Journal of Loss Prevention in the Process Industries 24 (2011) 85–89

Contents lists available at ScienceDirect

Journal of Loss Prevention in the Process Industries

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

Predicting the net heat of combustion of organic compounds from molecular structures based on ant colony optimization

Y. Pan [a,b,*], J.C. Jiang [a], R. Wang [a], J.J. Jiang [a]

[a] College of Urban Construction & Safety Engineering, Nanjing University of Technology, Nanjing 210009, China
[b] State Key Laboratory of Fire Science, University of Science and Technology of China, Hefei 230026, China

ARTICLE INFO

Article history:
Received 13 August 2010
Received in revised form
24 October 2010
Accepted 3 November 2010

Keywords:
Net heat of combustion
Prediction
Molecular structure
Ant colony optimization
Quantitative structure–property relationship (QSPR)

ABSTRACT

A quantitative structure–property relationship (QSPR) model for prediction of standard net heat of combustion was developed from molecular structures. A diverse set of 1650 organic compounds were employed as the studied dataset, and a total of 1481 molecular descriptors were calculated for each compound. The novel variable selection method of ant colony optimization (ACO) algorithm coupled with the partial least square (PLS) was employed to select optimal subset of descriptors that have significant contribution to the overall property of standard net heat of combustion from the large pool of calculated descriptors. As a result, four molecular descriptors were screened out as the input parameters, and a four-variable multi-linear model was finally constructed using multi-linear regression (MLR) method. The resulted squared correlation coefficient R^2 of the model was 0.995 for the training set of 1322 compounds, and 0.996 for the external test set of 328 compounds, respectively. The results showed that an accurate prediction model for the net heat of combustion could be obtained by using the ant colony optimization method. Moreover, this study can provide a new way for predicting the net heat of combustion of organic compounds for engineering based on only their molecular structures.

1. Introduction

The heat of combustion of a substance is the heat evolved when that substance is converted to its final oxidation products by means of molecular oxygen (Hshieh, 1999). The standard net heat of combustion (ΔH_c°) is defined as the increase in enthalpy when a substance in its standard state at 298.15 K of temperature and 1 atm of pressure, undergoes oxidation to defined combustion products. The combustion products are CO_2 (g), F_2 (g), Cl_2 (g), Br_2 (g), I_2 (g), SO_2 (g), N_2 (g), H_3PO_4, H_2O (g) and SiO_2 (*crystobalite*) (DIPPR, 2006). ΔH_c° is an important physicochemical property which can be used to calculate the reactive heat in many chemical engineering processes such as the processes of hydrogenation and dehydrogenation of hydrocarbon. In addition, the ΔH_c° values of reactive chemicals can be used to estimate the potential fire hazards of chemicals once they ignite and burn, and thus can be used to measure the risk level of chemicals in production, storage and transportation. Thus, reliable and accurate ΔH_c° data are always required and also considered to be absolutely necessary when preparing plant designs (Cardozo, 1986). Accordingly, it can be reasonably concluded that the reliable and accurate ΔH_c° data of various chemicals is quite important and valuable in engineering design (process and plant design) and risk assessment for loss prevention.

The experimental values are the main source of the ΔH_c° data used in production. However, as we known, the measurement of ΔH_c° is expensive and time consuming, and for toxic, volatile, explosive, and radioactive compounds, the measurement is more difficult and even impossible. Moreover, the experimentalists cannot always easily obtain a pure material for experiment nor have a complete reproducible combustion during experiment process. Therefore, in order to support and expand the ΔH_c° dataset used for industry, the development of theoretical prediction methods, which are desirably convenient and reliable for predicting the ΔH_c° of chemicals is required. Once a reliable model has been obtained, it is possible to use it to predict the ΔH_c° for other chemicals not yet measured or even not yet prepared.

It is well known that the property of a compound is dependent on its molecular structures, which is a basic law in chemistry studies. Quantitative structure–property relationship (QSPR) study which based on the law above has been widely applied in prediction of

* Corresponding author. Nanjing University of Technology, Mail Box 186, No.5 Xinmofan Road, Nanjing 210009, China. Tel./fax: +86 25 83587411.
E-mail addresses: ypannjut@gmail.com, ypnjut@126.com, hades44@126.com, njut_xf030209@126.com (Y. Pan).

0950-4230/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jlp.2010.11.001

various simple and complex physicochemical properties, such as boiling point, melting point, flash point, vapor pressure, critical properties, water solubility, auto-ignition temperatures, octanol/water coefficients, and so on, which have been extensively reviewed elsewhere (Katritzky & Fara, 2005; Katritzky, Lobanov, & Karelson, 1995; Katritzky, Maran, Lobanov, & Karelson, 2000; Pan, Jiang, Ding, Wang, & Jiang, 2010; Taskinen & Yliruusi, 2003). QSPR is a mathematical method that relates the properties of interest to the molecular structures of compounds which are represented by a variety of molecular descriptors, such as spatial, electronic, topological, thermodynamic, information-content, conformational, quantum mechanical, and shape descriptors. Thousands of descriptors could be generated in QSPR studies. But only a part of them is statistically significant in terms of correlation with physicochemical property for a particular analysis, and variable selection is necessary for producing a useful predictive model. The selection of variables that are really indicative of the physicochemical property concerned is becoming one of the key steps in QSPR studies. The benefit gained from variable selection in QSPR is not only the stability of the model but also the interpretability of relationship between the descriptors and physicochemical property.

The ΔH_c° property of compounds has already been studied by several researchers such as Cardozo (1986), Gharagheizi (2008), Hshieh (1999), Hshieh, Hirsch, and Beeson (2003), and Seaton and Harrison (1990), and several models have been developed by different methods. In this work, in order to attempt a new way of predicting the ΔH_c° of organic compounds, we combined ant colony optimization (ACO) algorithm, which was presented by Dorigo, (1992) for the first time, together with partial least square (PLS) technique as a powerful tool for variable selections. The main purpose is to develop a reliable QSPR model for predicting the ΔH_c° of various organic compounds from the molecular structures alone.

2. Method and materials

2.1. Methodology of QSPR

Quantitative structure–property relationship (QSPR) study links the value of a physicochemical property to the structures of molecules through computational means. A basic assumption in the QSPR approach is that physicochemical properties of a chemical substance are closely related to its molecular structures. By encoding the structures with numerical values, termed descriptors, an indirect mathematical relationship can be found which correlates structure to physicochemical properties.

The main task in QSPR studies is to establish a numerical relationship between certain molecular property and molecular descriptors by means of statistics or some other methods:

$$\text{Property} = f(\text{molecular structure}) = f(\text{molecular descriptors})$$

Generally speaking, the QSPR studies can be divided into the following two steps:

Step 1 – to design and generate molecular descriptors;
Step 2 – to construct QSPR models with some proper descriptors.

The success of the QSPR approach can be explained by the insight offered into the structural determination of physicochemical properties, and the possibility to estimate the properties of new compounds without the need to synthesize and test them.

2.2. Ant colony optimization algorithm

As have been stated above, one of the most important problems involved in QSPR studies is to select optimal subset of descriptors that have significant contribution to the desired property. ACO algorithm is just a powerful optimization method to search for the global or near global optimal solutions. This algorithm has emerged recently as a stochastic optimization approach, which is a population-based approach. This algorithm is inspired by the behavior of real ant colonies, and the advantages of which are simple and few parameters to be adjusted. As a novel computational approach, ACO algorithm have been successfully used in many fields (Atabati, Zarei, & Borhani, 2010; Jin & Ma, 2009; Schluter, Egea, Antelo, Alonso, & Banga, 2009; T'kindt, Monmarché, Tercinet, & Laugt, 2002; Zheng, Zhou, Wang, & Cen, 2008). The basic idea in the ACO algorithm is simulation of the natural metaphor of real ant colonies' behavior. Real ants are capable of finding the shortest path from a food source to their nest without using visual cues but exploiting a chemical substance called pheromone. While walking, ants deposit a pheromone trail on the ground which is added to that previously deposited by other ants. The more pheromone on a route, the more likely that it will be chosen by ants. ACO can be best described by using the traveling salesman problem (TSP). Given a set of n cities with known distances between each pair of them, the aim of the TSP is to find the shortest path to travel all the cities exactly once and return to the starting city. The detailed description of application of ACO algorithm to TSP can be found in reference (Dorigo, 1992).

In this study, the ACO algorithm was combined with PLS method (ACO-PLS) to find the optimal subset of descriptors that accurately represented the relationships between molecular structure and the ΔH_c° property. ACO-PLS is a hybrid approach that combines ACO as a powerful optimization method and variable selection applied to PLS as a popular statistical method for modeling. In this study, the program required to perform ACO-PLS was written in MATLAB M-file in our laboratory.

In the variable selection problem, m ants select one variable, then every ant moves to another variable according to the probability defined. After one selection, the amount of pheromone is updated. This process is terminated when the error do not decline obviously while adding variables. In the ACO algorithm, the increment of pheromone left on a certain variable is measured according to a pre-defined fitness function. The following objective function is applied to variable selection in the ACO:

$$\text{FF} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \quad (1)$$

here n is the number of samples in training set, Y_i is the observed value of the training set, $\overline{Y}$ is the average value of the observed values of training set, $\widehat{Y_i}$ is the predicted value of the training set by current model. The larger the fitness function is, the larger is the correlation coefficient of the model, and the higher the probability that the model based on a combination of regression values (molecular descriptors) is being selected.

2.3. Dataset

The accuracy of the prediction model can be directly affected by the reliability of experimental values. Thus, it is very important to select a reliable database for choosing the studied dataset. One of the mostly recommended databases for physical properties of organic compounds is the DIPPR 801 project, which is developed by AIChE (American Institute of Chemical Engineers) (DIPPR, 2006). In