Chemosphere 73 (2008) 1701–1707

Contents lists available at ScienceDirect

Chemosphere

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

# A new hybrid system of QSAR models for predicting bioconcentration factors (BCF)

Chunyan Zhao [a,b], Elena Boriani [b], Antonio Chana [b], Alessandra Roncaglioni [b], Emilio Benfenati [b,*]

[a] Department of Chemistry, Lanzhou University, Lanzhou 730000, China
[b] Istituto di Ricerche Farmacologiche "Mario Negri", Via La Masa 19, 20156 Milano, Italy

ARTICLE INFO

*Article history:*
Received 21 April 2008
Received in revised form 10 September 2008
Accepted 12 September 2008
Available online 26 October 2008

*Keywords:*
Bioconcentration factors
Hybrid model
QSAR validation
REACH

ABSTRACT

The aim was to develop a reliable and practical quantitative structure–activity relationship (QSAR) model validated by strict conditions for predicting bioconcentration factors (BCF). We built up several QSAR models starting from a large data set of 473 heterogeneous chemicals, based on multiple linear regression (MLR), radial basis function neural network (RBFNN) and support vector machine (SVM) methods. To improve the results, we also applied a hybrid model, which gave better prediction than single models. All models were statistically analysed using strict criteria, including an external test set. The outliers were also examined to understand better in which cases large errors were to be expected and to improve the predictive models. The models offer more robust tools for regulatory purposes, on the basis of the statistical results and the quality check on the input data.



## 1. Introduction

A large number of chemicals that resist degradation in the environment accumulate in body tissues and are intrinsically toxic to organisms in the environment or to humans. The possible effects of long-term and cumulative exposure to such chemicals are not always addressed adequately in risk assessment methods evaluating acute toxicity and short-term exposure (Pavan et al., 2006). Thus, bioconcentration is of great concern when defining toxic effects due to chronic exposure.

Bioconcentration usually refers to a situation, under laboratory conditions, where the chemical is absorbed from water only through the respiratory surface or the skin. Chemical bioconcentration is usually expressed as a bioconcentration factor (BCF), which can be defined as the ratio of the concentration of a chemical present in an aquatic organism to that in the environment (Lu et al., 2000). Among aquatic species, fish typically serve as a target for BCF assessments in view of their importance as food for many species, including humans, and the availability of standardized testing protocols (Barron, 1990). However, experimental determination is expensive and time-consuming, so estimation methods are needed to supply the missing data.

In Europe the recent legislation on industrial chemicals REACH (Registration, Evaluation, Authorisation and restriction of Chemicals) requires a huge amount of data, considering the tens of thousands of compounds to be evaluated in the EU (EC, 1907/2006, 2006). BCF is one of the endpoints which will require more chemical tests, hence costs. Quantitative structure–activity relationship (QSAR) models have been identified in scientific and policy communities as a major tool for obtaining this information and REACH foresees its extended use (Benfenati, 2007). This approach offers the advantage that it only requires knowledge of the chemical structure.

Many researchers have applied the QSAR method to investigate the correlations between structural descriptors and BCF and a review summarises them (Tao et al., 2000). One approach was to estimate a chemical's BCF based on its relationship with other physicochemical parameters such as the octanol/water partitioning coefficient ($K_{ow}$) (Devillers et al., 1996; Fisk et al., 1998), or the soil absorption coefficient ($K_{oc}$) (Sabljic et al., 1995). Log$K_{ow}$, also called log$P$, is widely held to be the most common and important descriptor to establish predictive models for BCF. These include linear regression (Veith et al., 1979; Mackay, 1982; Veith and Kosian, 1983), non-linear (Sabljic and Protic, 1982; Dimitrov et al., 2002), bilinear (Nendza, 1998) and polynomial (Connell and Hawker, 1988) models. However, these models have certain drawbacks, particularly because very large molecules with high log$K_{ow}$ may diffuse only slowly through membranes, resulting in considerable discrepancies in correlations for chemicals with log$K_{ow}$ > 7 (Mackay and Fraser, 2000; Papa et al., 2007). Some QSAR models use theoretical molecular descriptors including molecular weight (Govers et al., 1984), molecular connectivity indices (Sabljic and Protic, 1982; Sabljic, 1987; Lu et al., 2000), topological, geometrical (Tao et al., 2000), quantum-chemical descriptors (Wei et al., 2001) or combinations of different

* Corresponding author. Tel.: +39 02 39014420; fax: +39 02 39014735.
*E-mail address:* benfenati@marionegri.it (E. Benfenati).

0045-6535/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.chemosphere.2008.09.033

molecular descriptors (Gramatica and Papa, 2005). Many studies have obtained good results, but some models still lack robust validation, and the quality control of the data used for modelling is not always defined.

In the present study we used experimental data obtained only according to official guidelines (Dimitrov et al., 2005). A widely used model is EPI Suite (Environmental Protection Agency, http://www.epa.gov/oppt/exposure/pubs/episuite.htm), and we compare the results with this model and our models.

In many cases QSAR models implement a leave-one-out (or leave-some-out) cross-validation procedure and a high $q^2$ (for instance > 0.5) has been considered an indicator that the model is highly predictive (Golbraikh and Tropsha, 2002):

$$q^2 = 1 - \frac{\sum_i (y_i - \hat{y}_{i/i})}{\sum_i (y_i - \bar{y})} \quad (1)$$

where $y_i$, $\hat{y}_{i/i}$ and $\bar{y}$ are, respectively, the observed, estimated by cross-validation and mean values of activities. A high $q^2$ is necessary but not sufficient for a model to have strong predictive power (Golbraikh and Tropsha, 2002). Thus, beside the widely accepted $q^2$ criteria, the QSAR model needs stricter conditions to ensure good predictive ability for untested chemicals.

This study employed a set of statistical characteristics, already adopted in the DEMETRA project (Development of Environmental Modules for Evaluation of Toxicity of Pesticide Residues in Agriculture), to decide whether a QSAR model has acceptable predictive power (Benfenati et al., 2007). We now present new, reliable QSAR models, validated in strict conditions, for predicting BCF using different techniques. In addition, we applied a new hybrid model, which gave better prediction than single models. We also examined the structure of outliers, in order to identify fragments possibly related to larger errors. The discussion considers how fragments can be used to identify outliers.

## 2. Material and methods

### 2.1. Data set

The data set of 511 compounds and the measured log BCF values were from Dimitrov et al. (2005). The biological data are of high quality. Values were obtained only according to official guidelines, which makes the data suitable for regulatory purposes, such as REACH. For a quality check of the chemical information, using the molecule names and/or CAS numbers from the literature, we checked the two-dimensional (2D) chemical structures at five online databases: ChemFinder (http://chemfinder.cambridgesoft.com), ChemIDPlus (http://chem.sis.nlm.nih.gov/chemidplus/), Safe Nite (http://www.safe.nite.go.jp/english/kizon/KIZON_start_hazkizon.html), Biodegradability Database and Estimation (http://qsar.cerij.or.jp/cgi-bin/QSAR/index) and PubChem Compound (http://www.ncbi.nlm.nih.gov/sites/). Some ambiguities or errors were found. Some compounds were omitted according to the following criteria: (a) too little information to find the structure; (b) mixtures; (c) diastereo-isomers; (d) metal complexes; (e) some compounds were repeated in the original paper. We used the neutral form of salts. With diastereo-isomers we kept only one compound, using the average BCF value. The full list of chemicals omitted, and the specific reasons, is available in the Supplementary material (Table S1).

The final database of 473 compounds was created with ISIS BASE 2.5 SP2. The data set covers a wide range of log BCF and calculated $\log K_{ow}$ (log BCF from −1.00 to 4.85; $\log K_{ow}$ from −4.3 to 12.7), with molecular weights from 68 to 943. The 473 compounds were randomly split into a training ($n$ = 378) and a test set ($n$ = 95) using Statistica 6.0 random number generator (http://www.statsoft.com). Chemicals with their log BCF values are listed in the Supplementary material (Table S2).

### 2.2. Generation and selection of descriptors

We used 2D molecular descriptors, calculated with Dragon version 5.4 (759 descriptors) (http://www.talete.mi.it), MDL descriptors (249 descriptors), ACD labs (version 9.08) (13 descriptors), and KOWWIN (version 1.67) (1 descriptor) (http://www.syrres.com/eSc/est_kowdemo.htm), mainly including (a) constitutional descriptors; (b) functional groups, atom centered fragments; (c) topological, BCUTs (Burden–CAS–University of Texas eigenvalues), walk and path counts, autocorrelations, connectivity indices, information indices, topological charge indices, and eigenvalue-based indices. Thus, 1022 descriptors were obtained including different log $P$ and log $D$ values calculated with these programs. Constant or near-constant descriptors were omitted.

Heuristic (HM) (Zhao et al., 2005) and genetic algorithm (GA) methods were then used to select optimal descriptors. The software CODESSA version 2.21 was used, to give a complete search for the best multilinear correlations in the ordinary least squares regression (OLS) method. MobyDigs version 1.0 (http://www.talete.mi.it) was used for genetic algorithm-variable subset selection (GA-VSS).

### 2.3. Building predictive models

Multiple linear regression (MLR) was used to develop the linear model of the property of interest, with CODESSA software. Radial basis function neural network (RBFNN) (Wan and Harrington, 1999) was used with a Matlab function to build our models. Information on this function is available in the Supplementary material.

We used R code for support vector machines (SVM) (Burges, 1998) models (http://www.R-project.org). This learning system uses a hypothetical space of linear functions in a high-dimensional feature space, trained with a learning algorithm from the optimisation theory. SVM algorithms yield prediction functions that are expanded on a subset of training vectors, or support vectors.

Finally, we used a hybrid model approach based on the idea of more representations of the problem, more paradigms of knowledge representation, and more algorithms to find a solution. As in other cases (Lo Piparo et al., 2006; Amaury et al., 2007b; Porcelli et al., 2008), we used the outputs of the individual models as inputs of the hybrid model. Thus, the hybrid model is similar to the QSAR model, but its inputs are the outputs (predicted values) of the QSAR models it combines. We used in-house software made as a PC-Windows Excel macro to build combined models. The interval of the output of each individual model was divided into three areas, where the predicted output, maximum, minimum, or mean of the selected models, are used for the final model. In practice we used a non-continuous function that can be expressed as combinations of simple linear equations such as:

$$\log \mathrm{BCF} = k_n[\mathrm{Min}, \mathrm{Mean}, \mathrm{Max}(\text{value given by models to combine})] + a_n \quad (2)$$

where $n$ is the number of areas chosen to build the hybrid models, three in our case. In the present case the final expression is as follows:

If mean (value given by models to combine) > 2.410

$$\log \mathrm{BCF} = 1.052 * [\mathrm{Min}(\text{value given by models to combine})] - 0.065 \quad (3)$$