Chemosphere 213 (2018) 76–83

Contents lists available at ScienceDirect

Chemosphere

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

Prediction of biopersistence of hydrocarbons using a single parameter

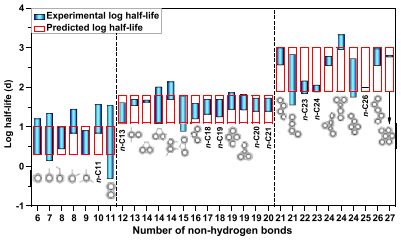
Feng Xiao*, Quinn E. Huisman

Department of Civil Engineering, University of North Dakota, Grand Forks, ND 58202-8115, United States

HIGHLIGHTS

- A new QSPR was built for the biodegradation half-lives of hydrocarbons.
- This new QSPR contains only one predictor variable.
- Number of bonds that do not contain hydrogen is the most significant predictor.
- This model was validated by an external test set of 64 hydrocarbons.
- The model shows a prediction accuracy of 95% over the external validation set.

GRAPHICAL ABSTRACT



(size: 5.0 cm in height and 8.6 cm in width)

ARTICLE INFO

Article history:
Received 15 July 2018
Received in revised form
1 September 2018
Accepted 4 September 2018
Available online 5 September 2018

Handling Editor: Keith Maruya

Keywords:
Petroleum hydrocarbons
Biodegradation
Molecular descriptors
Quantitative structure property relationship

ABSTRACT

Aerobic biodegradation is an important attenuation process for petroleum hydrocarbons (PHCs) in the natural environment. It has also been frequently used in engineered systems to remediate PHC-contaminated sites. A model such as a quantitative structure property relationship (QSPR) that can predict the biodegradation rate of PHCs would be helpful prior to implementing any extensive environmental measurements and bioremediation strategies. Existing QSPRs either have a large number of predictor variables that may cause overfitting or are based on a small dataset of PHCs. The goal of this study is to develop a simple, portable QSPR that has only a few predicator variables but can accurately predict the biodegradation half-lives of a large group of PHCs. To this end, more than 500 molecular variables were screened, and candidate variables were refined by a feature selection method and fitted to biodegradation data of a group of structurally heterogeneous PHCs ($n = 173$). The model was established by means of hierarchical clustering and classification and regression tree algorithms, which was optimized by an internal validation procedure and validated by an external dataset. The optimal QSPR model, containing only one predictor variable (the number of bonds that do not contain hydrogen), was able to accurately predict biodegradation half-lives for a wide variety of PHCs. The internal validation test indicated an overall prediction accuracy of 93%, and predictions applied to an independent external set of 64 PHCs yielded 95% accuracy. The new model is transparent and easily portable from one user to another.

1. Introduction

Over the past 15 years, the United States and other parts of the world have experienced a spectacular growth in the production of unconventional fossil fuels, thanks to technological innovations

* Corresponding author.
E-mail addresses: Feng.Xiao@UND.edu, fxiaoee@gmail.com (F. Xiao).

https://doi.org/10.1016/j.chemosphere.2018.09.035
0045-6535/

such as horizontal drilling and hydraulic fracturing. In the US state of North Dakota alone, approximately 33 million barrels of crude oil are produced each month (2017 data). The shale oil boom has brought economic expansion to oil-producing regions but also raised questions on potential environmental damages, including soil and water contamination by chemicals associated with the oil production, transportation, refining, and storage/distribution of crude oil and refined products. Petroleum hydrocarbons (PHCs) are a group of chemicals present in oil, gasoline, diesel, and a variety of solvents and penetrating oils. They may also originate from shale formation, which have been observed in the waste stream (or produced water) generated from hydraulic fracturing operations (Orem et al., 2014; Kassotis et al., 2016; Luek and Gonsior, 2017). PHCs are common soil and sediment contaminants (Ruiz-Fernández et al., 2016; Guarino et al., 2017; Guigue et al., 2017), a result of fuel combustion, pipeline leaking, and oil spills. A majority of PHCs, including long-chain alkanes, benzenes, naphthalenes, and polycyclic aromatic hydrocarbons (PAHs), are toxic to plants, animals, and humans (Eisler, 1987; Hotz, 1994; Milinkovitch et al., 2011; Abdel-Shafy and Mansour, 2016; Tran et al., 2018). Remediation of PHC-contaminated sites (Beškoski et al., 2011; Couto et al., 2011; Kronenberg et al., 2017; Tursi et al., 2018) is complicated involving not only geo-environmental factors but also a variety of practical management problems including socio-economic issues affecting the implementation of lasting solutions.

Biodegradation is a major pathway leading to the attenuation of PHCs in the natural environment (Lock et al., 1982; Lei et al., 2005). Biodegradation half-life ($t_{1/2}$) is among the most commonly used criteria for assessing the environmental fate of PHCs and for developing a feasible bioremediation strategy. The experimental measurement of $t_{1/2}$ can be time-consuming, difficult, and expensive, and there is a need for models that can predict $t_{1/2}$ of PHCs based on available data. Establishing quantitative structure property relationships (QSPRs) between molecular features and $t_{1/2}$ would fill this need and also aid in our understanding of the relationships between the chemical structure and biopersistence of PHCs.

Only a few QSPRs have been developed to predict the aerobic biodegradation rate of PHCs. Howard et al. built a QSPR to predict $t_{1/2}$ of PHCs using multiple linear regression (MLR) against 31 molecular fragments (Howard et al., 2005). Their model can explain 91% of the variance in the training set and 81% of the variance in the independent validation set (Howard et al., 2005), and has become the BioHCwin program in the US EPA's EPI Suite™ software. Their work certainly offers a very knowledgeable discussion on this topic. However, MLR or ordinary least-squares regression has a set of assumptions (e.g., homoscedasticity and linearity) to be imposed before data mining (Jambu, 1991; Mundry and Nunn, 2009). It is important to examine whether the assumptions have been violated too greatly while using MLR (Cronin and Schultz, 2003; Dearden et al., 2009). In addition, with 31 independent variables, multicollinearity may arise, introducing instability to a model. It has been found that the coefficient estimates in a QSPR can change erratically in response to a small change in the data with multicollinear parameters (Xiao et al., 2013). Furthermore, an MLR-derived QSPR can include insignificant/unnecessary independent variables. There is always a temptation to add many predictor variables in a QSPR just to increase R^2 by small amounts. Another example of this case is a QSPR containing seven quantum-chemical predictors against only 20 aliphatic PHCs (Eriksson et al., 1995). In addition to wasting degrees of freedom and generating a cumbersome model, insignificant predictor variables can cause overfitting, modeling the noise (outliers) rather than the real pattern of the data (Martens and Martens, 2001).

In another study, Xu et al. developed a QSPR for predicting $t_{1/2}$ of 17 PHCs with the molecular vibration frequency as the predictor variable that was calculated by a density functional theory method (Xu et al., 2012). Cvetnic et al. developed a QSPR with 3–5 predictors based on a dataset of 36 single-benzene ring compounds (Cvetnic et al., 2017). QSPRs in both studies predicted that low-molecular-weight PHCs are more readily degraded than PHCs with a higher molecular weight. Both QSPRs, however, were built based on a small dataset of PHCs. In Xu's study, single-benzene ring compounds and alkanes were not included, and molecular vibration frequencies are not easily calculated without the essential density functional theory software. Similarly, Cvetnic's algorithms may not be applicable to PHCs other than single-ring aromatics.

The purpose of this study is to develop and validate a new QSPR with only a few independent variables for predicting $t_{1/2}$ of a wide variety of PHCs. To this end, we extended the database compiled by Howard and coworkers (Howard et al., 2005). The model was established by means of a hierarchical cluster analysis and classification and regression tree (CRT) algorithms against molecular descriptors. Both hierarchical clustering and CRT are nonlinear, nonparametric techniques, which are well suitable for data mining that involves both numerical and categorical data, nonlinear relationships, high-order interaction, and missing values (Breiman, 1984; Burow et al., 2010; Papathomas et al., 2011; Benninghoff et al., 2012; Xiao et al., 2012). In addition, unlike MLR, they require no *priori* knowledge of the data to be processed and no previous assumptions about the underlying statistical distribution of the data. Molecular descriptors are state-of-the-art mathematical expressions of the structured information contained in a molecule. They are believed to be “the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment” (Todeschini and Consonni, 2000).

2. Methods

2.1. Database

The database compiled by Howard et al. (2005) contains 167 PHCs including alkanes, simple aromatics, and PAHs, mainly from soil and sediment biodegradation studies. In order to extend their database, we reviewed approximately 50 relevant references published since their study. We employed the same screening strategies established by Howard et al. (2005). In brief, aerobic biodegradation half-lives were gleaned from studies conducted under environmentally relevant conditions: (1) mixed-culture of bacteria; (2) environmental relevant concentrations of PHCs; (3) the temperature was within the environmental range; (4) the inoculum was not preacclimated to PHCs; and (5) additional microorganisms were not added. By these strategies, a few references were not considered. An example is a study that bacteria were taken from an enrichment culture and preacclimated to PAHs (Wammer and Peters, 2005). A recent study (Birch et al., 2018) focusing on biodegradation of PHCs in the aqueous phase (e.g., lakes and seawater) was also not included. Furthermore, we have examined outliers of half-lives in the database. An example of this is a study (Morasch et al., 2011a) in which the half-lives of PHCs were at least 20 times longer than those reported in other studies. In addition, four complicated bicyclic PHCs (endo-dihydrodi(norbonadiene), CAS 66289-74-5; diadamantane, CAS 2292-79-7; exo-tetrahydrodi(cyclopentadiene), CAS 2825-82-3; adamantine, CAS 281-23-2) in Howard's database (Howard et al., 2005) were found to be structural outliers. There were not enough bicyclic PHCs in the database to generate reliable regression results, and thus these four bicyclic PHCs were removed during the