Chemometrics and Intelligent Laboratory Systems 159 (2016) 187–195

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Interpretable linear and nonlinear quantitative structure-selectivity relationship (QSSR) modeling of a biomimetic catalytic system by particle swarm optimization based sparse regression

Lu Xu[a,b,1], Hai-Yan Fu[c,1], Qiao-Bo Yin[c], Yao Fan[a], Mohammad Goodarzi[d], Yuan-Bin She[a,*]

[a] State Key Laboratory of Green Chemistry Synthesis Technology, College of Chemical Engineering, Zhejiang University of Technology, Hangzhou 310014, Zhejiang, PR China
[b] Institute of Applied Chemistry, College of Material and Chemical Engineering, Tongren University, Tongren 554300, Guizhou, PR China
[c] The Modernization Engineering Technology Research Center of Ethnic Minority Medicine of Hubei Province, College of Pharmacy, South-Central University for Nationalities, Wuhan 430074, PR China
[d] KU Leuven, Department of Biosystems, MeBioS Division, Kasteelpark Arenberg 30, Box 2456, B-3001 Leuven, Belgium

ARTICLE INFO

Keywords:
Quantitative structure-activity relationship (QSAR)
Quantitative structure-selectivity relationship (QSSR)
Metalloporphyrin catalysts
Selective oxidation
Sparse regression (SR)

ABSTRACT

A particle swarm optimization (PSO) based sparse regression (PSO-SR) strategy was proposed to study the quantitative structure-selectivity relationship (QSSR) of a biomimetic catalytic system, where the selectivity in the mild oxidation of *o*-nitrotoluene to *o*-nitrobenzaldehyde was related to the molecular descriptors of 48 metalloporphyrin catalysts. PSO was used to obtain an optimal variable combination for linear or nonlinear models. For nonlinear modeling, a set of 44 nonlinear transforms were developed for each single descriptor. To enable model interpretability and reduce the risk of overfitting, the total descriptors were divided into subclasses and the selected variables were forced to be sparsely distributed in each subclass. Model complexity was controlled by adjusting the maximum total number of variables included. Accurate linear and nonlinear PSO-SR models were developed using multiple linear regression (MLR) and partial least squares (PLS) and validated by randomly and repeatedly splitting the data into training and test objects for 500 times. The best predictions were obtained with 10 variables with linear (Q^2=0.9460) and nonlinear (Q^2=0.9505) models. The results indicate PSO-SR could provide an effective and useful strategy for modeling and interpreting complex QSSR problems. The proposed nonlinear modeling method could provide more information for model interpretation by probing and catching the unknown nonlinear relationship between a descriptor and the observed selectivity.

1. Introduction

Over the past half century, quantitative structure-activity/property relationship (QSAR/QSPR) modeling methods have been widely used in many fields, including chemistry, pharmacology, biology, materials, and environmental sciences. QSAR builds a bridge between a specific activity of interest for a group of molecules and their structure derived descriptors by means of a mathematical model. Originating from the seminal linear regression QSAR models by Hansch and Fujita [1,2], the current arsenal of QSAR regression techniques consists of various mathematical methods such as multiple linear regression (MLR) [3,4], partial least squares (PLS) [5,6], principal component regression (PCR) [7,8], artificial neural networks (ANNs) [9–16], kernel regression [17,18], support vector machines (SVMs) [19–24], classification and regression trees (CARTs) [25,26], and others [27–30].

During the history of QSAR, the interpretability and predictive ability have been two important aspects of a QSAR model, from which two main branches of QSAR have evolved [31–33]. On one hand, the interpretability of a QSAR model emphasizes that the mechanisms of molecular activities should be explicitly explained in physicochemical sense [34]. As in the "pure" or classical QSAR studies, a descriptive and relatively simple model is developed, which is often based on free energy relationships and combination of single mechanisms, such as the electronic, hydrophobic, lipophilic, and steric properties. Because the main aim of this type of QSAR models is interpretation rather than prediction, they are usually local models developed on a small set of similar molecules and their application domain is limited. Besides the advantage of obtaining mechanism insights, these interpretation-

* Corresponding author.
E-mail address: sheyb@zjut.edu.cn (Y.-B. She).
[1] Equally contributed to this work.

http://dx.doi.org/10.1016/j.chemolab.2016.10.016
Received 11 August 2016; Received in revised form 13 October 2016; Accepted 25 October 2016
Available online 05 November 2016

oriented models are more likely to reveal causative relationships and to reduce the risk of chance correlations [35]. On the other hand, the predictive ability of a QSAR model focuses on making reliable predictions of properties of new molecules [36,37]. A QSAR model primarily for predictions is usually trained and validated using large data sets with considerable chemical diversity. Because the chemicals in these data sets are usually non-congeneric and do not necessarily have a common core structure, it is very likely that multiple rather than single modes of action or mechanisms will be involved. Therefore, non-empirical and computational molecular descriptors and much more complex nonlinear modeling techniques are increasingly used to model the relationship between structures and activities [38–40]. Although it is more difficult to explain the models, with sufficient validation of predictive power, these models will enable reliable predictions of the properties for a wide range of new molecules, which is very useful for optimizing leads and predicting or screening for new molecules of interest from a large library. Both of the above two types of QSAR models have been widely and successfully used and the pursuit of more predictive and interpretable QSAR models is still a major challenge of QSAR field.

Metalloporphyrins (Fig. 1), a class of organometallic complexes, have been a hot research field for decades due to their biological significance and applications [41–43]. Motivated by the bioactivities of cytochrome P450 enzymes, the effects of modified metalloporphyrins in catalyzing the oxidative hydroxylation of alkanes (C-H bond) using oxygen under mild conditions have been intensively investigated [44,45]. During the process of seeking more effective metalloporphyrin catalysts, it has been recognized that the catalytic activity of metalloporphyrins largely depends on their structural characteristics, such as the type of the central metal ion, axial ligands, and peripheral substituents [46]. Therefore, in our previous studies [47], efforts have been made to obtain improved catalytic activity by modifying the structures of metalloporphyrins. However, it was found that such efforts do not necessarily bring satisfying results, because the catalytic activity is not always consistent with the selectivity of target products, which will largely influence the economic viability and resource

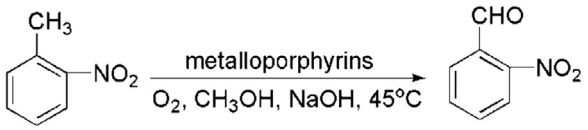


Fig. 2. Metalloporphyrin catalysts used in mild selective oxidation of *o*-nitrotoluene to *o*-nitrobenzaldehyde.

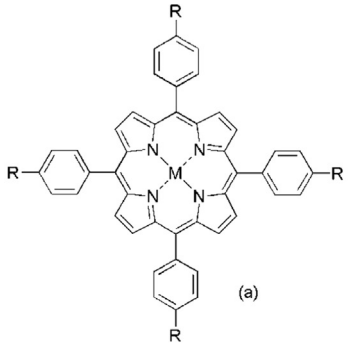
availability of a reaction. Unlike the widely investigated enantiomeric selectivity, where only two chiral isomers are involved and the observed enantiomeric excess (ee) can usually be explained by electronic or/and steric effects around the chiral center [48–51], the selectivity of a general and non-chiral catalysis system would be more complex and may involve multiple, parallel or/and subsequent reaction paths.

Therefore, a quantitative structure-selectivity relationship (QSSR) [52–56] model is required to link the structures of metalloporphyrin catalysts to the selectivity of target products. The QSSR model should present insights into the mechanisms and major influencing factors of selectivity, provide clues about how to design novel and efficient metalloporphyrin catalysts, as well as to enable the predictions of selectivity for new catalyst molecules. The objective of this work was to develop a QSSR model for the metalloporphyrins in catalytic oxidation of *o*-nitrotoluene to *o*-nitrobenzaldehyde as shown in Fig. 2. In order to ensure the interpretability and predictive ability, a novel sparse regression (SR) strategy was proposed using particle swarm optimization (PSO) algorithm [57–61] to model and interpret the complex selective mechanisms of metalloporphyrins.

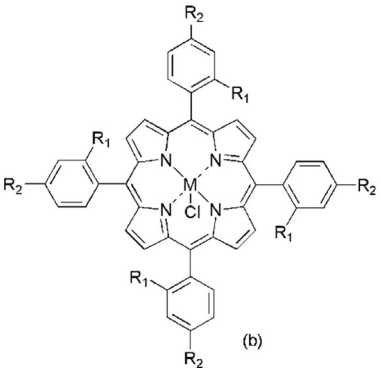
2. Materials and methods

2.1. Experimentals

A set of 48 metalloporphyrins as shown in Fig. 1 and Table 1 were synthesized using the method in references [62,63]. *O*-nitrotoluene (0.06 mol), NaOH (0.75 mol), metalloporphyrin catalyst (3×10^{-6} mol) and methanol (300 mL) were added to a 600 mL high-pressure laboratory-scale reactor. The reaction was carried out at 45 ℃ under



mol12: R=Cl;M=Fe
mol13: R=H;M=Fe
mol14: R=CH_3;M=Fe
mol15: R=OCH_3;M=Fe
mol16: R=OH;M=Fe
mol28: R=NO_2;M=Co
mol29: R=Cl;M=Co
mol30: R=H;M=Co
mol31: R=CH_3;M=Co
mol32: R=OCH_3;M=Co
mol33: R=OH;M=Co
mol45: R=Cl;M=Mn
mol46: R=H;M=Mn
mol47: R=CH_3;M=Mn
mol48: R=H;M=Cu



mol1: R_1=H;R_2=NO_2;M=Fe
mol2: R_1=H;R_2=Cl;M=Fe
mol3: R_1=H;R_2=H;M=Fe
mol4: R_1=H;R_2=CH_3;M=Fe
mol5: R_1=H;R_2=OCH_3;M=Fe
mol6: R_1=H;R_2=OH;M=Fe
mol7: R_1=NO_2;R_2=H;M=Fe
mol8: R_1=Cl;R_2=H;M=Fe
mol9: R_1=CH_3;R_2=H;M=Fe
mol10: R_1=OCH_3;R_2=H;M=Fe
mol11: R_1=OH;R_2=H;M=Fe
mol17: R_1=H;R_2=NO_2;M=Co
mol18: R_1=H;R_2=Cl;M=Co
mol19: R_1=H;R_2=H;M=Co
mol20: R_1=H;R_2=CH_3;M=Co
mol21: R_1=H;R_2=OCH_3;M=Co
mol22: R_1=H;R_2=OH;M=Co
mol23: R_1=NO_2;R_2=H;M=Co
mol24: R_1=Cl;R_2=H;M=Co
mol25: R_1=CH_3;R_2=H;M=Co
mol26: R_1=OCH_3;R_2=H;M=Co
mol27: R_1=OH;R_2=H;M=Co
mol34: R_1=H;R_2=NO_2;M=Mn
mol35: R_1=H;R_2=Cl;M=Mn
mol36: R_1=H;R_2=H;M=Mn
mol37: R_1=H;R_2=CH_3;M=Mn
mol38: R_1=H;R_2=OCH_3;M=Mn
mol39: R_1=H;R_2=OH;M=Mn
mol40: R_1=H;R_2=OH;M=Mn
mol41: R_1=NO_2;R_2=H;M=Mn
mol42: R_1=Cl;R_2=H;M=Mn
mol43: R_1=CH_3;R_2=H;M=Mn
mol44: R_1=OCH_3;R_2=H;M=Mn

Fig. 1. Parent structures and substitutes of 48 metalloporphyrin catalysts.