Science of the Total Environment xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Science of the Total Environment

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

The index of ideality of correlation: A criterion of predictability of QSAR models for skin permeability?☆

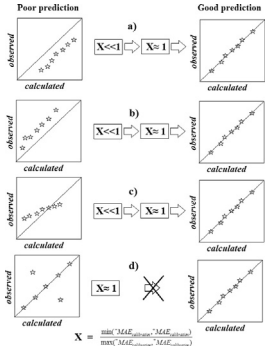
Alla P. Toropova *, Andrey A. Toropov

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

HIGHLIGHTS

- Index of ideality of correlation (IIC) can sbe a criterion of predictive potential of a model.
- The IIC is sensitive to correlation coefficient and mean absolute error.
- Utilization of IIC improves predictive potential of models for human skin permeability.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 25 November 2016
Received in revised form 24 January 2017
Accepted 29 January 2017
Available online xxxx

Keywords:
QSAR
Predictive potential
Index of ideality of correlation
Monte Carlo method
Human skin permeability
CORAL software

ABSTRACT

New criterion of the predictive potential of quantitative structure–property/activity relationships (QSPRs/QSARs) is suggested. This criterion is calculated with utilization 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 utilization of this criterion improves the predictive potential of QSAR models of dermal permeability coefficient, logKp (cm/h).

1. Introduction

The idealization is the widely used approach to describe and studying complex phenomena in the field of natural sciences, for instance, ideal gas (Feinberg, 1977), ideal solution (Yalkowsky, 1993), ideal crystals (Hill and Milstein, 1977), ideal symmetry (Toropov and Toropova, 1997).

In fact, quantitative structure–property/activity relationships (QSPR/QSAR) are a part of modern natural sciences at the

☆ Authors have equivalent contributions into the development of suggested models, preparation of the text, and design of tables and figures.

* 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.scitotenv.2017.01.198
0048-9697/

intersection of physics, chemistry, biology, and medicine (Achary, 2014a, 2014b; Toropova et al., 2014; Ghaedi, 2015; Veselinovic et al., 2015; Toropova et al., 2016). The aim of the QSPR/QSAR is the estimation of values of different endpoints for substances, which are not studied experimentally.

The Wiener's pioneer works, which are the historical basis of the modern QSPR/QSAR modeling, were aimed to establish correlations between descriptors, calculated from molecular structure, with various physicochemical endpoints (Wiener, 1947, 1948).

However, further QSPR/QSAR studies have shown that real predictive potential and correlation are not synonyms (Golbraikh and Tropsha, 2002; Roy, 2007; Chirico and Gramatica, 2011), because the cluster of dots-images is not placed symmetrically 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 with taking into account not only correlation coefficient but also the arrangement of the cluster of dots-images relatively of the diagonal in coordinates "observed calculated" values of endpoint.

The estimation of the influence of various chemicals to skin is important task of ecology and medicine (Ahmed et al., 2016; Dimitrov et al., 2016; Jung et al., 2016; Strickland et al., 2016). One of the important indicators to estimate the influence is skin permeability (Cadavona et al., 2016).

The structure of the skin presents some alterations according to the region of the body as physiological adaptations. The skin is structurally divided into three layers: the outside layer or epidermis, the dermis and the inside layer. The skin is an organ composed of many different cell types and structures with different functions. The main role of the skin is to act as a protective barrier between the inside and outside of the body. The skin also ensures physiological functions such as hydro-regulation, thermo-regulation, and sensory information. Therefore, its integrity and functionality are very important for the health of an individual. Exposure of the skin to chemicals or other materials may result in a variety of pathological effects ranging from surface effects to deeper topical effects or even systemic effects if penetration through the skin occurs (Chilcott and Price, 2008).

The QSAR models are a tool to assess the skin permeability of drugs and chemicals based on molecular structure. As with other experimental methods, QSAR models present limitations and variation between models. In fact, the majority of models make use of the physicochemical characteristics of the substances in order to estimate their skin permeability (Lee et al., 2010; Neely et al., 2009).

The aim of this study is estimation of the QSAR models (based solely on the molecular structure represented by SMILES) for skin permeability built up by means of the CORAL software (http://www.insilico.eu/coral) with using of the above-mentioned novel criterion of predictability (IIC).

2. Method

2.1. Data

Data on human skin permeability for 186 compounds expressed in logKp (cm/h) were taken in the literature (Alves et al., 2015). One compound has been rejected as a stable outlier (1-Dodecyl glycidyl ether: the only molecule which contains three-member ring which includes oxygen). The remained compounds (n = 185) are distributed into the training (≈35%), the invisible training (≈35%), the calibration (≈15%), and the validation (≈15%) sets. In addition these distributions are considerably different (non-identical) and the ranges of endpoint for the above-mentioned four sets should be as similar as possible (Table 1).

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

	Set	Split2	Split3	Split4	Split5	Split6	Split7
Split 1	Training	29.7	32.0	27.7	18.0	24.8	38.1
	Invisible training	28.6	25.7	29.1	32.1	41.6	33.7
	Calibration	38.6	16.3	31.3	26.8	17.1	26.8
	Validation	32.2	7.2	24.1	22.0	17.1	36.6
Split 2	Training	100	30.9	32.7	49.5	21.6	27.5
	Invisible training	100	26.9	35.8	42.2	28.8	36.5
	Calibration	100	23.0	28.6	16.9	24.1	26.5
	Validation	100	31.7	19.5	17.3	29.6	27.2
Split 3	Training		100	39.2	25.0	27.7	23.8
	Invisible training		100	25.5	38.1	20.0	32.0
	Calibration		100	23.0	16.3	18.6	32.6
	Validation		100	23.8	26.5	26.5	28.9
Split 4	Training			100	18.6	21.6	27.5
	Invisible training			100	33.6	31.4	25.5
	Calibration			100	28.9	16.9	21.7
	Validation			100	38.6	41.0	21.7
Split 5	Training				100	21.8	27.7
	Invisible training				100	32.4	32.4
	Calibration				100	29.3	29.3
	Validation				100	24.4	9.8
Split 6	Training					100	32.1
	Invisible training					100	30.0
	Calibration					100	14.6
	Validation					100	24.4

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.

Gray indicates diagonal and confirms that each split is identic to itself.

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:

$$\Delta_k = \text{observed}_k - \text{calculated}_k \quad (1)$$

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):

$$^{-}\text{MAE}_{\text{calibration}} = \frac{1}{^{-}\text{N}} \sum_{k=1}^{^{-}\text{N}} |\Delta_k| \quad \Delta_k<0, {}^{-}\text{N is the number of } \Delta_k<0 \quad (2)$$

$$^{+}\text{MAE}_{\text{calibration}} = \frac{1}{^{+}\text{N}} \sum_{k=1}^{^{+}\text{N}} |\Delta_k| \quad \Delta_k\geq 0, {}^{+}\text{N is the number of } \Delta_k\geq 0 \quad (3)$$

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

$$\text{IIC} = r_{\text{calibration}} \times \frac{\min(^{-}\text{MAE}_{\text{calibration}}, {}^{+}\text{MAE}_{\text{calibration}})}{\max(^{-}\text{MAE}_{\text{calibration}}, {}^{+}\text{MAE}_{\text{calibration}})} \quad (4)$$

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.

Please cite this article as: Toropova, A.P., Toropov, A.A., The index of ideality of correlation: A criterion of predictability of QSAR models for skin permeability?, Sci Total Environ (2017), http://dx.doi.org/10.1016/j.scitotenv.2017.01.198