Chemometrics and Intelligent Laboratory Systems 110 (2012) 177–181

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Short Communication

CORAL: QSAR modeling of toxicity of organic chemicals towards *Daphnia magna*

A.P. Toropova [a], A.A. Toropov [a,*], S.E. Martyanov [b], E. Benfenati [a], G. Gini [c], D. Leszczynska [d], J. Leszczynski [e]

[a] *Istituto di Ricerche Farmacologiche Mario Negri, 20156, Via La Masa 19, Milano, Italy*
[b] *Teleca OOO, 603093, 23, Rodionova st, Nizhny Novgorod, Russia*
[c] *Department of Electronics and Information, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italy*
[d] *Interdisciplinary Nanotoxicity Center, Department of Civil and Environmental Engineering, Jackson State University, 1325 Lynch St, Jackson, MS 39217-0510, USA*
[e] *Interdisciplinary Nanotoxicity Center, Department of Chemistry and Biochemistry, Jackson State University, 1400 J. R. Lynch Street, P.O. Box 17910, Jackson, MS 39217, USA*

ARTICLE INFO

Article history:
Received 24 August 2011
Received in revised form 4 October 2011
Accepted 7 October 2011
Available online 15 October 2011

Keywords:
QSAR
SMILES
Molecular graph
CORAL software
Toxicity to *Daphnia magna*

ABSTRACT

Convenient to apply and available on the Internet, CORAL software (http://www.insilico.eu/CORAL) has been used to build up quantitative structure–activity relationships (QSAR) for prediction of toxicity to *Daphnia magna*. The QSARs developed in this study are one-variable models based on the optimal descriptors calculated with the Monte Carlo method. The toxicity has been modeled with the following representations of the molecular structure: (i) by hydrogen-suppressed graph (HSG); (ii) by simplified molecular input line entry system (SMILES); and (iii) by hybrid representation, i.e. the HSG together with SMILES. Four random splits into the sub-training, calibration, and test sets were examined. The hybrid version of the representation of the molecular structure provided the best accuracy of the prediction for the considered endpoint.

1. Introduction

Toxicity of a compound towards *Daphnia magna* represents well-known and important ecological indicator of potential environmental hazards of chemicals [1–8]. Experimental evaluations of environmental properties (e.g. toxicity to *D. magna*) for all new synthetic substances which are being used in everyday life (marketing, cosmetics, medicine, industry, etc.) become impossible owing to increasing number of these substances. Under such circumstances, the development of efficient quantitative structure–activity relationships (QSAR) represents the only compelling alternative to the experiment.

European Union REACH (Registration, Evaluation and Authorisation of Chemicals) explicitly encourages the use of computational methods for estimation of environmental parameters of all new and existing chemicals. Obviously, QSAR will play an important role in addressing of this task [9–14].

Recently, CORAL software has been suggested as an efficient tool for the QSAR analysis [15]. The CORAL models represent one-variable correlations between an endpoint and optimal descriptors. The optimal descriptors are calculated with special coefficients related to presence of various molecular features (molecular fragments and physicochemical characteristics of molecules). These coefficients (correlation weights) are obtained by the Monte Carlo method. One can use as the representation of the molecular structure for the optimal descriptors hydrogen-suppressed molecular graph (HSG) [16], simplified molecular input line entry system (SMILES) [17–19], or a hybrid representation which includes both the HSG and SMILES.

The comparison of aforementioned three representations of the molecular structure in the development of QSAR approaches devoted to toxicity towards *D. magna* is the aim of the present study.

2. Methods

2.1. Data

The descriptions of organic chemicals related to 48 h *D. magna*-toxicity expressed in negative decimal logarithm of the dose that kills 50% of organisms i.e. pLC50 were taken from the literature [1]. The data set covers range of octanol/water partition coefficient from −2 to 8. The range of toxicity (daphnia) is from 0.46 to 10.09. In regard to the chemical domain, the data set includes hydrocarbons, aliphatic alcohols, phenols, ethers, and esters; anilines, amines, nitriles, nitroaromatics, amides, and carbamates; urea and thiourea derivatives; iso-thiocyanates; thiols; phosphorothionate and phosphate esters; and halogenated derivatives. The list of compounds represented by CAS numbers and SMILES with their daphnia toxicity values are shown in Supplementary Materials.

2.2. Molecular descriptors

CORAL software can generate three kinds of optimal descriptors: graph-based, SMILES-based, and hybrid descriptors which are calculated

* Corresponding author.
E-mail address: andrey.toropov@marionegri.it (A.A. Toropov).

0169-7439/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2011.10.005

with both graph and SMILES. Accordingly, the CORAL software can generate three kinds of molecular graphs: the above-mentioned HSG, hydrogen filled graph (HFG), and graph of atomic orbitals (GAO).

The graph-based optimal descriptors are calculated as the following:

$$^{Graph}\mathrm{DCW}\left(\mathrm{Threshold}, N_{\mathrm{epoch}}\right) = \sum \mathrm{CW}(A_k) + \alpha \sum \mathrm{CW}\left({}^{0}EC_k\right) + \beta \sum \mathrm{CW}\left({}^{1}EC_k\right) + \gamma \sum \mathrm{CW}\left({}^{2}EC_k\right) + \delta \sum \mathrm{CW}\left({}^{3}EC_k\right) \quad (1)$$

where A_k is chemical element, such as, C, N, O, etc., for HSG and HFG; or atomic orbitals, such as $1s^1$, $2p^3$, $3\,d^{10}$, etc., for GAO; ${}^{0}EC_k$, ${}^{1}EC_k$, … ${}^{3}EC_k$ represents the hierarchy of the Morgan extended connectivity; α, β, γ, and δ can be 1 or 0: combinations of their values gives possibility to define various versions of the graph-based optimal descriptor; CW(x) is the correlation weight of a molecular feature (encoded by A_k or ${}^{x}EC_k$).

The SMILES-based optimal descriptors are calculated as the following:

$$^{SMILES}\mathrm{DCW}(\mathrm{Threshold}, N_{\mathrm{epoch}}) = \alpha \sum \mathrm{CW}(S_k) + \beta \sum \mathrm{CW}(SS_k) + \gamma \sum \mathrm{CW}(SSS_k) + x \cdot \mathrm{CW}(NOSP) + y \cdot \mathrm{CW}(\mathrm{HALO}) + z \cdot \mathrm{CW}(\mathrm{BOND}) \quad (2)$$

where S_k, SS_k, and SSS_k are one-, two-, and three-component SMILES attributes, respectively; the component of SMILES represents one symbol (e.g. C, c, N, n, =, #, etc.) or two symbols which cannot be separated (e.g. Cl, Br, @@, etc.); NOSP, HALO, and BOND are indices calculated according to presence or absence of chemical elements: nitrogen, oxygen, sulfur, and phosphorus (NOSP); fluorine, chlorine, and bromine (HALO). The BOND symbolizes a mathematical function related to the presence or absence of double (=), triple (#), or stereo chemical bonds (@ or @@); α, β, γ, x, y, and z can be 1 or 0: combinations of their values provide possibility to define various versions of the SMILES-based optimal descriptor. CW(x) is the correlation weight of a molecular feature (encoded by $S_kS_kS_k$ or ${}^{x}EC_k$).

The hybrid optimal descriptors are calculated with taking into account both representations of the molecular structure by graph and by SMILES.

$$^{Hybrid}\mathrm{DCW}(\mathrm{Threshold}, N_{\mathrm{epoch}}) = {}^{SMILES}\mathrm{DCW}(\mathrm{Threshold}, N_{\mathrm{epoch}}) + {}^{Graph}\mathrm{DCW}(\mathrm{Threshold}, N_{\mathrm{epoch}}) \quad (3)$$

Threshold and N_{epoch} (in Eqs. (1)–(3)) are parameters of the Monte Carlo optimization. Threshold is criterion for classification of components of the representation of the molecular structure into two classes: rare (noise) and active (not rare). The correlation weight of a rare component is fixed as zero; hence rare component is not involved in the building up of the model. N_{epoch} is the number of epochs of the Monte Carlo optimization. Fig. 1 shows the theoretical influence of the threshold and of the number of epochs of the Monte Carlo method optimization for the correlation coefficient between the experimental and calculated values of an endpoint.

One can see (Fig. 1) that the increase of the threshold is accompanied by decrease of the correlation coefficient between experimental and calculated values of an endpoint for the sub-training and calibration set, whereas the correlation coefficient for the external test set has a maximum (Threshold = 2). The increase of the number of epochs of the Monte Carlo method optimization is accompanied by

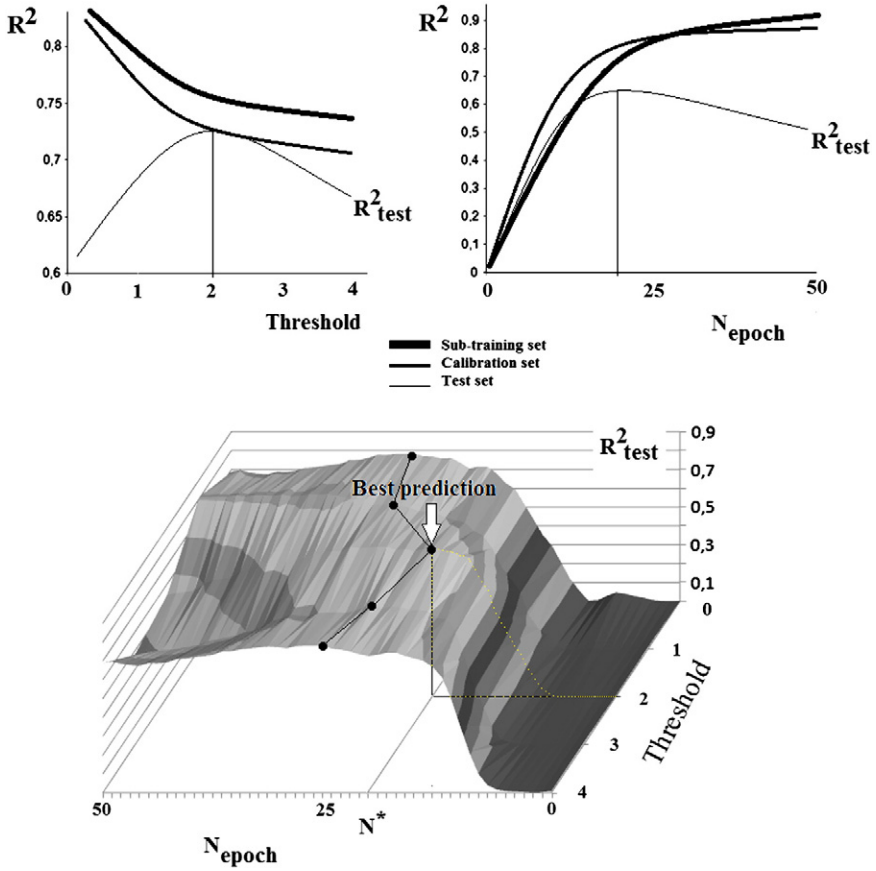

Fig. 1. Correlation coefficient between experimental and predicted values of an endpoint for the external test set as a mathematical function of the threshold and the number of epochs of the Monte Carlo method optimization.