using CORINA program [37], and the coarse-grained 3D conformation was then minimized with the MM2 force field [38] to achieve its energetic minimum. Subsequently, the electrostatic, steric and hydrophobic potentials were calculated for the chlorobenzene based on its minimized 3D structure. The chlorobenzene contains five hydrogen atoms, six carbon atoms and a chlorine atom, corresponding to classes I, II and V, respectively. The combination of classes I, II and V presents six possibilities, that is, I–I, I–II, I–V, II–II, II–V and V–V. Here, the chlorobenzene is unable to define the V–V combination since it has only one atom (chlorine atom) belonging to the class V. Therefore, five descriptors were finally generated for each nonbonded term, totally resulting in 15 non-zero nonbonded descriptors for the chlorobenzene molecule.

2.3. Machine learning methods

Seven sophisticated machine learning methods, including three linear and four nonlinear methods, namely, multivariable linear regression (MLR) [39], partial least squares (PLS) regression [40], back-propagation artificial neural network (BPANN) [41], support vector machine (SVM) [42], least square-support vector machine (LSSVM) [43], random forest (RF) [44] and Gaussian process (GP) [45,46], were employed here to perform statistical modeling. Here, a coarse-grained grid-searching scheme using root-mean-square error of cross-validation as the objective function was carried out to determine the optimum combination of model parameters.

Matlab toolboxes of ChemoAC_MLR, ChemoAC_PLS, NNET, SVMKM, LSSVMlab, RFTX and GPML were used to implement MLR, PLS, BPANN, SVM, LSSVM, RF and GP modeling, respectively.

2.4. Statistics

The statistical quality of built regression models is measured using the coefficients of determination of fitting (r_f^2) and 10-fold cross-validation (r_c^2) on training set, the coefficients of determination

Please cite this article as: X. Jin, et al., Three dimensional quantitative structure–toxicity relationship modeling and prediction of acute toxicity for organic contaminants to algae, Comput. Biol. Med. (2014), http://dx.doi.org/10.1016/j.compbiomed.2014.05.009