Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Research paper

The Monte Carlo technique as a tool to predict LOAEL

Jovana B. Veselinović [a], Aleksandar M. Veselinović [a], Alla P. Toropova [b, *], Andrey A. Toropov [b]

[a] University of Niš, Faculty of Medicine, Department of Chemistry, Niš, Serbia
[b] IRCCS- Istituto di Ricerche Farmacologiche Mario Negri, Milano, Italy

ARTICLE INFO

Article history:
Received 3 June 2015
Received in revised form
24 March 2016
Accepted 25 March 2016
Available online 30 March 2016

Keywords:
QSAR
LOAEL
SMILES
Ecology
Drug toxicity
Optimal descriptor

ABSTRACT

Quantitative structure – activity relationships (QSARs) for the Lowest Observed Adverse Effect Level (LOAEL) for a large set of organic compounds (n = 341) are suggested. The molecular structures of these compounds are represented by Simplified Molecular Input-Line Entry Systems (SMILES). A criteria for the estimation quality of split into the "visible" training set (used for developing a model) and "invisible" external validation set is suggested. The correlation between the above criterion and the predictive potential of developed QSAR model (root-mean-square error for "invisible" validation set) has been detected. One-variable models are built up for several different splits into the "visible" training set and "invisible" validation set. The statistical quality of these models is quite good. Mechanistic interpretation and the domain of applicability for these models are defined according to probabilistic point of view. The methodology for defining applicability domain in QSAR modeling with SMILES notation based optimal descriptors is presented.

1. Introduction

In recent years a considerable efforts have been made to assess genotoxic effect of various pharmaceutical products as well as industrial pollutants in general, because nowadays, people in the majority of industrial countries are under the influence of various substances [1,2]. Toxic effects that substance can show are different and they include an adverse alteration of morphology, function, capacity, growth, development, or lifespan of a target organism distinguished from normal organisms of the same species under defined conditions of exposure. It is inconvenient to use human for biochemical and/or medicinal observations and therefore databases on potential risk of different substances are gradually increasing with experiments on animals [3,4]. However, all experiments with animals have serious ethical issues. On the other hand, the definition of endpoints which are the reliable measure of the harmfulness of substances is a task which requires long time and expensive equipment [5]. Further, chronic studies are designed to obtain a dose–response covering overt toxic effects, mild effects (the Lowest Observed Adverse Effect Level, LOAEL), and no effects (the No Observed Effect Level, NOAEL). The numerical data on these endpoints (LOAEL and NOAEL) are not available for hundreds of thousands or millions of substances which can enter the food chain and result in human exposure. For all stated reasons, the risk assessment in the absence of sufficient experimental data is a challenge for scientists, so the search for mathematical approaches which are capable to estimate the harmfulness of various substances (without direct experiment) is an attractive alternative of the experimental definition of risk assessment [6,7].

The quantitative structure – property/activity relationships (QSPRs/QSARs) based on the molecular descriptors are a computational tool used to predict various endpoints and they can be used for risk assessment [8–11]. Therefore, QSAR models for these endpoints can be useful from points of view of medicinal chemistry and ecology [25]. Optimal descriptors give possibility to establish specific one-variable QSPR/QSAR model using the Monte Carlo method [12–15]. Recently, the optimal descriptors calculations become available with CORAL software [16], where Simplified Molecular Input-Line Entry System (SMILES) [17–19] were used for representation of the molecular structure [20–24].

The aim of the present study is the estimation of SMILES-based optimal descriptors calculated with the CORAL software as a tool to predict of the LOAEL of various organic compounds. Also, in this

* Corresponding author. Laboratory of Environmental Chemistry and Toxicology, IRCCS – Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy.
E-mail address: alla.toropova@marionegri.it (A.P. Toropova).

http://dx.doi.org/10.1016/j.ejmech.2016.03.075
0223-5234/© 2016 Elsevier Masson SAS. All rights reserved.

research the methodology for defining applicability domain in QSAR modeling with SMILES notation based optimal descriptors is presented.

2. Method

2.1. Data

Experimental data on LOAEL (logarithmic scale, mg/kg body weight per day) were taken from literature [25]. These values were converted into negative decimal logarithm, i.e. the pLOAEL is the endpoint examined in this work. It has to be notated that the database from literature contains large number of duplicates. After the extracting of the duplicates, the total number of compounds available for the QSAR analysis was 341. The supplementary materials contains the lists of compounds involved into building up models (n=341) from the above mentioned source [25]. There is large variety of molecular structures, which contain oxygen, nitrogen, sulphur, phosphorus, chlorine, bromine, iodine, different combines of rings with 3, 5, 6, 7 members, as well as acyclic compounds (Table S1). Five random splits into the training set, invisible training set, calibration set, and the validation set were prepared according to the following principles: (i) these splits are random; (ii) these splits are not identical (Table 1); and (iii) the number of compounds in the external validation set is about 50 or more.

2.2. Optimal descriptors

The Monte Carlo method simulations, based on iterative algorithms, are run for obtaining the distribution of an unknown probabilistic entity. Therefore, Monte Carlo method develops QSAR model by generating suitable random numbers and observing how that fraction of numbers obeys a property or some properties. Further, a numerical correlation weight value (CW) is randomly assigned to SMILES-based descriptors in each independent Monte Carlo run and for a defined endpoint. Descriptors of Correlation Weights (DCW) for SMILES notation are calculated as the following [26]:

$$\mathrm{DCW(T,N)} = \sum \mathrm{CW}(\mathrm{S_k}) + \sum \mathrm{CW}(\mathrm{SS_k}) + \sum \mathrm{CW}(\mathrm{SSS_k}) \qquad (1)$$

where $\mathrm{S_k}$ is SMILES atoms, i.e. one symbol (e.g. ‘C’, ‘N’, ‘=’, etc.) or two symbols which cannot be examined separately (e.g. ‘Cl’, ‘Br’, etc.); $\mathrm{SS_k}$ and $\mathrm{SSS_k}$ are compositions of two and three SMILES atoms, respectively; $\mathrm{CW(S_k)}$, $\mathrm{CW(SS_k)}$, and $\mathrm{CW(SSS_k)}$ are the correlation weights for the $\mathrm{S_k}$, $\mathrm{SS_k}$, and $\mathrm{SSS_k}$, respectively; the numerical data on the correlation weights for above-mentioned SMILES attributes (i.e. $\mathrm{S_k}$, $\mathrm{SS_k}$, and $\mathrm{SSS_k}$) are calculated by the Monte Carlo method where their values should provide the maximum of the target function (TF):

$$\mathrm{TF} = \mathrm{R} + \mathrm{R'} - \mathrm{abs}(\mathrm{R} - \mathrm{R'}) \times \mathrm{dR_w} \qquad (2)$$

where R and R′ are correlation coefficients between pLOAEL and DCW(T,N) for the active training and invisible training sets, respectively; $\mathrm{dR_w}$ (= 0.01) is an empirical constant. The parameter T is a threshold that is used to define rare and active SMILES attributes, e.g. T=3 means that if an attribute ‘x’ is represented only in two (or less) SMILES of the sub-training set, the ‘x’ is rare and CW(x) is fixed equal to zero (i.e. the ‘x’ is not involved in the model).

The parameter N is the number of epochs of the Monte Carlo optimization for the TF which gives maximum of correlation coefficient between LOAEL and DCW(N,T) for test set. The values of T=T* and N=N*, which gives maximum of correlation coefficient between LOAEL and DCW(T,N) for the test set are to be preferable in order to build up a model:

$$\mathrm{pLOAEL} = \mathrm{C_0} + \mathrm{C_1} \times \mathrm{DCW(T^*,N^*)} \qquad (3)$$

The external validation set (no information on these substances is used in the modeling process) is involved in the final checking up of the predictive potential of the model calculated with Eq. (3).

There are two ways to build up a model using the optimal descriptor [26]: (i) classical method, which is based on three sets, namely, training set, calibration set, and validation set; and (ii) balance of correlations which is based on four sets, namely, training set, invisible training set, calibration set, and validation

Table 1
The percentage of identity for splits 1–5 and defects of distributions “Sub-training/Test” together with defects of distributions “Sub-training/Validation” (indicated by bold).

Split	Set	Defect	n	Split 1	Split 2	Split 3	Split 4	Split 5
1	Training	216.9	111	100*	80.9	30.9	39.5	29.9
	Invisible training		126	100	81.1	38.7	35.8	36.0
	Calibration		52	100	15.1	24.8	27.3	21.0
	Validation		52	100	16.8	15.4	27.5	18.9
2	Training	202.0	114		100	35.0	38.1	33.0
	Invisible training		118		100	35.9	36.2	30.6
	Calibration		54		100	16.8	21.4	22.4
	Validation		55		100	22.4	12.5	18.3
3	Training	188.9	109			100	33.5	38.4
	Invisible training		127			100	35.7	36.7
	Calibration		53			100	19.8	20.8
	Validation		52			100	16.5	18.9
4	Training	217.9	112				100	32.4
	Invisible training		114				100	29.4
	Calibration		58				100	23.4
	Validation		57				100	12.6
5	Training	218.9	110					100
	Invisible training		124					100
	Calibration		53					100
	Validation		54					100

*) $Identity(\%) = \frac{N_{i,j}}{0.5*(N_i+N_j)} \times 100$
where

$N_{i,j}$ is the number of substances which are distributed into the same set for both i-th split and j-th split (set =sub-training, calibration, test, validation).
N_i is the number of substances which are distributed into the set for i-th split.
N_j is the number of substances which are distributed into the set for j-th split.