Ecotoxicology and Environmental Safety 157 (2018) 128–133

Contents lists available at ScienceDirect

Ecotoxicology and Environmental Safety

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

Development of chicken and fish muscle protein – Water partition coefficients predictive models for ionogenic and neutral organic chemicals

Wang Ou[a], Huihui Liu[a,*], Junyi He[a], Xianhai Yang[a,b,**]

[a] Jiangsu Key Laboratory of Chemical Pollution Control and Resources Reuse, School of Environmental and Biological Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

[b] Nanjing Institute of Environmental Science, Ministry of Environmental Protection, Nanjing 210042, China

ARTICLE INFO

Keywords:
Muscle protein
Quantitative structure-activity relationship (QSAR)
Muscle protein-water partition coefficient (log$K_{MP/w}$)
Chicken
Fish

ABSTRACT

Muscle protein was one of critical accumulation protein for anthropogenic chemicals. However, few predictive models were constructed for muscle protein up to now. In addition, some ionizable chemicals classes e.g. sulfonates were not successfully modeled in previously models, indicating considerable work would be needed. The major objective of this study was to develop quantitative structure-activity relationship (QSAR) models for predicting the muscle protein-water partition coefficient (log$K_{MP/w}$) of chicken and fish. In the modeling, the n-octanol/water distribution coefficient (log$D$), functional groups, atom-centred fragments and chemical form adjusted descriptors were used to construct the models. The application domain of the derived models was defined by the Euclidean distance-based method and Williams plot. The modeling results indicated that the determination coefficient (R^2), leave-one out cross validation Q^2 (Q^2_{LOO}) and bootstrapping coefficient (Q^2_{BOOT}) of the QSAR models for chicken and fish were 0.882 and 0.929, 0.844 and 0.906, 0.779 and 0.792, respectively, implying the models had good goodness-of-fit and robustness. The coefficient determination (R^2_{EXT}) and external validation coefficient (Q^2_{EXT}) of the validation set for the two models were 0.874 and 0.937, 0.869 and 0.915, respectively, indicating the models had good predictive ability. The predictor variables selected to construct the log$K_{MP/w}$ models of chicken and fish included log$D$, the function groups, and the fraction of the ionized species (δ_I). Considering the molecular descriptors used here can be calculated from their molecular structures directly, the developed models could be easily used to fill the log$K_{MP/w}$ data gap for other chemicals within the applicability domain.

1. Introduction

Organisms are inevitably exposed to the environmental medium contained various anthropogenic chemicals (Schwarzenbach et al., 2006; Ding et al., 2017a). Unfortunately, these compounds may find their ways into organisms, such as through skin contact, ingestion, and inhalation (Judson et al., 2011; Zhang et al., 2014). Among the chemicals into organisms, a part of them would be metabolized and excreted; while another part would be stored in organisms. Over a long time, more and more chemicals will be remained in organisms, which are known as the bioaccumulation process. As a result, it prolongs the residual time of chemicals in the body, which means the organisms may suffer from continuing exposures to those chemicals. This may cause a variety of adverse effects on organisms (Henriksson et al., 2017; Zhuang et al., 2017). Therefore, measuring the bioaccumulation potential of a chemical in organisms is essential for characterizing the chemicals' persistence and assessing exposure risk. Generally, lipid was regarded as the most important storage phase for anthropogenic chemicals (Arnot and Gobas, 2006; Letcher et al., 2010; van der Heijden and Jonker, 2011).

It was reported that proteins such as serum albumin, muscle protein, membrane protein, and so on, are also responsible for the storage of chemicals in organism (Kelly et al., 2009; Endo et al., 2012, 2013; Henneberger et al., 2016a, 2016b). Even though serum albumin is often used as a representative protein for the bulk protein fraction of an organism, more and more studies documented that muscle protein may be also an important accumulation phase, because of its abundance in organisms' lean tissues (Endo et al., 2012). Additionally, it constitutes the main edible part by predators, which is relevant to the food-web bioaccumulation of chemicals. Unlike lipids, due to the polar nature,

* Corresponding author.
** Corresponding author at: Jiangsu Key Laboratory of Chemical Pollution Control and Resources Reuse, School of Environmental and Biological Engineering, Nanjing University of Science and Technology, Nanjing 210094, China.
E-mail addresses: hhliu@njust.edu.cn (H. Liu), xhyang@njust.edu.cn (X. Yang).

https://doi.org/10.1016/j.ecoenv.2018.03.064
Received 22 January 2018; Received in revised form 13 March 2018; Accepted 23 March 2018

various proteins are prone to accumulate the hydrophilic compounds, such as phenols, sulfonates (Chen and Guo, 2009; Henneberger et al., 2016b). Those hydrophilic compounds generally contain dissociable groups, which may ionize depending on the pH of the environment and their pK_a. It is well known that the partitioning behavior, toxicity, mechanisms of action of the dissociated form is different from that of the undissociated form, which behaved as form-dependent (Nishikawa and Kondo, 2006; Abraham and Acree, 2010; Crisan et al., 2014). For example, Yang et al., (2013, 2017) founds that the anionic forms of the ionizable compounds bind stronger than the corresponding neutral forms with human transthyretin. Among the pre-registered 0.145 million substances by REACH, more than 10% of the substances may ionize into anions and/or cations in the normal physiological pH conditions, e.g.7.4 for human plasma (Daginnus, 2010). If we were not aware of the form-dependent phenomenon, it will result in inaccurate measurement or prediction for the accumulation capability of muscle protein.

Muscle protein-water partition coefficient ($K_{MP/w}$) is a key parameter for quantifying the accumulation potential of muscle protein. However, only about 100 chemicals have experimental $K_{MP/w}$ data up to now. It is unrealistic to fill the $K_{MP/w}$ data gap for more than 140,000 commercially chemicals by experimental method alone, as it often is time-consuming, cost-intensive, equipment-dependent and error-prone. Therefore, it is important to develop rapid methods to help in filling data gap. *In silico* methods (e.g. quantitative structure-activity relationship (QSAR) model) have been demonstrated as useful and efficient in this application. For example, predictive models were successful developed for multiple distribution coefficients (Goss and Schwarzenbach, 2001; Yang et al., 2003; Wu et al., 2008; Endo and Goss, 2014; Ng et al., 2015). For $K_{MP/w}$ of muscle protein, Endo et al. (2012) and Henneberger et al. (2016b) developed polyparameter linear free energy relationship (PP-LFER) models. However, some ionizable chemicals classes e.g. sulfonates were not successfully included in their models due to the method was not appropriate for those compounds, indicating considerable work would be needed.

Therefore, the objectives of present study were: (1) to develop new predictive log$K_{MP/w}$ model that include both ionogenic and neutral organic chemicals; (2) to use the descriptors that could characterize the ionization effects and could be calculated from their molecular structures directly; (3) to develop the predictive models following the OECD QSAR models validation guideline (OECD, 2007).

2. Materials and method

2.1. Data sets

In this study, muscle proteins of chicken and fish were used as model proteins. The experimental data of $K_{MP/w}$ were obtained from the studies of Endo and co-workers (Endo et al., 2012; Henneberger et al., 2016b). The logarithm of partition coefficient (log$K_{MP/w}$) was employed as the modeling endpoint. It is defined as:

$$\log K_{MP/w} = \log\left(\frac{C_{MP}}{C_W}\right) \quad (1)$$

where C_{MP} and C_{water} are the equilibrium concentrations of a given chemical at experimental conditions in muscle protein and water, respectively.

In original studies, muscle proteins isolated from fish and chicken were used. For fish muscle protein, the data set contained 45 compounds, including 6 anionic compounds and 3 cationic compounds. For chicken muscle protein, the data set contained 97 compounds, including 48 anionic compounds and 16 cationic compounds. In the modeling, the data set of fish and chicken was divided into a training set and a validation set at random with a ratio of 3:1, respectively. The chemicals' names, CAS numbers and the observed log$K_{MP/w}$ value of each compound for chicken and fish were presented in the Table S1 and S2 of the Supporting Information, respectively.

2.2. Calculation of the molecular descriptors

Firstly, the molecular structures of model compounds were optimized by the minimize energy method in the ChemBio3D Ultra (Version 12.0). Then, the compounds' geometry was further optimized using Gaussian 16 program at the B3LYP/6–31 + G (d,p) level (Frisch et al., 2016). In addition, solvent effect of water was considered implicitly by the polarized continuum model (Li et al., 2010).

On the basis of the output file from Gaussian 16, fourteen quantum chemical descriptors including the molecular polarizability, dipole moment, molecular volume (V), the highest occupied molecular orbital energy (E_{HOMO}), the lowest unoccupied molecular orbital energy (E_{LUMO}), the most positive and negative net atomic charge on a carbon atom (qC^+ and qC^-), the most positive net atomic charge on a hydrogen atom (qH^+) and the most negative charge on electron acceptor atom (e.g. O, N, S, X) (Q_{min}), the average potentials on the molecular surface (V_s), the average deviation of surface potential (Π), the chemical potential (μ), the chemical hardness (η), the electrophilicity index (ω) were selected to characterize the hydrogen bonding, electrostatic interactions and electron donor-acceptor interactions (Li et al., 2010; Yang et al., 2013, 2016, 2017; Ding et al., 2017b).

To consider the contribution of both the neutral and ionized species, those quantum chemical descriptors were adjusted by the following formula (Yang et al., 2013):

$$X_{adj} = \delta_n \cdot X_n + \delta_i \cdot X_i \quad (2)$$

where X_n and X_i are the descriptors for neutral and ionized compounds, respectively; δ_n and δ_i are the fractions of neutral and ionized species for a certain compound, respectively.

To characterize the hydrophobic interactions of neutral molecular, the logK_{OW} was usually used. For ionized compounds, the n-octanol/water distribution coefficient (log$D$) is more reliable than logK_{OW} to characterize the hydrophobicity because it could take the impact of ionic effect into account (Kah and Brown, 2008). In the modeling, the log$D$ value of neutral compounds is equivalent to its logK_{OW} value (Bharate et al., 2016).

In addition, due to the main difference between all compounds is their molecular structure, the functional group counts and atom-centred fragments were also introduced in the modeling.

The molecular surface potential descriptors were obtained from the Gaussian 16 output files by GsGrid V1.7 (Lu, 2010). The experimental or predicted logK_{OW} values were searched and calculated using the KOWIN module within the EPI Suit 4.1™ (U.S. EPA, 2012). The δ_n, δ_i and log$D$ were calculated by the Calculator Plugins from MarvinSketch 15.6.29.0, 2015, ChemAxon (http://www.chemaxon.com) at the experimental temperature and pH condition. The atom-centred fragments and functional group counts descriptors were calculated using DRAGON 6.0 (Talete srl, 2012).

2.3. QSAR development and validation

Stepwise multiple linear regressions (MLRs) analysis was used to select the predictive variable and construct the QSAR models employing the SPSS 19.0 software. The internal and external prediction performances of derived models were assessed by following the OECD QSAR models validation guideline (OECD, 2007). Specifically, the goodness-of-fit and robustness were evaluated by determination coefficient (R^2), leave-one out cross validation Q^2 (Q^2_{LOO}) and bootstrapping coefficient (Q^2_{BOOT}). The external predictive ability was assessed by coefficient determination (R^2_{EXT}) and external validation coefficient (Q^2_{EXT}) of the validation set (Golbraikh and Tropsha, 2002; OECD, 2007; Schüürmann et al., 2008; Roy and Ambure, 2016). The models also comply with the QUIK rules, i.e., K_{XX} (inter correlation of selected descriptors) < K_{XY} (the correlation of the X block with Y), where X is the