Contents lists available at ScienceDirect

Mutat Res Gen Tox En

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

The index of ideality of correlation: A criterion of predictive potential of QSPR/QSAR models?

Andrey A. Toropov, Alla P. Toropova*

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

ARTICLE INFO

Keywords:
QSAR
Predictive potential
Index of ideality of correlation
Mutagenicity
Salmonella typhimurium
CORAL software

ABSTRACT

The index of ideality of correlation (IIC) is a new criterion of the predictive potential of quantitative structure–property/activity relationships (QSPRs/QSARs). This IIC is calculated with using of the correlation coefficient between experimental and calculated values of endpoint for the calibration set, with taking into account the positive and negative dispersions between experimental and calculated values. The mutagenicity is well-known important characteristic of substances from ecological point of view. Consequently, the estimation of the IIC for mutagenicity is well motivated. It is confirmed that the utilization of this criterion significantly improves the predictive potential of QSAR models of mutagenicity. The new criterion can be used for other endpoints.

1. Introduction

The majority of phenomena of natural sciences are complex. The idealization (or simplification) is one of the most common approaches to studying complex phenomena in the field of natural sciences, e.g. ideal gas [1], ideal solution [2], ideal crystals [3], ideal symmetry [4].

Quantitative structure − property/activity relationships (QSPRs/QSARs) are a specific fragment of natural sciences. The main aim of this field of the science is to predict values of different endpoints for substances, which were not studied in direct experiment.

At the first stages of the QSPR/QSAR-theory, the establishing of correlations between descriptors, which are originated from molecular structure, and endpoints was considered as the main aim of the researches [5,6].

Further QSPR/QSAR studies have shown that real predictive potential of a model for the training set and correlations outside training set sometimes (or even usually) have considerable disagreement [7–9]. The essence of the problem is the unsymmetrical superposition of the dots-images relatively to diagonal in coordinates "observed-predicted" values of an endpoint. The index of ideality of correlation (IIC) is a criterion to estimate the predictive potential of QSPR/QSAR models by means of estimation of the above-mentioned "asymmetry".

The mutagenicity is important characteristic of substances from ecological point of view. Consequently, there are works dedicated to the development of QSAR models to predict this endpoint [10–17]. The aim of this study is estimation of the QSAR models for mutagenicity, which are building up by means of the CORAL software (http://www.insilico.eu/coral) with using the above-mentioned novel criterion of predictability (IIC).

2. Method

2.1. Data

Data on mutagenic potentials of the set of 95 aromatic and heteroaromatic amines were taken from the literature [18]. The mutagenic activity in *Salmonella typhimurium* TA98 + S9 microsomal reparation is expressed as the natural logarithm of R, where R is the number of revertants per nanomole (lnR). Seven splits into the training (≈35%), invisible training (≈35%), calibration (≈15%), and external validation (≈15%), sets are examined in this work. These splits are random and definitely non-identic (Table 1). There are a few compounds, which are characterized by large value for described below delta (Eq. (1)) even if these are distributed in the training set. These compounds are distributed solely into the training or invisible training sets (not into the calibration or validation sets).

2.2. Index of ideality of correlation

Fig. 1 shows possible defects of a QSPR/QSAR model in aspect of their applying to predict values of endpoint. The quality of prediction for one substance from a set can be estimated as the following:

* 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).

http://dx.doi.org/10.1016/j.mrgentox.2017.05.008
Received 29 November 2016; Received in revised form 11 May 2017; Accepted 15 May 2017
Available online 17 May 2017

Table 1
Percentage of identity for splits #1–#7.

	Set	Split2	Split3	Split4	Split5	Split6	Split7
Split 1	Training	30.3	33.8	35.5	26.7	26.7	20.3
	Invisible training	22.6	25.8	30.8	27.3	32.8	38.8
	Calibration	12.9	0.0	25.0	31.3	12.9	18.8
	Validation	0.0	18.8	25.8	12.5	12.5	18.8
Split 2	Training	100	47.9	47.1	33.3	36.4	46.2
	Invisible training	100	34.5	36.1	38.7	19.0	19.0
	Calibration	100	20.0	19.4	25.8	20.0	25.8
	Validation	100	19.4	20.0	12.9	19.4	12.9
Split 3	Training		100	47.8	33.8	36.9	50.0
	Invisible training		100	32.8	29.0	28.6	47.6
	Calibration		100	6.5	19.4	20.0	6.5
	Validation		100	12.9	25.0	31.3	25.0
Split 4	Training			100	32.3	22.6	39.3
	Invisible training			100	30.8	39.4	48.5
	Calibration			100	37.5	25.8	18.8
	Validation			100	25.8	32.3	25.8
Split 5	Training				100	30.0	20.3
	Invisible training				100	38.8	23.9
	Calibration				100	0.0	12.5
	Validation				100	31.3	25.0
Split 6	Training					100	33.9
	Invisible training					100	47.1
	Calibration					100	19.4
	Validation					100	37.5

*The percentage of identity for i-th and j-th splits is calculated as the following:

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

$N_{i,j}$ 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.

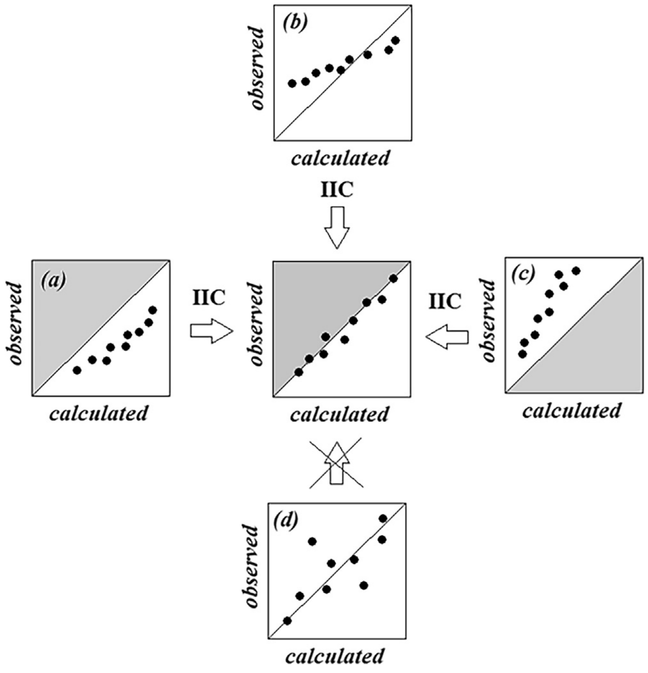

Fig. 1. Possible defects of a QSPR/QSAR model in aspect of their applying to predict values of endpoint.

$$\Delta_k = observed_k - calculated_k \tag{1}$$

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

$$^{-}MAE_{calibration} = \frac{1}{^{-}N}\sum_{k=1}^{^{-}N}|\Delta_k|\ \Delta_k < 0,\ ^{-}\text{N is the number of}\Delta_k < 0 \tag{2}$$

$$^{+}MAE_{calibration} = \frac{1}{^{+}N}\sum_{k=1}^{^{+}N}|\Delta_k|\ \ \Delta_k \geq 0,\ ^{+}\text{N is the number of}\Delta_k \geq 0 \tag{3}$$

The index of ideality of correlation (*IIC*) is calculated with the following formula:

$$IIC = r_{calibration} \times \frac{\min(^{-}MAE_{calibration},\ ^{+}MAE_{calibration})}{\max(^{-}MAE_{calibration},\ ^{+}MAE_{calibration})} \tag{4}$$

The $\mathrm{r}_{\mathrm{calibration}}$ is the correlation coefficient value between experimental and calculated values of an endpoint for the calibration set.

The IIC can be an alternative of traditional correlation coefficient. One can see from Fig. 1, there is the probability of improving poor models expressed as the cases 'a', 'b', 'c' if to use IIC instead of the traditional correlation coefficient. However, in the case of 'd', the IIC cannot improve the model (in fact, the case 'd' probably is characterized by equivalent values of the $^{-}MAE_{calibration}$ and $^{+}MAE_{calibration}$).

2.3. Apply IIC to build up predictive model

The balance of correlations is a technique described in the literature [13,14,19–24]. The essence of the approach is building up of a model via the Monte Carlo optimization of the following target function (*TF*)

$$TF = R_{training} + R_{invisible-training} - |R_{training} - R_{invisible-training}| \times Const \tag{5}$$

The $R_{training}$ and $R_{invisible-training}$ are correlation coefficients between observed and calculated values of an endpoint for the training and invisible training sets, respectively. The *Const* is an empirical constant which usually fixed equal 0.01 [25].

In this study, modified target function (TF_m) for the balance of correlation has been used

$$TF_m = TF + IIC \tag{6}$$

The optimal descriptor of the correlation weights (*DCW*) for molecular features extracted from SMILES is calculated as the following

$$DCW\,(T^*,\,N^*) = CW\,(HARD) + \sum CW\,(F_k) \tag{7}$$

The *T* and *N* are parameters of the Monte Carlo optimization. The *T* is threshold to classify molecular features into two classes (i) rare; and (ii) not rare. The *N* is the number of epochs of the optimization [13,14,19–24]. The *T** and *N** are values of these parameters which give the best statistical characteristics for the calibration set. The F_k is a molecular feature expressed in SMILES by one (e.g. 'C', 'N', 'F', etc.) or two symbols ('Cl', 'Br', etc.). The HARD is global molecular feature (physicochemical situation) extracted from SMILES. Table 2 contains example of building up the HARD, which is a superposition of described in the literature BOND, NOSP, and HALO [26]. Lines of twelve symbols represent these SMILES attributes. The '0′ indicates that a molecular feature (e.g. oxygen, double bond, chlorine, etc.) is absent; the '1′ indicates that a molecular feature is present in a molecule (Table 2).

The general scheme of building up a model with the balance of correlations contains two phases: (i) definition of *T** and *N**; and (ii) building up and validation of model based on the *DCW(T*,N*)*. Fig. 2 contains the graphical representation of the scheme of building up a model by means of the balance of correlations.

Thus, the CORAL model is one-variable correlation [25]: