Ecotoxicology and Environmental Safety 138 (2017) 92–97

Contents lists available at ScienceDirect

Ecotoxicology and Environmental Safety

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

Development and evaluation of predictive model for bovine serum albumin-water partition coefficients of neutral organic chemicals

Guangcai Ma, Quan Yuan, Haiying Yu*, Hongjun Lin, Jianrong Chen, Huachang Hong

College of Geography and Environmental Sciences, Zhejiang Normal University, Yingbin Avenue 688, 321004, Jinhua, PR China

ARTICLE INFO

Keywords:
Bovine serum albumin – water partition coefficients
Quantitative structure – property relationship
QSAR
PM6
Quantum chemical descriptors

ABSTRACT

The binding of organic chemicals to serum albumin can significantly reduce their unbound concentration in blood and affect their biological reactions. In this study, we developed a new QSAR model for bovine serum albumin (BSA) – water partition coefficients ($K_{BSA/W}$) of neutral organic chemicals with large structural variance, log$K_{BSA/W}$ values covering 3.5 orders of magnitude (1.19–4.76). All chemical geometries were optimized by semi-empirical PM6 algorithm. Several quantum chemical parameters that reflect various intermolecular interactions as well as hydrophobicity were selected to develop QSAR model. The result indicates the regression model derived from logK_{ow}, the most positive net atomic charges on an atom, Connolly solvent excluded volume, polarizability, and Abraham acidity could explain the partitioning mechanism of organic chemicals between BSA and water. The simulated external validation and cross validation verifies the developed model has good statistical robustness and predictive ability, thus can be used to estimate the log$K_{BSA/W}$ values for chemicals in application domain, accordingly to provide basic data for the toxicity assessment of the chemicals.

1. Introduction

When exogenous organic chemicals enter into the organisms, the partitioning in tissues and organs significantly affects their bioaccumulation, toxicodynamics and toxicity. It was generally believed that the liposome component play an important role in the adsorption and absorption of organic pollutants in living organism (Arnot and Gobas, 2004; Czub and McLachlan, 2004). However, more and more studies have identified the contribution of protein to the enrichment of organic chemicals must not be ignored (DeBruyn and Gobas, 2007; Endo and Goss, 2011; Endo et al., 2012; Kelly et al., 2007). The chemicals can bind to serum constituents, especially serum albumin, which is the most abundant carrier protein in the blood serum, accordingly reduce the unbound (or freely available) concentration in blood and affect the cell uptake, metabolic reactions and other biological processes (Armitage et al., 2014; Endo et al., 2013; Groothuis et al., 2015; Madureira et al., 2014). Furthermore, the serum albumin is usually taken as a model protein to investigate the interactions of organic chemicals with protein because of the abundance of binding constant and the necessity of drug development (Bischel et al., 2010; Ghuman et al., 2005; Vandenbelt et al., 1972).

The bovine serum albumin (BSA) is widely used to study the binding affinity of organic chemicals to serum albumin (Bischel et al., 2010; Zhang et al., 2016), and the equilibrium partitioning is characterized by BSA-water partition coefficients ($K_{BSA/W}$), which is defined as:

$$K_{BSA/W} = C_{BSA}/C_{water} \quad (1)$$

where C_{BSA} and C_{water} are the equilibrium concentration of chemicals at certain temperature in BSA and water, respectively. The $K_{BSA/W}$ values of some volatile organic compounds such as alkylbenzens, polycyclic aromatic hydrocarbons (PAHs), and aliphatic alcohols has been determined by solid phase microextraction (Cross et al., 2003; Kramer et al., 2007; van Eijkeren et al., 2004; Yuan et al., 1999). Endo and Goss reported log$K_{BSA/W}$ values for 83 neutral organic chemicals, so far the largest structurally diverse dataset that were measured by the same research group (Endo and Goss, 2011).

As is well known, the experimental determination of log$K_{BSA/W}$ is subject to time cost and high-purity of standards, and there is an urgent need to understand the molecular structure properties that govern the partition behavior of chemicals between BSA and water. All of this promotes the theoretical study of log$K_{BSA/W}$. Hansch et al. found that the log$K_{BSA/W}$ of some small apolar chemicals strongly correlated with their octanol-water partition coefficients (logK_{ow}) (Jahnke et al., 2008; Mayer et al., 2009), similarly to the distribution of chemicals in lipid phase (Helmer et al., 1968; Vandenbelt et al., 1972), and pointed out that non-specific intermolecular forces dominated the binding of neutral chemicals to BSA. However, Endo et al. revealed that the linear correlation between log$K_{BSA/W}$ and logK_{ow} became less significant when

* Corresponding author.
E-mail address: yhy@zjnu.cn (H. Yu).

http://dx.doi.org/10.1016/j.ecoenv.2016.12.022
Received 3 July 2016; Received in revised form 13 November 2016; Accepted 16 December 2016

the molecular structure diversity enlarged, and a break point arose at $\log K_{ow}$ = 4–5, the slope thereafter became gentler (Endo and Goss, 2011). At the same time, they declared that the polyparameter linear free energy relationships (pp-LFER) (Abraham et al., 2001; Goss, 2005), which are capable of describing accurately the non-specific intermolecular interactions cannot satisfactorily fit $\log K_{BSA/W}$, so they also suggested that new model that fully capture the interactions between the chemicals and BSA is needed to improve the regression accuracy.

To date, quantitative structure-activity relationships (QSARs) that correlate the molecular structure descriptors and chemical properties have been widely used in predicting the toxicity and metabolism of organic chemicals (Cherkasov et al., 2014). The QSAR models also can help us to understand the toxicity mechanism through the physicochemical information carried by the molecular descriptors. In 2007, OECD promulgated a guidance document for developing, validating and defining application domain of QSAR models for regulatory purpose (OECD, 2007). The present QSAR models for $\log K_{BSA/W}$ either have no satisfied statistics or do not fully comply with the OECD guidance. Accordingly, the purpose of this study is (1) to develop a new QSAR model for predicting $\log K_{BSA/W}$ values of neutral organic chemicals by quantum chemical computation according to OECD guidance; (2) to identify the physicochemical characteristics that dominate the binding affinity of chemicals to BSA.

2. Materials and methods

2.1. Experimental $\log K_{BSA/W}$ values and chemical structures

The $\log K_{BSA/W}$ values for 83 neutral organic chemicals were taken from the *reference* (Endo and Goss, 2011). All of the values were measured by headspace sampling (volatile compounds) or fiber extraction method (not volatile compounds) coupled with GC/MS at 37 °C. The chemicals in the data set include alkanes, alkenes, alkyl halides, alcohols, phenols, ethers, ketones, amines, anilines, nitro compounds, polycyclic hydrocarbons, heterocyclic compounds, benzene derivatives, and the $\log K_{BSA/W}$ values vary from 1.19 (4-ethyl-3-hexanol) to 4.76 (pyrene), covering 3.6 orders of magnitude. All data is summarized in Table 1.

2.2. Molecular descriptors

Intermolecular interactions, such as dispersion interactions, dipole-dipole interactions, dipole-induced dipole interactions, electrostatic interactions and hydrogen bonding may dominate the partitioning of organic chemicals in BSA and water, so we selected average molecular polarizability (α, a.u.) and dipole moment (μ, a.u.) to characterize van der Waals forces, and choose Abraham parameters A and B to quantify the hydrogen bond donor and acceptor strength, respectively. As BSA contains a lot of amino-acid residues (the isoelectronic point is 4.7), it may form positive or negative charges at the physiologic pH values and then polarize nearby molecules and induce favorable dipoles. Such charge-induced dipole interactions were quantitatively expressed by the most negative net atomic charges on an atom (q^-, a.c.u.) and the most positive net atomic charges on an atom (q^+, a.c.u.). Furthermore, the cavities should be formed for the partitioning of organic molecules in the solvent, so Connolly solvent excluded volume (*CSEV*, Å^3) was selected to reflect the molecular size. The hydrophobicity also plays an important role in the solvation of chemicals in organic phase, and $\log K_{ow}$ was chosen.

All of the molecules were optimized by semi-empirical PM6 methodology, which was implemented by MOPAC (2012), to calculate α, μ, q^- and q^+. *CSEV* was quantified by CS Chem3D Ultra in ChemOffice 2010 (Cambridge, UK). Abraham parameters A and B were collected from the reference (Endo and Goss, 2011). The $\log K_{ow}$ values were obtained from EPI Suite 4.11 (EPA, 2011) and all of them are experimental measurement for the purpose of model accuracy. Thus, the chemicals have no experimental $\log K_{ow}$ values were removed and finally the dataset contains 74 chemicals.

2.3. Model calibration and validation

The multilinear regression (MLR) is employed to filter the descriptors and develop predictive $\log K_{BSA/W}$ model, which is carried out by SPSS 17.0. The squared correlation coefficient r^2 was used to quantify the calibration performance:

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i^{fit} - y_i^{obs})^2}{\sum_{i=1}^{n}(y_i^{obs} - y^{mean})^2} \quad (2)$$

where y_i^{fit}, y_i^{obs} and y^{mean} denote the regression-fitted, observed, and observed mean values (in our case: $\log K_{BSA/W}$), respectively, and n is the number of data points. The predictive performance was evaluated by q^2 (Schüürmann et al., 2008):

$$q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i^{pred} - y_i^{obs})^2}{\sum_{i=1}^{n}(y_i^{obs} - y^{mean})^2} \quad (3)$$

In Eq. (3), y_i^{fit} is replaced by y_i^{pred}, the predictive $\log K_{BSA/W}$ value by the fitting model. Differently from r^2 value that varies between 0 and 1, q^2 ranges from $-\infty$ to 1 (perfect agreement) and $q^2 = 0$ represents the case that the predictive performance is as good as taking the average $\log K_{BSA/W}$ value as the predicted value for all compounds. In addition, root mean square error (rms), mean error (me), maximum negative error (mne, the largest underestimation), and maximum positive error (mpe, the largest overestimation), were also calculated to assess the statistical performance of the developed model.

The multicollinearity among the selected molecular descriptors was diagnosed by a variable inflation factor (VIF). If VIF varies from 1.0 to 5.0, the regression model is suitable without the existence of collinearity; if VIF is larger than 10.0, the correlation coefficients of variables must be rechecked.

The statistical robustness and predictive ability of regression model were evaluated by simulated external validation and cross validation. In simulated external validation, the whole dataset was randomly divided into training set that contains 70% compounds and the test set for the left 30% compounds. For the feasibility of validation, both subsets should cover the descriptor space and the $\log K_{BSA/W}$ range of the whole data set. Then a new model was re-calibrated for the training set using the descriptors selected in the original model, and then used to predict the $\log K_{BSA/W}$ values for the test set. The parameters r^2, q^2 and rms were calculated to characterize the predictive capacity.

2.4. Influential compounds and outliers

The compounds that own extremely large or small molecular descriptors can significantly affect the model regression. Such chemicals are diagnosed by Hat (h_i) value, which is calculated as:

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (4)$$

where X is a descriptor matrix in which the rows and columns represent the compounds and their descriptors, respectively, X^T is the transposition of X, $(X^T X)^{-1}$ is the inverse of matrix $X^T X$, and x_i is the descriptor row-vector of the query compound. The warning value for h_i is $h^* = 3p/n$, where n and p is the number of chemicals and descriptors, respectively (Chatterjee and Hadi, 1986). Thus the chemicals with h_i higher than h^* are identified as influential compounds, but not necessarily outliers, which are determined by the absolute standard deviation larger than 3. The outliers should be removed from the application domain to improve the predictive reliability. In this study, the influential compounds and outliers are visualized by a Williams plot of standardized prediction residuals (SR) versus leverage h_i.