Chemical Physics Letters 701 (2018) 137–146

Contents lists available at ScienceDirect

Chemical Physics Letters

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

Research paper

Application of the Monte Carlo method for building up models for octanol-water partition coefficient of platinum complexes

Andrey A. Toropov, Alla P. Toropova *

IRCCS, Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy

ARTICLE INFO

Article history:
Received 25 January 2018
In final form 4 April 2018
Available online 25 April 2018

Keywords:
QSPR
Octanol/water partition coefficient
Pt(II) and Pt(IV) complexes
Index of ideality of correlation
Monte Carlo method

ABSTRACT

Predictive model of logP for Pt(II) and Pt(IV) complexes built up with the Monte Carlo method using the CORAL software has been validated with six different splits into the training and validation sets. The improving of the predictive potential of models for six different splits has been obtained using so-called index of ideality of correlation. The suggested models give possibility to extract molecular features, which cause the increase or vice versa decrease of the logP.

1. Introduction

Platinum complexes are a well-known class of substances used in cancer treatment [1]. There is a large variety in the pharmacological behavior of platinum complexes [1,2]. The logP is the useful information about behaviour of platinum complexes from point of view of drug discovery. This stimulates works dedicated to establishing of quantitative structure - property relationships (QSPRs) for logP of platinum complexes [3–7], e.g. the data on logP is used in drug discovery related to anti-HIV activity [8]; for treatment of skin diseases [9]; for establishing of the anticancer potential of crown ethers [10]; for searching of new antibiotics [11]; etc. It should be noted, that lipophilicity, expressed as the octanol-water partition coefficient, constitutes the most important property in drug action, influencing both pharmacokinetic and pharmacodynamics processes as well as drug toxicity [12]. Computational models for logP can be attractive alternative of the experimental measurements at least as preliminary estimation of the parameter for new substances. Thus, the development of theoretical computational methods, which can predict the logP is the important task of contemporary science [13–22].

The estimation of predictive potential of models for logP of platinum complexes, which are built up using the Monte Carlo technique via the CORAL software (www.insilico.eu/coral), is the aim of the present study.

2. Method

2.1. Data

The experimental data on logP for Pt(II) and Pt(IV) complexes (n = 46) together with simplified molecular input-line entry system (SMILES) [23,24] were taken in the literature [1]. The total set has been six times randomly split into the training (≈25%), invisible training (≈25%), calibration (≈25%), and validation (≈25%) sets. Table 1 shows that these six splits are not identical [25].

2.2. Optimal descriptor

QSPR-models of the logP for Pt(II) and Pt(IV) complexes suggested in this work are calculated with descriptors based on correlation weights (CW) of molecular features extracted from SMILES:

$$logP = C_0 + C_1 * DCW(T*, N*) \quad (1)$$

The optimal descriptor used in this work is calculated as the following:

$$DCW(T*, N*) = \sum CW(S_k) + \sum CW(SS_k) \quad (2)$$

The S_k are SMILES-atoms, i.e. one symbol of the SMILES notation (e.g. ‘c’, ‘C’, ‘N’, etc.) or two symbols which cannot be examined separately (e.g. ‘Cl’, ‘Br’, etc.). The SS_k is combine of two SMILES-atoms. The $CW(Sk)$ and $CW(SSk)$ are correlation weights which are involved to calculate descriptor with Eq. (2). Table 2 contains an example of $CW(Sk)$ and $CW(SSk)$.

* Corresponding author at: 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).

https://doi.org/10.1016/j.cplett.2018.04.012
0009-2614/© 2018 Elsevier B.V. All rights reserved.

Table 1
Percentage of identity for six random splits into the training and validation sets.

	Set	Split 2	Split 3	Split 4	Split 5	Split 6
Split 1	Training	17.4*	18.2	26.1	27.3	26.1
	Invisible training	16.7	33.3	26.1	16.7	34.8
	Calibration	8.7	25.0	26.1	26.1	8.3
	Validation	9.1	27.3	26.1	17.4	36.4
Split 2	Training	*100*	17.4	25.0	26.1	8.3
	Invisible training	*100*	33.3	26.1	41.7	17.4
	Calibration	*100*	17.4	18.2	36.4	8.7
	Validation	*100*	18.2	43.5	34.8	18.2
Split 3	Training		*100*	17.4	9.1	17.4
	Invisible training		*100*	26.1	25.0	34.8
	Calibration		*100*	26.1	34.8	50.0
	Validation		*100*	26.1	26.1	18.2
Split 4	Training			*100*	26.1	33.3
	Invisible training			*100*	34.8	36.4
	Calibration			*100*	63.6	26.1
	Validation			*100*	41.7	17.4
Split 5	Training				*100*	34.8
	Invisible training				*100*	17.4
	Calibration				*100*	17.4
	Validation				*100*	17.4

*$Identity\,(\%) = \frac{N_{i,j}}{0.5*(N_i+N_j)} \times 100$
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 = training, invisible training, calibration, and 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.

The T is threshold to separate the S_k and SS_k into the categories: (i) rare; and (ii) non-rare. The rare attributes are not used to build up a model: their correlation weights are equal to zero. The N is the number of epochs of the Monte Carlo optimization. One epoch is sequence of modifications of all correlation weights of non-rare (according to selected threshold) SMILES attributes. The T = T* and N = N* are numerical data on these parameters which provide the best statistical quality of the model for the calibration set [25–27]. The S_k and SS_k are described in the literature [26].

Each the above-mentioned set (Table 1) is aimed to solve the “individual” task:

(i) The training set is aimed to calculate correlation weights, which give maximal correlation coefficient between *DCW (T*,N*)* and logP for complexes distributed into the training set;
(ii) The invisible training set is aimed to confirm that the correlation “*DCW(T*,N*)* – logP” is more or less satisfactory for similar substances which are not involved in the training set;
(iii) The calibration set is aimed to inform about the moment of the beginning of overtraining; and
(iv) The validation set is aimed to estimate factual predictive potential of the model for substances which are unknown during the optimization process. Thus, factually, the training, invisible training, and calibration sets are the structured training set.

Table 3 shows an example of calculation of the *DCW(T*,N*)*. The above-mentioned correlation weights are calculated with optimization by the Monte Carlo method. The target function of the optimization can be [25]:

$$TF = R + R' - |R - R'| \times 0.1 \tag{3}$$

The R and R' are correlation coefficients between experimental and predicted endpoint for the training and invisible training sets. The Monte Carlo optimization provides numerical data on the correlation weights $CW(S_k)$ and $CW(SS_k)$ which provide maximum of the *TF*.

In addition, the modified target function can be used for building up the predictive model calculated with Eq. (1) [27,28]:

$$TF_m = TF + IIC \times 0.1 \tag{4}$$

The 0.1 for Eqs. (3) and (4) is defined “empirically” as coefficient that gives more or less satisfactory predictive potential for models of different endpoints [14,23,25,26].

The *IIC* is the so-called index of ideality of correlation (*IIC*) [27,28]. The index is defined according to the following logic. The quality of prediction for one arbitrary compound can be estimated as the following:

$$\Delta_k = observed_k - calculated_k \tag{5}$$

Having data on all Δ_k for the calibration set, one can calculate sum of negative and positive values of Δ_k similar to mean absolute error (MAE):

$$^{-}MAE_{calibration} = \frac{1}{^{-}N}\sum_{k=1}^{^{-}N}|\Delta_k|\Delta_k < 0,\ ^{-}N \text{ is the number of } \Delta_k < 0 \tag{6}$$

$$^{+}MAE_{calibration} = \frac{1}{^{+}N}\sum_{k=1}^{^{+}N}|\Delta_k|\Delta_k \geqslant 0,\ ^{+}N \text{ is the number of } \Delta_k \geqslant 0, \tag{7}$$

The *IIC* is calculated with the following formula:

$$IIC = r_{calibration} \times \frac{\min(^{-}MAE_{calibration},\ ^{+}MAE_{calibration})}{\max(^{-}MAE_{calibration},\ ^{+}MAE_{calibration})} \tag{8}$$

According to Eq. (8), the diapason of *IIC* is (−1, 1). The *IIC* is not identic to traditionally used criteria of predictive potential of QSPR models. Table 3 contains a group of widely used criteria of predictive potential of QSPR models [29–31]. All these criteria obey the principle “larger value of a criterion means better predictive potential”. Consequently, the quality of the choice of model (model-1 or model-2) according to the above-mentioned criteria can be compared. Fig. 1 shows the scheme how to select better model according to one of the listed criteria.