Journal of Molecular Liquids 209 (2015) 404–412

Contents lists available at ScienceDirect

Journal of Molecular Liquids

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

Predicting acetyl cholinesterase enzyme inhibition potential of ionic liquids using machine learning approaches: An aid to green chemicals designing

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

[a] KanbanSystems Pvt. Ltd., Laxmi Nagar, Delhi 110092, India
[b] CSIR-Indian Institute of Toxicology Research, Environmental Chemistry Division CSIR-Indian Institute of Toxicology Research, Academy of Scientific & Innovative Research (AcSIR), Lucknow-226001, India

ARTICLE INFO

Article history:
Received 24 February 2015
Received in revised form 25 May 2015
Accepted 1 June 2015
Available online 11 June 2015

Keywords:
AChE inhibition
Ionic liquids
SARs
Structural diversity
Molecular descriptors

ABSTRACT

The ionic liquids (ILs) constitute a group of novel chemicals that have potential industrial applications. Designing of safer ILs is among the priorities of the chemists and toxicologists today. Computational approaches have been considered appropriate methods for prior safety assessment of the chemicals. The present study is an attempt to investigate the chemical attributes of a wide variety of ILs towards their inhibitory potential of acetyl cholinesterase enzyme (AChE) through the development of predictive qualitative and quantitative structure–activity relationship (SAR) models in light of the OECD principles. Here, machine learning based cascade correlation network (CCN) and support vector machine (SVM) SAR models were established for qualitative and quantitative prediction of the AChE inhibition potential of ILs. Diversity and nonlinearity of the considered dataset were evaluated. The CCN and SVM models were constructed using simple descriptors and validated with external data. Predictive power of these SAR models was established through deriving several stringent parameters recommended for QSAR studies. The developed SAR models exhibited better statistical confidence than those in the previously reported studies. The models identified the structural elements of the ILs responsible for the AChE inhibition, and hence could be useful tools in designing of safer and green ILs.

1. Introduction

Ionic liquids (ILs) constitute an important class of chemicals and are considered as green solvents (more environmentally safe) for various industrial applications. These are the salts with a melting temperature below the boiling point of water [1]. These compounds exhibit several beneficial properties such as negligible vapor pressure, high heat capacity, high thermal conductivity, high thermal stability, a wide temperature range of liquids, and relatively low toxicities. The properties of ILs could be altered by modifying the structures of the cations or anions. However, due to their high aqueous solubility, these compounds easily disperse into the aquatic media, and may have significant detrimental ecotoxicological consequences for aquatic organisms. Recently, the biological effects of a number of ILs have been reported on various aquatic organisms [2–5]. This warrants for a systematic toxicological assessment of such chemicals already in use and of those yet coming to the market. Inhibition of the acetyl cholinesterase enzyme (AChE) is considered an appropriate test for toxicity assessment of the chemicals. The AChE plays an important role in the nerve response and function which catalyzes the degradation of the neurotransmitter acetylcholine. An inhibition of the AChE leads to various adverse effects in neuronal processes [6]. AChE inhibitors are used in the treatment of various neuro-muscular disorders and have provided the first generation of drugs for the treatment of several diseases [7]. Therefore, AChE has been frequently used in cytotoxicity assays of the ILs [8]. Hence, the design of ILs with a reduced propensity towards inhibition of AChE can improve their greenness in the maintenance of a suitable chemical system. Compared to the huge number of ILs, the toxicity data in the literature are relatively scarce. Although, in vivo or in vitro risk assessments have significantly improved nowadays, these methods are very time and resource consuming beside this fact that these do not respond to the large number of different chemicals. Thus, the aid of the computational methods can be especially helpful to estimate the technical properties of candidate molecules within safe boundaries. Therefore, the models capable to estimate the expected toxicity of an IL formed by the combination of certain cation and anion can be useful screening tools within the IL design process. The quantitative structure–activity relationship (QSAR) study provides a good alternative approach of developing predictive methods by correlating the molecular chemistry with an end-point of interest. However, there is a paucity of predictive modeling studies of ILs on this end-point, AChE inhibition. Only a few previous studies [9–12] reported linear (multiple linear regression, MLR) as well as nonlinear (artificial neural networks, ANNs) regression

* Corresponding author. Tel.: +91 522 4069006; Mobile: 09450400200.
E-mail addresses: kpsingh_52@yahoo.com, kunwarpsingh@gmail.com (K.P. Singh).

http://dx.doi.org/10.1016/j.molliq.2015.06.001
0167-7322/

based quantitative models on ILs towards the AChE inhibition data. However, most of the studies were based on limited data and lacked in proper validation, thus limiting their scope for future applications. However, most critical limitations of many QSAR studies are their low external predictive power when applied to predict toxicity of new chemicals not included in model building phase, which could be due to the incorrect usage or lack of external validation during the modeling process. The OECD principle 4 stresses for the importance and need for the statistical validation of the QSARs [13]. Accordingly, several useful parameters for statistical checks on the robustness and predictivity of the QSARs have been proposed in the literature [14].

Therefore, there is a need for developing reliable SAR models for qualitative and quantitative prediction of AChE inhibition by the ILs properly validated for screening the chemicals. In recent years, several machine learning (ML) methods such as the cascade correlation network (CCN) and support vector machines (SVMs) have emerged as tools for establishing QSARs for various data analyses. CCN is a new architecture and supervised learning algorithm for ANNs. It learns very quickly and the network determines its own size and topology [15], and retains the structures, it has built even if the training set changes, and it requires no back propagation of error signals through the connections of the network. The SVM is a relatively new method based on the structural risk minimization (SRM) that has emerged as one of the techniques for the pattern classification and function approximation [16]. SVM can simultaneously minimize estimation errors and model dimensions. It has good generalization ability and is less prone to over-fitting. SVM adopts kernel functions which make the original nonlinear inputs linearly separable in mapped high dimensional feature space [17].

In this study, we established the QSARs (CCN, SVMs) based on the ML approaches for qualitative and quantitative prediction of the AChE inhibition of ILs using simple descriptors directly derived from the chemical structure. Accordingly, the classification and regression models were constructed to predict the toxicity classes (four-categories) and the toxicity end-point (log EC_{50}) of the diverse ILs using a set of selected molecular descriptors as estimators. The predictive and generalization abilities of the SAR models constructed here were evaluated using several statistical criteria parameters and the predictive power of these models was tested using external dataset.

2. Materials and methods

The qualitative and quantitative SAR models here were developed for predicting the AChE inhibition activity of diverse ILs in accordance with the OECD guidelines [13] for the QSAR model development. Accordingly strict rules were followed for a definite dataset selection with a defined end-point (principle 1), an easily explainable model building strategy (principle 2), a defined applicability domain (principle 3), appropriate validation strategies corresponding to the goodness of fit, robustness and predictivity (principle 4), and finally offering a possible mechanistic interpretation of the developed models (principle 5).

2.1. Dataset

A set of 292 diverse ILs containing both the qualitative and quantitative toxicities (log EC_{50}, μM) data towards the inhibition of *Electrophorus electricus* AChE has been taken from the literature [8] and used for the predictive QSAR modeling. The toxicity values considered here were determined by Ellman's method [18]. In this dataset, the majority of ILs belong to the imidazolium, pyridinium, pyrrolidinium, morpholinium, piperidinium, quinolinium, ammonium, phosphonium based salts, together with a range of anions. In qualitative SAR analysis, all ILs were considered, whereas in quantitative analysis, 232 compounds, for which the exact end-point values are reported, were used. In the dataset, the toxicity values (log EC_{50}, μM) ranged between 0.30 and 3.11. A histogram of the experimental toxicity values plotted in Fig. 1 shows a near normal distribution pattern.

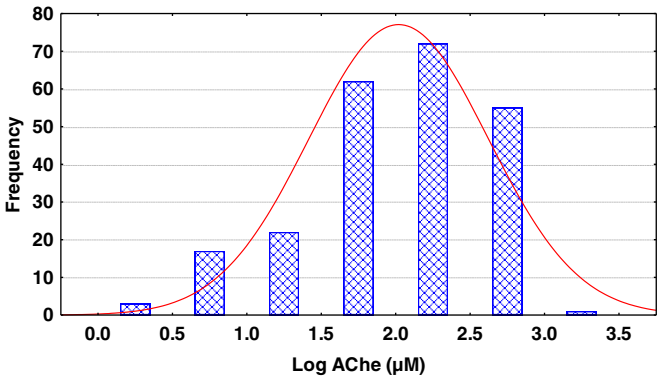

Fig. 1. Histogram of AChE inhibition values of ILs in complete dataset.

2.2. Molecular descriptors

The molecular descriptors of the considered ILs were computed using the Moses Descriptor Community Edition [19]. Here, a total of 211 descriptors (physico-chemical, constitutional, geometrical, topological, and spatial) were calculated for each IL. Among these, descriptors with low variation (<1.0) were excluded from the pool. Model fitting approach was then adopted for the selection of relevant descriptors for QSAR development. Both the qualitative and quantitative SAR models were trained using the remaining set of features computing the respective scoring functions (classification accuracy and mean squared error) to rank the contribution of the features in the current set. The lowest ranked features were then removed [20]. The SAR models were developed by using the remaining set of features, and the corresponding prediction accuracies were computed by means of 10-fold cross validation (CV). A final set of four descriptors were retained in both the qualitative and quantitative SAR modeling (Table 1). The range of the selected descriptors is presented in the radar chart (Fig. 2).

2.3. Data characteristics and processing

The structural diversity of the considered chemicals and nature of the end-point toxicity dependence on the estimators are two important factors to be considered in QSAR studies. The structural diversity of the ILs in the considered dataset was assessed using the Tanimoto Similarity Index (TSI) calculated based on the molecular descriptors. TSI for topology-based chemical similarity studies is calculated using Tanimoto similarity between the fingerprint of a chemical and a consensus fingerprint, which is 1024 bit fingerprint (Toxmatch, Ideaconsult Ltd.). The fingerprint generation is based on the fingerprint implementation of the open source chemoinformatics library [21]. A good cutoff for chemically similar molecules is 0.8. Smaller TSI (<0.8) means that compounds have good diversity [20]. A histogram showing the distribution of the TSI values of the considered ILs in AChE data set is given in Fig. 3. The average TSI values of the ILs (0.04) in the data suggest that the compounds considered in this work represent sufficiently high structural diversity.

On the other hand, the linear or nonlinear dependence in data allows selecting the appropriate modeling approach. Prior to the SAR model development, nonlinearity in the experimental toxicity data was tested using the Brock–Dechert–Scheinkman (BDS) statistics [22]. It tests the null hypothesis of independent and identically distributed (I.I.D.) data against an unspecified alternative. If the computed BDS statistics exceed the critical value at the conventional level, the null hypothesis of linearity is rejected, which reveals the presence of