European Journal of Medicinal Chemistry 44 (2009) 2544–2551

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

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

Original article

QSPR modeling bioconcentration factor (BCF) by balance of correlations

A.A. Toropov [a,b,*], A.P. Toropova [a,b], E. Benfenati [b]

[a] Institute of Geology and Geophysics, Khodzhibaev Street 49, 100041 Tashkent, Uzbekistan
[b] Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy

ARTICLE INFO

Article history:
Received 14 July 2008
Received in revised form
18 January 2009
Accepted 22 January 2009
Available online 31 January 2009

Keywords:
QSPR
SMILES
Bioconcentration factor
Optimal descriptor
Balance of correlations

ABSTRACT

In many cases, quantitative structure–property/activity relationships (QSPRs/QSARs) are built according to the following scheme: (1) a split of the chemicals into a training and test sets; (2) selection of a model satisfactory for the training set; (3) validation of the model with the test set. Balance of correlations is an approach we used with the following scheme: (1) a split into a subtraining set, a calibration set, and a test set; (2) selection of a model that is satisfactory for both the subtraining and calibration sets; (3) validation of the model with the external test set. Comparison of these schemes for optimal descriptors calculated with simplified molecular input line entry system (SMILES) has shown that the use of the correlation balance gives better prediction of bioconcentration factor for the test set. The calculations were carried out for three random splits into a subtraining, a calibration, and a test set.

1. Introduction

The bioconcentration factor (BCF) is useful to characterize the environmental behaviour of a chemical substance and in particular if a chemical is likely to accumulate. BCF defines the ratio between the concentration in the organism and the medium. Typically, BCF is measured in fish, but other organisms can be used. Different thresholds are defined for BCF. For instance, a chemical may be considered potentially bioaccumulative if the BCF is greater than 500 [1]. However, the EU legislation REACH defines a threshold of 2000 for a bioaccumulative compound, and of 5000 for a very bioaccumulative [2] compound. The same legislation solicits to assess the possibility of using alternative methods to evaluate chemical properties, in order to reduce the animal use. Quantitative structure–property/activity relationships (QSPRs/QSARs) are also mentioned in this legislation, and QSAR methods for BCF have been developed and used for regulatory purposes in the USA [3]. To validate the developed model is important. Indeed the model's utility depends on the possibility to quantify the model's capability to predict unknown chemicals with a clear degree of certainty [4]. The absence of this possibility leads to serious criticism of QSPR/QSAR results [5].

The genesis of the QSAR and components such as the descriptors used for the model has both heuristic and technical significance. For instance, descriptors such as LOMO and HUMO energies [6] are available for a limited number of laboratories. On the other hand, the number of internet databases providing simplified molecular input line entry system (SMILES) [7–10] is gradually increasing. Thus, the search for SMILES-based methods for QSPR/QSAR analyses becomes useful for both theoretical and practical aspects.

A good model for the training set can be a poor one for an external test set. A method of correlation balance described in Ref. [11] was effective to avoid this situation.

The aim of the present study is an estimation of predictive ability of the models of **log BCF** which have been obtained by means of the SMILES-based optimal descriptors calculated by the method of correlation balance [11]. The significant number of articles which is dedicated to the bioconcentration factor indicates importance of this endpoint [12–26].

2. Method

2.1. Data

In this study the numerical data on the bioconcentration factor (**log BCF**) was taken from Ref. [27].

An usual practice for the QSAR is scheme: (1) a split of an available experimental data into two sets: a training set and a test set; (2) construction of a model using the training set; and (3) an evaluation of the predictive potential of the model, generated with the training set, for the test set.

[*] Corresponding author. Institute of Geology and Geophysics, Khodzhibaev Street 49, 100041 Tashkent, Uzbekistan.
E-mail address: aatoropov@yahoo.com (A.A. Toropov).

0223-5234/$ – see front matter © 2009 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2009.01.023

Table 1
List of the global SMILES attributes.

SA_k	Definition	Reason
!-01________ !-02________ … !000________ !001________ … !010________	Difference between the numbers of ‘c’ (lowercase letter) and the number of ‘C’ (capital letter)	Carbons in sp^2 state generate a rigid molecular fragment. Carbons in sp^3 state generate a flexible molecular fragment. The chlorine atom also is a provider of the ‘C’ (capital letter), but chlorine also is a flexible (not rigid) fragment. Thus this difference is an indicator of behaviour of a molecule in mechanical aspect (presence/absence of rigid and flexible fragments)
(000________ (001________ … (025________	Number of branches in the molecular skeleton. (000_____ is indicator of absence of the branching	Branching has influence on many physico-chemical properties (normal boiling and melting points, viscosity, density, etc.)
1000________ 1002________ … 1006________ 2000________ 2002________ … 5002________	Numbers of cycles encoded in the SMILES with ‘1’, ‘2’, … ‘5’ The 1000______, 2000______, …5000____ are indicators of the absence of the cycles	Cycles have influence on many physico-chemical and biological parameters
=000________ =001________ … =006________	Number of the double covalent bonds in molecule. = 000____ is an indicator of absence of double bonds; = 001 is an indicator of presence of one double bond, etc.	Double bonds have a effect upon many biochemical phenomena
C000________ C001________ … C027________	Number of carbon atoms in the sp^3 electronic state	Each chemical element theoretically can have an effect to any endpoint
F000________ … F027________	Number of fluorine atoms in molecule	Each chemical element theoretically can have an effect to any endpoint
Br00________ … Br10________	Number of bromine atoms in molecule	Each chemical element theoretically can have an effect to any endpoint
Cl00________ … Cl08________	Number of chlorine atoms in molecule	Each chemical element theoretically can have an effect to any endpoint
N000________ … N006________	Number of nitrogen atoms in molecule	Each chemical element theoretically can have an effect to any endpoint
O000________ … O010________	Number of oxygen atoms in molecule	Each chemical element theoretically can have an effect to any endpoint
P000________ P001________	Number of phosphorus atoms in molecule	Each chemical element theoretically can have an effect to any endpoint
S000________ … S003________	Number of sulphur atoms in molecule	Each chemical element theoretically can have an effect to any endpoint
c000________ … c022________	Number of carbon atoms in the sp^2 electronic state	Each chemical element theoretically can have an effect to any endpoint

As an alternative, the training set can be split into two sets: a subtraining set and a calibration set and the model obtained with the subtraining set can be preliminarily evaluated with substances of the calibration set. In fact, the classical training phase gives a maximum of the correlation coefficient between the actual and predicted values for an endpoint of interest. Using the subtraining and calibration sets, one can calculate the combinatorial measure of correlation between actual and predicted values, as the following:

$$B = R_s^2 + R_c^2 - \mathrm{abs}\left(R_s^2 - R_c^2\right) \tag{1}$$

where R_s and R_c are correlation coefficients between the actual and predicted values of an endpoint for the subtraining set and calibration set, respectively. It is likely that a model, obtained by the scheme [subtraining → calibration → prediction] (i.e., by the correlation balance), is more robust in comparison with the model, obtained by the ‘classic’ scheme [training → test], without the preliminary calibration [11].

One-variable models examined in this study are based on optimal descriptors, calculated as the following:

$$\mathrm{DCW(LimN)} = \prod \mathrm{CW}(\mathrm{SA_k}) \tag{2}$$

where $\mathbf{SA_k}$ is an attribute of the SMILES notation; the $\mathbf{CW}(\mathbf{SA_k})$ is the correlation weight for the $\mathbf{SA_k}$; the **LimN** described in Ref. [11] is a parameter that defines two categories: “rare SMILES attribute” and “non rare SMILES attribute”. In other words, the **LimN** is the minimal number of $\mathbf{SA_k}$ in the training (or subtraining for case of the correlation balance) set. For instance, if **LimS** = 4 and only three (or less) $\mathbf{SA_k}$ exist in the training (or subtraining) set then this attribute should be classified as rare. The correlation weights of the rare SMILES attributes are equal to 1. The correlation weights of the non-rare SMILES attributes are calculated by optimization of the Monte Carlo method: the target function can be the correlation coefficient between **DCW**(**limN**) and **log BCF** for the training set or the criterion that is calculated with Eq. (1) for the subtraining set and calibration set.

The use of different **LimNs** is accompanied by different statistical qualities of the models. If **LimN** tends to zero, the correlation coefficient between **DCW**(**limN**) and **log BCF** asymptotically tends to unit for the training set. An increase of the **LimN** leads to a decrease of the correlation coefficient for the training set, owing to the decrease of the number of optimized parameters. However, the increase of the **limN** can lead to an increase of the correlation coefficient for the test set [11,28].

The local and global SMILES attributes are used in this study. The local attributes are constructed with elements of the SMILES. The element can be one character of the SMILES or two characters, which cannot be separated, without the loss of the physical sense: for instance, ‘Cl’, ‘Br’, ‘Si’, etc. Combinations of one, two and three SMILES elements are used as local SMILES attributes. Combinations of two (AB) and three (ABC) SMILES elements are organized according to ASCII codes. In other words these local attributes have only one version in the list of SMILES attributes (e.g., AB, not both AB and BA; similarly ABC, not ABC and CBA: it should be noted that the positions of A and C can be changed, B in any case has middle position). The global SMILES attributes are calculated with the SMILES. These are the number of a chemical element (C, O, N, etc.) or other molecular features, such as branching in molecular skeleton, cycles, etc. Table 1 contains the list of global SMILES attributes.

The split into training and test sets has an influence on the results of the QSPR/QSAR modeling. Hence, the robustness of the model should be estimated for several splits.

Three splits into subtraining, calibration, and test set for modeling by the scheme [subtraining → calibration → test] are examined. Supplementary material section contains CAS numbers of substances randomly extracted in the external test sets into splits A, B, and C. For modeling by scheme [training → test], the association of the subtraining and calibration sets was used as the united training set.

Optimization of the correlation weights has been carried out by algorithm based on the Monte Carlo method [11]. For each intermediate list of the correlation weights, the **DCW**(**limN**) values for all compounds have been calculated. One can, using these data, calculate the R_s and R_c for the [subtraining → calibration → test] scheme or the correlation coefficient for the united subtraining set