Available online at www.sciencedirect.com
ScienceDirect
Analytica Chimica Acta 589 (2007) 150–158
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

Prediction of ozone tropospheric degradation rate constants by projection pursuit regression

Yueying Ren [a], Huanxiang Liu [a], Xiaojun Yao [a,b,*], Mancang Liu [a]

[a] *Department of Chemistry, Lanzhou University, Lanzhou 730000, China*
[b] *State Key Laboratory of Applied Organic Chemistry, Lanzhou University, Lanzhou 730000, China*

Received 12 November 2006; received in revised form 30 January 2007; accepted 16 February 2007
Available online 1 March 2007

Abstract

Quantitative structure–property relationship (QSPR) models were developed to predict degradation rate constants of ozone tropospheric and to study the degradation reactivity mechanism of 116 diverse compounds. DUPLEX algorithm was utilized to design the training and test sets. Seven molecular descriptors selected by the heuristic method (HM) were used as inputs to perform multiple linear regression (MLR), support vector machine (SVM) and projection pursuit regression (PPR) studies. The PPR model performs best both in the fitness and in the prediction capacity. For the test set, it gave a predictive correlation coefficient (R) of 0.955, root mean square error (RMSE) of 1.041 and absolute average relative deviation (AARD, %) of 4.663, respectively. The results proved that PPR is a useful tool that can be used to solve the nonlinear problems in QSPR. In addition, methods used in this paper are simple, practical and effective for chemists to predict the ozone degradation rate constants of compounds in troposphere.

Keywords: Quantitative structure–property relationship; Ozone tropospheric degradation rate constants; Heuristic method; Support vector machine; Projection pursuit regression

1. Introduction

In the troposphere, the reactions with OH radical and ozone during the day and the NO_3 radical at night are the main abiotic degradation process of volatile organic compounds (VOCs) [1]. Therefore, the impact of organic persistence on the ecosystem can be assessed by determining their reaction rate constants with OH radical, ozone and NO_3 radical. In the case of ozone degradation, the lifetime of organic is given by $\tau_{O_3} = (k[O_3])^{-1}$, where k is the rate constant for the reaction of O_3 with the organic compounds and $[O_3]$ is the O_3 concentration. The temperature decreases with the altitude increasing throughout the troposphere, and hence the rate constant k also decreases with the altitude increasing. The use of the room temperature rate constant allows the estimation of lower limit lifetimes due to reaction with O_3 [2]. However, the experimental determination of such reaction rate constants is difficult, costly and time-consuming, and there are many uncertainties in chamber conditions [3]; furthermore, the increasing number of chemicals emitted or formed into the troposphere is a great challenge over the laboratory determination. For the reasons above, reliable theoretical models to estimate rate constants of the abiotic degradability of chemicals are strongly required. Among them, quantitative structure–property relationships (QSPR) study is a useful and effective alternative approach to predict rate constants of this process.

In recent years, several tropospheric degradation QSPR models have been published and most of them deal with OH and/or NO_3 radical degradation. Relatively few works concern the ozone degradation. Based on theoretical molecular descriptors calculated by CODESSA program [4], Pompe et al. developed a six-parameter MLR model to predict the ozone tropospheric degradation of 116 chemicals. The final model results were evaluated by 10-fold cross-validation procedure and the average root mean squared error (RMSE) value was 0.99 log unit. However, this model may not be appropriate for compounds not included in the data set, i.e. its generalization capacity is uncertain, as it was evaluated not by the external test set but

* Corresponding author at: Lanzhou University, Department of Chemistry, 222, Tianshui South Road, Lanzhou, China. Tel.: +86 931 891 2578; fax: +86 931 891 2582.
E-mail address: xjyao@lzu.edu.cn (X. Yao).

0003-2670/$ – see front matter © 2007 Published by Elsevier B.V.
doi:10.1016/j.aca.2007.02.058

by the internal performance statistics derived from the training set alone [5]. Gramatica et al. also built a MLR model, which was able to predict rate constants with 82–88% accuracy for the internal validation (leave-many-out) and a somewhat higher (i.e. 90%) accuracy for the external validation. Genetic algorithm-variable subset selection (GA-VSS) procedure was used to screen molecular descriptors calculated by Dragon package. The final 6-parameter model has a RMSE of 0.73 log unit [6]. Recently, Fatemi successfully developed a neural network (NN) model using six parameters screened by GA-VSS procedure from CODESSA computed descriptors. The RMSE of calculated $\log k(O_3)$ based on the neural network model are 0.357, 0.460 and 0.481 for the training, prediction and validation set, respectively, which are smaller than those obtained by MLR model [7], and thus confirmed that there exists a nonlinear relationship between the molecular descriptors and the rate constant. To date, this is the only model that utilized nonlinear feature mapping method such as NN to predict the ozone tropospheric degradation rate constant.

In addition to neural networks, a number of nonlinear modeling methods, such as genetic algorithm, support vector machine and projection pursuit regression have been developed in the field of statistics to handle nonlinearity exhibited in a given data set. The support vector machine (SVM) has attracted attention and gained extensive applications in recent QSPR/QSAR analysis owing to its remarkable generalization performance. The projection pursuit regression (PPR) is another important nonparametric statistical technique, which seeks the "interesting" projections of data from high-dimensional to lower-dimensional space to try to find the intrinsic structure information hidden in the high-dimensional data [8]. With the obtained interesting projections direction, it can be used for further study of visual pattern recognition and regression (projection pursuit regression, PPR) [9,10]. Similar to SVM, it can also effectively overcome the curse of dimensionality. In our previous work, we have successfully applied PPR in the QSPR study of ion mobility [11]. In the present work, SVM and PPR are used to establish the quantitative relationship between molecular structure and ozone rate constant for the same data used by Pompe et al. [5]. We used Duplex algorithm on the calculated structure descriptors to generate QSPR sets, i.e. the training and test sets. Then, heuristic method was utilized to reduce the number of descriptors space and to select the structural features of the molecules relevant to the ozone degradation. Finally, using the selected variables as inputs, QSPR models were constructed by MLR, SVM and PPR. The ultimate objective is to establish reliable QSPR models for reaction rate constants prediction and, to obtain knowledge of the reactivity with ozone of substances not yet tested or for which reliable experimental data are not available as well.

2. Materials and methods

The data set for this investigation are listed in Supplementary Material 1. All the experimental rate constants, reported in $cm^3\ s^{-1}$ molecule^{-1}, were determined in carbon tetrachloride solution at 298.15 K and 101.3 kPa. Rate constants cover a range of over 12 orders of magnitude. They were transformed to logarithmic units marked as $\log k(O_3)$.

2.1. Calculation of the structural descriptors by CODESSA

To obtain a QSPR model, compounds are often represented by the molecular descriptors. The calculation process of the molecular descriptors is described as following. The structures of the compounds studied were drawn with the ISIS/Draw 2.3 program [12]. The pre-optimization of these compounds were performed using MM+ molecular mechanics force field in HyperChem 6.0 program [13] followed by a more precise optimization with AM1 method. All calculations were carried out at restricted Hartree–Fock level with no configuration interaction using the Polar–Ribiere algorithm until the root mean square gradient norm was 0.001. The resulted geometries were transferred into program package MOPAC 6.0 to calculate optimized structural co-ordinates and net atomic charges [14], which formed the inputs of CODESSA software to calculate constitutional, topological, geometrical, electrostatic, and quantum chemical descriptors. These descriptors contain the information about the connections between atoms, shape, branching, symmetry, distribution of charge, and quantum-chemical properties of the molecule.

2.2. Selection of descriptors using heuristic method (HM)

After the calculation of the descriptors, the heuristic method in CODESSA was used to search the best set of descriptors for multi-linear correlations. The heuristic method has advantages of the high speed and no restrictions on the size of the data set. It can either quickly give a good estimation about what quality of correlation to expect from the data, or derive several best regression models. Besides, it will demonstrate which descriptors have bad or missing values, which descriptors are insignificant, and which descriptors are highly intercorrelated. This information will be helpful to reduce the number of descriptors in QSPR studies.

First of all, all descriptors were checked to ensure that: (a) values of each descriptor were available for each structure; and (b) there was a variation in these values. Descriptors for which values were not available for every structure in the data set were discarded. Descriptors having a constant value for all structures were also discarded. Thereafter, all possible one-parameter regression models were tested and insignificant descriptors were removed. As a next step, the program calculated the pair correlation matrix of descriptors and further reduced the descriptors pool by eliminating highly correlated descriptors. All two-parameter regression models with remaining descriptors were subsequently developed and ranked by the regression correlation coefficient. A stepwise addition of further descriptor scales was performed to find the best multi-parameter regression models with the optimum values of statistical criteria (highest values of R^2, the cross-validated R^2_{CV}, and the F-value). The influence of the dimension of the model on its prediction capability was tested by the leave-one-out cross-validation procedure. Detailed discussion about HM can be found in article [15].