Toxicology and Applied Pharmacology 275 (2014) 198–212

Contents lists available at ScienceDirect

Toxicology and Applied Pharmacology

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

In silico prediction of toxicity of non-congeneric industrial chemicals using ensemble learning based modeling approaches

Kunwar P. Singh *, Shikha Gupta

Academy of Scientific and Innovative Research, Anusandhan Bhawan, Rafi Marg, New Delhi 110 001, India
Environmental Chemistry Division, CSIR-Indian Institute of Toxicology Research, Post Box 80, Mahatma Gandhi Marg, Lucknow 226 001, India

ARTICLE INFO

Article history:
Received 5 August 2013
Revised 4 January 2014
Accepted 13 January 2014
Available online 23 January 2014

Keywords:
Toxicity
Diverse chemicals
Ensemble learning models
Interspecies model
Molecular descriptors
Regulatory toxicology

ABSTRACT

Ensemble learning approach based decision treeboost (DTB) and decision tree forest (DTF) models are introduced in order to establish quantitative structure–toxicity relationship (QSTR) for the prediction of toxicity of 1450 diverse chemicals. Eight non-quantum mechanical molecular descriptors were derived. Structural diversity of the chemicals was evaluated using Tanimoto similarity index. Stochastic gradient boosting and bagging algorithms supplemented DTB and DTF models were constructed for classification and function optimization problems using the toxicity end-point in *T. pyriformis*. Special attention was drawn to prediction ability and robustness of the models, investigated both in external and 10-fold cross validation processes. In complete data, optimal DTB and DTF models rendered accuracies of 98.90%, 98.83% in two-category and 98.14%, 98.14% in four-category toxicity classifications. Both the models further yielded classification accuracies of 100% in external toxicity data of *T. pyriformis*. The constructed regression models (DTB and DTF) using five descriptors yielded correlation coefficients (R^2) of 0.945, 0.944 between the measured and predicted toxicities with mean squared errors (MSEs) of 0.059, and 0.064 in complete *T. pyriformis* data. The *T. pyriformis* regression models (DTB and DTF) applied to the external toxicity data sets yielded R^2 and MSE values of 0.637, 0.655; 0.534, 0.507 (marine bacteria) and 0.741, 0.691; 0.155, 0.173 (algae). The results suggest for wide applicability of the inter-species models in predicting toxicity of new chemicals for regulatory purposes. These approaches provide useful strategy and robust tools in the screening of ecotoxicological risk or environmental hazard potential of chemicals.

Introduction

Chemicals are today an essential part of human life and comfort and consequently, a large number of industrial chemicals are manufactured and used every day. However, many of these have been identified as potentially toxic to the humans. The regulatory agencies have been emphasizing their safety assessment prior to their manufacture and use (Casalegno et al., 2005). A 40-h toxicity test, expressed in terms of the median growth inhibition concentration (IGC_{50}) value to the fresh water ciliate *Tetrahymena pyriformis* is considered appropriate for toxicological testing and safety evaluation of the chemicals (Cronin et al., 2002). *T. pyriformis* is an ubiquitous ciliated protozoan belonging to free living, fresh water genus and are sensitive to growth conditions. It has several advantageous characteristics for toxicity studies, such as they play key role in the transfer of energy and matter within the microbial loop. They occupy one of the first trophic levels in aquatic ecosystems and thus, are early warning of a toxic danger (Lukacinova et al., 2007). In *T. pyriformis*, the toxicological end-points have usually been restricted to a measure of growth inhibition rather than of cell viability (Schultz, 1997). The main relevance of this end-point is to characterize aquatic toxicity. According to Duchowicz and Ocsachoque (2009), the IGC_{50} encapsulates the most aqueous toxicity information. Toxicity evaluation of the high production volume compounds, such as industrial chemicals and pesticides is a top global regulatory priority. Accordingly, several experimental databases on IGC_{50} of different groups of chemicals have been developed by various research groups (Aptula et al., 2005; Cronin and Schultz, 2001; Cronin et al., 2002, 2004; Devillers, 2004; Deweese and Schultz, 2001; Melagraki et al., 2005; Netzeva and Schultz, 2005; Netzeva et al., 2003; Ren, 2003; Ren and Schultz, 2002; Roy et al., 2005; Schultz, 1999; Schultz et al., 2005; Schuurmann et al., 2003; Serra et al., 2001). Because the experimental determination of toxicological properties is a costly and time consuming process, it is essential to develop predictive mathematical relationships to theoretically quantify toxicity (Melagraki et al., 2006). Quantitative structure–activity relationships (QSARs) are increasingly being used as a tool to assess regulatory agencies in toxicological assessment of chemical substances (Cronin et al., 2002). Desirable qualities of QSAR include the model being transparent, easily portable, and developed with interpretable descriptors. Several quantitative structure–toxicity relationship (QSTR) models, based on multiple linear regression (MLR), partial least squares regression (PLSR), k-nearest neighbor (k-NN), artificial neural networks (ANN), support vector machines (SVM), decision tree (DT) approaches have been proposed by various research groups

* Corresponding author. Fax: +91 522 2628227.
E-mail addresses: kpsingh_52@yahoo.com, kunwarpsingh@gmail.com (K.P. Singh).

0041-008X/$ – see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.taap.2014.01.006

(Cheng et al., 2011; Ivanciuc, 2004; Jalali-Heravi and Kyani, 2008; Melagraki et al., 2006; Schuurmann et al., 2003; Toropov et al., 2010; Zhu et al., 2008) for predicting *T. pyriformis* toxicity of chemicals. However, the main problem in any QSAR analysis is the evaluation and control of the predictive ability of the developed model (Toropov et al., 2010). Moreover, these local models could provide acceptable predictive accuracy for a very limited chemical domain; they were not applicable to assess a large diverse set of chemical structures (Cheng et al., 2011). Also, all these approaches considered different types of molecular descriptors as estimators, and selection and computation of relevant descriptors to extract information from compound structures is the major limitation of this research field.

In recent years ensemble learning (EL) methods (Snelder et al., 2009) have emerged as unbiased tools for modeling the complex relationships between set of independent and dependent variables and have been applied successfully in various research areas (Yang et al., 2010). In general, these methods are designed to overcome problems with weak predictors (Hancock et al., 2005) and have the advantage of alleviating the small sample size problem by averaging and incorporating over multiple classification models to reduce the potential for over-fitting the training data (Dietterich, 2000). Decision tree forest (DTF) and decision treeboost (DTB) implementing bagging and boosting techniques, respectively are relatively new methods for improving the accuracy of a predictive function (Yang et al., 2010). These techniques are inherently non-parametric statistical methods and make no assumption regarding the underlying distribution of the values of predictor variables and can handle numerical data that are highly skewed or multi-model in nature (Mahjoobi and Etemad-Shahidi, 2008). To our knowledge, ensemble learning methods have not yet been applied to the toxicity prediction modeling.

Selection of appropriate molecular descriptors in toxicity prediction is yet another important issue. A large number and variety of such descriptors have been used in several earlier studies, generally derived through highly complicated semi-empirical and empirical methods based on quantum mechanical calculations (Eroglu et al., 2007; Wang et al., 2010; In et al., 2012). Hence, it would be desirable to develop toxicologically relevant QSTRs using simple properties that can be derived directly from a chemical's structure. Moreover, in view of the regulatory toxicology requirements, models discriminating compounds merely between toxic and non-toxic classes are not enough and it is very much desirable to have more efficient screening tools capable of classifying compounds in several toxicity classes, such as highly toxic, toxic, harmful, and non-harmful; as well as capable of predicting the toxicity end-points in a quantitative manner.

In this study, the basic objectives were to construct the ensemble learning based models (DTB and DTF) for predicting the toxicity of the diverse chemical compounds using simple molecular descriptors, Accordingly, classification and regression models were constructed to predict the toxicity classes and the toxicity end-point ($-\log IGC_{50}$) of the diverse chemicals using a set of selected molecular properties/descriptors as estimators. The predictive and generalization abilities of the DTB and DTF classification and regression models constructed here were evaluated using several statistical criteria parameters and performance of these models were tested using external datasets. Moreover, the predictive ability of the DTB and DTF regression models was compared with kernel partial least squares regression (KPLSR), a basic modeling approach.

Materials and methods

Data set

For developing the ensemble learning based QSTR models for toxicity prediction of chemicals in *T. pyriformis*, data from multiple sources were considered (Cheng et al., 2011; Tetko et al., 2008; Xue et al., 2006). The chemically heterogeneous dataset is comprised of 1450 chemicals representing different groups. The reported 50% growth-inhibitory concentration values (IGC_{50}) are based on the standard *T. pyriformis* test protocol (Schultz and Netzeva, 2004). For these compounds, values of 40-h exposure based median growth inhibition concentration (IGC_{50}) for *T. pyriformis* are reported as $-\log IGC_{50}$ ($pIGC_{50}$, mmol L^{-1}), where the logarithm is taken to contract the dataset to a computationally efficient range (Serra et al., 2001). Toxicity level generally increases with increasing value of $-\log IGC_{50}$ (mmol L^{-1}), and compounds with positive values are generally considered to be toxic or weakly toxic. Complete dataset of 1450 compounds with end-point values are presented in Table S1 (Supporting Information). Since, validation aims to stimulate the predictivity of a model towards new, unknown chemicals, external datasets were collected from literature for model validation (Zhang et al., 2010). Accordingly, the external datasets (Table S2 of Supporting Information) contained comparative toxicity (IGC_{50}) data of chemicals to *T. pyriformis*, and median effective concentration (EC_{50}) values of chemicals in *Vibrio fischeri* (marine bacterium) and *Scenedesmus obliguue* (algae). The EC_{50} values in view of their high correlation reported with IGC_{50} values in different species (Zhang et al., 2010), were also considered for external validation of the constructed models.

Molecular descriptors and feature selection

In toxicological studies, the molecular descriptors represent structural and physicochemical properties of compounds. The descriptors were calculated for each molecule using Toxmatch (Ideaconsult Ltd.). Molecular descriptors (physical, constitutional, geometrical, and topological) were computed by 2D structures of the molecules, which were taken in the form of SMILES (simplified molecular input line entry system). A set of 60 different molecular properties of each of the compound were selected initially. Since, all the molecular properties may not be relevant to the modeling; elimination of less significant descriptors can improve the accuracy of prediction, and facilitate the interpretation of the model through focusing on the most relevant variables. Initial features were selected by model-fitting approach. EL modeling was performed. For optimal values of the model parameters, the EL models were trained by using the complete set of features computing the respective scoring functions to rank the contribution of features in the current set. The lowest ranked features were then removed (Xue et al., 2006). The EL systems were retained by using the remaining set of features, and the corresponding prediction accuracies (misclassification rate, and mean squared error of prediction) were computed by means of 10-fold cross validation. The selected descriptors are logarithmic form of octanol–water partition coefficient (log P), molecular weight (MW), molecular surface area (MSA), charge partial surface area 22 (CPSA-22), connectivity index order one (CIOO), eccentric connectivity index (ECI), number of atoms in largest chain (NALC), and number of atoms in largest pi-system (NALPS). Finally a set of six descriptors for classification and five descriptors for regression modeling were considered in this study. The basic statistics of the selected descriptors for different datasets are given in Table 1 and Table S3 (Supporting information).

Data processing

Since the aim of present study is to build a robust model capable of making accurate and reliable predictions of toxicity of new compound, the model derived from a training set should be validated/tested using new chemical moieties for checking its predictive ability. The validation strategies check the reliabilities of models for their possible application on a new dataset, and confidence in the prediction can thus be judged. For predictive modeling the end point (IGC_{50}) values were expressed as $-\log IGC_{50}$ ($pIGC_{50}$, mmol L^{-1}). Categorization of compounds as *T. pyriformis* toxic (TPT) and non-toxic to *T. pyriformis* (Non-TPT) was based on the criteria of Xue et al. (2006). According to the criteria