Chemosphere 120 (2015) 680–689

Contents lists available at ScienceDirect

Chemosphere

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

QSTR modeling for predicting aquatic toxicity of pharmacological active compounds in multiple test species for regulatory purpose

Kunwar P. Singh [a,b,*,1], Shikha Gupta [a,b,1], Nikita Basant [c,1]

[a] Academy of Scientific and Innovative Research, Anusandhan Bhawan, Rafi Marg, New Delhi 110 001, India
[b] Environmental Chemistry Division, CSIR-Indian Institute of Toxicology Research, Post Box 80, Mahatma Gandhi Marg, Lucknow 226 001, India
[c] KanbanSystems Pvt. Ltd., Laxmi Nagar, Delhi 110 092, India

HIGHLIGHTS

- QSTRs constructed for toxicity prediction of diverse PACs in multiple test species.
- Proposed QSTRs used a few simple non-quantum mechanical molecular descriptors.
- QSTRs validated using OECD recommended procedures using multi-species toxicity data.
- QSTRs passed validation criteria; precisely predicted PACs toxicity in multi-species.

ARTICLE INFO

Article history:
Received 4 March 2014
Received in revised form 25 September 2014
Accepted 12 October 2014

Handling Editor: A. Gies

Keywords:
Pharmaceutically active compounds
QSTRs
Decision treeboost
Decision tree forest
Multi-species
Regulatory

ABSTRACT

High concentrations of pharmacological active compounds (PACs) detected in global drinking water resources and their toxicological implications in aquatic life has become a matter of concern compelling for the development of reliable QSTRs (qualitative/quantitative structure–toxicity relationships) for their risk assessment. Robust QSTRs, such as decision treeboost (DTB) and decision tree forest (DTF) models implementing stochastic gradient boosting and bagging algorithms were established by experimental toxicity data of structurally diverse PACs in daphnia using molecular descriptors for predicting toxicity of new untested compounds in multiple test species. Developed models were rigorously validated using OECD recommended internal and external validation procedures and predictive power tested with external data of different trophic level test species (algae and fish). Classification QSTRs (DTB, DTF) rendered accuracy of 98.73% and 97.47%, respectively in daphnia and 84.38%, 85.94% (algae), 78.46% and 79.23% (fish). On the other hand, the regression QSTRs (DTB, DTF) yielded squared correlation coefficient values of 0.831, 0.852 (daphnia), 0.534, 0.556 (algae) and 0.620, 0.637 (fish). QSTRs developed in this study passed the OECD validation criteria and performed better than reported earlier for predicting toxicity of PACs, and can be used for screening the new untested compounds for regulatory purpose.

1. Introduction

Pharmacological active compounds (PACs) are an important class of chemicals as they constitute active ingredient of already established drugs or may be molecules of interest for future drug development process. PACs are often designed molecules with a specific effect and receptor in mind, and thus have potentially specific toxicological targets. In recent years, the number of PACs detected in the environment has increased spectacularly, reaching a broad number of consumed drugs and including virtually all the existing therapeutic classes (Guillén et al., 2012). Several studies have already reported the presence and considerably higher levels of PACs in various aquatic systems worldwide (Luo et al., 2011; Ascar et al., 2013). Aquatic organisms are particularly important targets, as they are exposed via wastewater residues over their whole life (Fent et al., 2006). The presence of increased levels of PACs in the environment raises concerns about their propensity to select for resistant bacteria and facilitate the establishment and amplification of pathogen reservoirs that threaten public health (Luo et al., 2011). PACs concentrations once build up in water sources, may allow them to pass over to various processed food items leading to exposure of consumers, especially infants,

* Corresponding author at: Environmental Chemistry Division, CSIR-Indian Institute of Toxicology Research, Post Box 80, Mahatma Gandhi Marg, Lucknow 226 001, India. Tel.: +91 522 2476091; fax: +91 522 2628227.
E-mail addresses: kpsingh_52@yahoo.com, kunwarpsingh@gmail.com (K.P. Singh).
[1] All the authors contributed equally.

http://dx.doi.org/10.1016/j.chemosphere.2014.10.025
0045-6535/

causing immunity to these molecules, rendering them ineffective in later years. Thus, the PACs need to be screened for their aquatic toxicities. Ecotoxicological assessment of the pharmaceuticals has been based on acute toxicity experiments performed by standard tests according to existing guidelines (OECD) using laboratory organisms belonging to different trophic levels, such as algae, zooplankton, other invertebrates and fish (Christen et al., 2010). These three organism groups (crustaceans, algae, fish) represent different trophic levels of the aquatic food-web. The choice of these three trophic levels is considered to be relevant in order to protect aquatic ecosystems (Wei et al., 2006).

Daphnia magna is widely used as a standard test organism in aquatic toxicology. It is an important primary consumer of primitive plant life and itself a major food source for vertebrate and invertebrate predators and has been used as a representative for other freshwater animals in standard tests of toxicity (OECD, 1984) because of its high sensitivity, easy handling and high reproductive rate (Katritzky et al., 2009). Daphnia acute toxicity (48 h) test is used for short term toxicity (LC_{50}) assessment of chemicals. Algae, the essential components of aquatic ecosystems produce oxygen and organic substances on which most other life forms depend by providing food for other organisms, including fish and invertebrates. The algal toxicity test involves estimating the lethal concentration (LC_{50}) based on 24 h exposure to a test chemical. The 96 h LC_{50} test was used to measure the susceptibility and survival potential of fish to toxic substances, which is the last chain in the aquatic food cycle (Castano et al., 1996). However, the experimental approach for the *in vivo* toxicity testing of the chemicals is very expensive, labor intensive and time consuming (Tan et al., 2010). Therefore, there is an urgent need for reliable quantitative-structure toxicity relationships (QSTRs) based models capable of predicting the toxicities of the PACs using their structural properties.

PACs toxicity prediction models based on interspecies correlation estimation (ICE) (Kar and Roy, 2010; Das et al., 2013), ECOSAR (Sanderson and Thomsen, 2007) and multiple linear regression (MLR) (Jiang et al., 2010; Tugcu et al., 2012) approaches have been proposed earlier. ICE method, although, yielded results in the acceptable range of accuracy, it correlates endpoint toxicities of chemicals in different test species along with a few selected properties of the compounds. Thus, while predicting toxicity of new untested PACs, the ICE method requires experimental toxicity values of these molecules in some other test species as well, hence limiting its applicability. ECOSAR model using single property of chemical ($\log K_{ow}$) yielded toxicities of PACs with very low correlations with the measured values (Sanderson and Thomsen, 2007). MLR using a few molecular descriptors successfully predicted toxicities of PACs in the acceptable range, however, the set of compounds considered were small, limiting its applicability over larger data set. Moreover, all these methods are based on linear relationships and cannot capture the nonlinearities in data, which is a common feature of the experimental toxicity data (Singh et al., 2013a). Moreover, these studies, in general consider single test species and no attempts have yet been made to develop a multi-species QSTRs (model developed using toxicity data in one species and capable of predicting toxicity in other test species). Therefore, it would be desirable to develop global toxicity prediction QSTR models which are extendable to predict toxicity of diverse PACs in multiple test species.

In recent years ensemble learning (EL) methods have emerged as unbiased tools for modeling the complex relationships between a set of independent and dependent variables and have successfully been applied for establishing relationships between molecular descriptors of diverse chemicals and their endpoint toxicities (Singh and Gupta, 2014a,b). These methods overcome the problems with the weak predictors, alleviate the small sample size problem and reduce the over-fitting the training data (Dietterich, 2000). Decision trees (DTs) commonly used as base predictors in EL modeling, supplemented with bagging and stochastic gradient boosting techniques improve the prediction accuracy of the weak learners (Breiman, 1996). Decision tree forest (DTF) and decision treeboost (DTB) implementing the bagging and boosting techniques, respectively are relatively new methods for improving the accuracy of a predictive function (Dietterich, 2000). These techniques are inherently non-parametric statistical methods and make no assumption regarding the underlying distribution of the values of predictor variables (Singh et al., 2014). Here, we have considered the EL approaches to develop QSTRs (classification and regression) for predicting the toxicity of PACs.

In this study, we developed multi-species QSTRs for predicting the acute toxicity of diverse PACs using simple molecular descriptors. Here, classification and regression QSTRs (DTB and DTF) were constructed using daphnia toxicity data (LC_{50}) to predict the toxicity classes and endpoint values of PACs. The predictive and generalization abilities of the constructed QSTRs were evaluated using various statistical criteria parameters and performance of these models was validated with the toxicity data of other test species (algae and fish).

2. Materials and methods

2.1. Data sets

Multi-species toxicity data for diverse PACs (Sanderson and Thomsen, 2009) were considered for analysis in this study. Thorough examination of the data set showed that it consisted of a wide variety of classes of PACs. The acute toxicity data in the compiled database were obtained using OECD toxicity testing guidelines and were reported in mg L^{-1}, which were transformed into their corresponding molar concentration and finally were converted to their negative logarithmic scale values (pLC_{50}, M) for regression QSTR model development. However, for the classification model development, toxicity values expressed in mg L^{-1} were used. For developing the predictive QSTR models, the daphnia toxicity (LC_{50}) data were used. The chemically heterogeneous dataset is comprised of 244 PACs representing different groups reporting their acute toxicity in daphnia, algae, and fish. A total of 50 salts and charged molecules were removed from the dataset. In this study, a total of 158 compounds for daphnia, 64 for algae, and 130 for fish toxicity were selected. The toxicities (pLC_{50}, M) of the selected PACs in three datasets varied between 0.59 and 7.04 (daphnia), 1.95 and 6.38 (algae), and 1.29 and 6.67 (fish).

Since, the structural diversity of the considered PACs is very important for global model development (Zhao et al., 2006), it was measured using the Tanimoto similarity index (TSI), which is an appropriate distance metric for topology-based chemical similarity studies. TSI is calculated using Tanimoto similarity between the fingerprint of a chemical and a consensus 1024 bit fingerprint. The fingerprint generation is based on the implementation of the open source chemoinformatics library (Steinbeck et al., 2003). For a given molecule all possible paths for a predefined length are generated, the path is submitted to a hash function which uses it as a seed to a pseudorandom generator, the hash function outputs a set of bits, which is added to the fingerprint. For a molecule, TSI is calculated as; $TSI_{AB} = 2Z_{AB}[Z_{AA} + Z_{BB} - Z_{AB}]^{-1}$, where Z is the similarity matrix, A and B are the two molecules being compared. A good cutoff for biologically similar molecules is 0.7 or 0.8 (Singh et al., 2013a). The distribution of the TSI values of PACs considered in three test species is given in Fig. SI1 (Supplementary materials). TSI values for the PACs in daphnia, algae, and fish data are 0.132, 0.204 and 0.144, respectively. These values suggest that the compounds considered in this work represent sufficiently high