Chemosphere 122 (2015) 280–287

Contents lists available at ScienceDirect

Chemosphere

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

In silico prediction of chemical toxicity on avian species using chemical category approaches

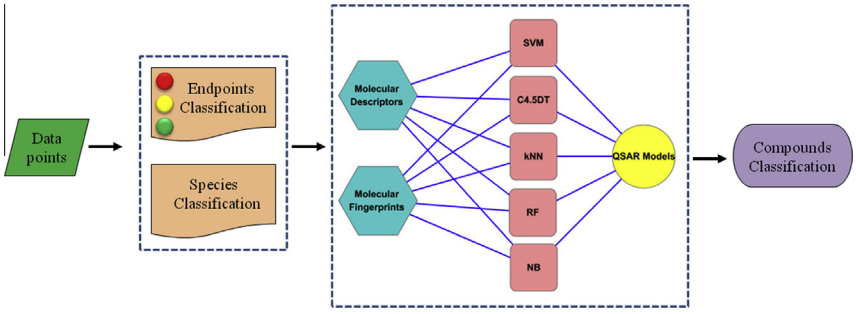
Chen Zhang [a], Feixiong Cheng [a,1], Lu Sun [a], Shulin Zhuang [b], Weihua Li [a], Guixia Liu [a], Philip W. Lee [a,*], Yun Tang [a,*]

[a] Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China
[b] Institute of Environmental Sciences, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China

HIGHLIGHTS

- Robust classification models were developed by machine learning methods.
- Different avian toxicity data points were discussed by category approaches.
- Privileged substructures were identified via the information gain analysis.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 6 July 2014
Received in revised form 28 November 2014
Accepted 1 December 2014
Available online 19 December 2014

Handling Editor: J. de Boer

Keywords:
Avian toxicity
In silico prediction
Chemical category approach
Support vector machine
Information gain

ABSTRACT

Avian species are sensitive to pesticides and industrial chemicals, and hence used as model species in evaluation of chemical toxicity. In present study, we assessed the toxicity of more than 663 diverse chemicals on 17 avian species. All the chemicals were classified into three categories, i.e. highly toxic, slightly toxic and non-toxic, based on the toxicity classification criteria of the United States Environmental Protection Agency (EPA). To evaluate these chemicals, the toxicity prediction models were built using chemical category approaches with molecular descriptors and five commonly used fingerprints, in which five machine learning methods were performed on two standard test species: aquatic bird mallard duck and terrestrial bird northern bobwhite quail. The support vector machine (SVM) method with Pubchem fingerprint performed best as revealed by 5-fold cross-validation and the external validation set on Japanese quail. No species difference existed in our database despite several chemicals with different toxicity on some avian species. The best model had an overall accuracy at 0.851 for the prediction of toxicity on avian species, which outperformed the work of Mazzatorta et al. Furthermore, several representative substructures for characterizing avian toxicity were identified via information gain (IG) method. This study would provide a new tool for chemical safety assessment.

1. Introduction

Avian species are sensitive to industrial chemicals and pesticides, and hence are used as model organisms to evaluate chemical toxicity in the ecotoxicological field. Birds are exposed to toxic substances directly through food or dermal exposure, preening and grooming. Oral intake is considered as the most significant route

* Corresponding authors. Tel.: +86 21 64251052; fax: +86 21 64251033.
E-mail addresses: philiplee2007@gmail.com (P.W. Lee), ytang234@ecust.edu.cn (Y. Tang).
[1] Current address: Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, USA.

http://dx.doi.org/10.1016/j.chemosphere.2014.12.001
0045-6535/

of exposure for avian species. Therefore, test for oral toxicity is one of the most important steps in determining the toxicological significance of any compound under investigation on avian species. Tests are usually conducted on either northern bobwhite quail (*Colinus virginanus*) or mallard duck (*Anas platyrhynchos*); while Japanese quail, red-necked pheasant, house sparrow and others can also be used as alternatives. However, a comprehensive study of the mechanism of avian oral toxicity is still missing. In particular, there is little study on the relationships between chemical structures and their avian toxicity (Mazzatorta et al., 2006).

However, *in vitro* or *in vivo* evaluation of a large number of compounds is costly and time-consuming. Thus, it is necessary to develop alternative methods and tools for toxicity assessment. Computational (*in silico*) techniques, such as quantitative structure–activity relationship (QSAR), provide suitable methods for early evaluation in the development of new chemicals (Cronin et al., 2003; Jaworska et al., 2003). Hence, computational toxicology significantly reduces the cost of experimental toxicity assessment and accelerates the environmental hazard assessment (Hengstlera et al., 2006).

Recently, several models were built to evaluate the environmental hazard assessment using QSAR models or related computational techniques (Cheng et al., 2011, 2012a). The latest researches on evaluating avian toxicity were done by Toropov and Benfenati (2006) and Mazzatorta et al. (2006). Toropov and Benfenati used graphs of atomic orbitals (GAOs) to represent molecular structures and built linear QSAR models to evaluate quail dietary toxicity, which used 110 pesticides data for the quail toxicity dietary exposure and randomly split into a training set ($n = 91$) and a test set ($n = 19$). After evaluating different rules to convert the labeled hydrogen-filled graphs (LHFGs) into GAOs by correlation coefficient of each model, the models based on the optimization of correlation weights of local invariants (OCWLI) of GAO are better than those based on the OCWLI of the LHFGs. Mazzatorta et al. collected 116 pesticides with oral toxicity LD_{50} data for the bobwhite quail from several sources. Then the final model was obtained using support vector machines combined with genetic algorithms for feature selection, after calculating physico-chemical and structural descriptors using OpenMolGRID (Sild et al., 2006). The model had a good predictive ability with 0.021 error rate for the training set and 0.158 error rate for the validation set. However, all the published models were based on a small number of chemicals and had limited applicability domains. Meanwhile, avian species exposure in different chemicals, the toxicity level of the chemicals, and chemical cross species toxicity were not well explained. Therefore, it is urgent to develop new computational methods and QSAR models for avian toxicity assessment.

In this study, a large data set containing 663 diverse compounds with eight-day dietary LC_{50} (Lethal Concentration 50, defined as the concentration of a chemical to cause death in 50% of tested animals) values for different bird species was collected. Both molecular descriptors and fingerprints were then calculated to represent the compounds. Afterwards, binary classification models were developed using five machine learning methods based on the US EPA toxicity classification criteria. Meanwhile, different toxicity data points for the avian species and cross species toxicity of the chemicals were discussed based on the predictive models. Furthermore, the information gain (IG) method (Shen et al., 2010) was applied to identify privileged substructures that might be responsible for the avian toxicity. Our study provided a useful tool for chemical safety assessment.

2. Materials and methods

The workflow of this study was shown in Fig. S1 of Supplementary Material S4.

2.1. Data preparation

Oral toxicity data of chemicals on 17 avian species, in LC_{50} values tested in 8-day dietary, were collected from the latest version (06_14_2013) of the EPA Ecotox database (http://cfpub.epa.gov/ecotox/) and the EPA Pesticide database May 2004.

For each avian species, the data set was divided into three classes, namely highly toxic, slightly toxic and nontoxic, based on the EPA toxicity classification criteria (http://www.epa.gov/oppefed1/ecorisk_ders/toera_analysis_eco.htm#content) (shown in Table 1) and labeled in traffic lights as red, yellow and green, respectively. Duplicated compounds of each avian species were removed. For a given compound, if it has several data points for the same avian species, the most toxic one, namely the smallest LC_{50} value, was kept. If several LC_{50} values are conflicted, the compound will be double checked in the IUPAC Pesticide Properties Database (PPDB). If the values were ambiguous, all the data points were removed. Inorganic compounds were also omitted and salt fragments of the compounds were removed. Finally, only those molecules with molecular weight greater than 40 but less than 800 were kept in the data set (Xu et al., 2012).

2.2. Molecular representation

1D, 2D molecular descriptors and five types of molecular fingerprints were calculated using PaDEL-Descriptor software (Yap, 2011). The five fingerprints are CDK Fingerprint (FP), Estate Fingerprint (Estate), MACCS Fingerprint (MACCS), Pubchem Fingerprint (Pubchem) and Substructure Fingerprint (SubFP). The detailed description of these descriptors and fingerprints can be found in their original literature (Klekota and Roth, 2008; Yap, 2011). The F-score method (Chen and Lin, 2006) was used to select correlated molecular descriptors, which was validated in our previous studies (Cheng et al., 2012b). The larger the F-score value is, the more likely the descriptor is discriminative.

2.3. Machine learning methods

Five machine learning methods including support vector machine (SVM), C4.5 decision tree (C4.5 DT), *k*-nearest neighbors (*k*NN), random forest (RF) and Naive Bayes (NB) were used for model building. The SVM algorithm was provided by the open source LIBSVM (LIBSVM2.9 package) (Chang and Lin, 2011). The others were performed using software Orange (version 2.6a2, freely available at http://orange.biolab.si/).

2.3.1. Support vector machine (SVM)

The SVM algorithm was applied to build the classification models using both molecular descriptors and fingerprints. The common used kernel function radial basis function was used to seek the penalty parameter C and different kernel parameter γ, using grid search strategy based on a 5-fold cross-validation in classification models.

2.3.2. C4.5 decision tree (C4.5 DT)

Decision tree is a hierarchical structure composed of nodes and directed edges. The version used here is Quinlan's C4.5 (Salzberg, 1994) which was incorporated in Orange. All parameters of C4.5 in Orange used the default values.

Table 1
Classification criteria of chemical avian toxicity.

LC_{50} (mg kg^{-1})	Label	Class
$\leqslant$500	Red	Highly toxic
(500, 2000]	Yellow	Slightly toxic
>2000	Green	Nontoxic