Computers and Chemical Engineering 106 (2017) 690–698

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

Hybrid genetic algorithm-decision tree approach for rate constant prediction using structures of reactants and solvent for Diels-Alder reaction

Shounak Datta, Vikrant A. Dev, Mario R. Eden*

Department of Chemical Engineering, Auburn University, Auburn, AL 36849, USA

ARTICLE INFO

Article history:
Received 5 October 2016
Received in revised form 23 January 2017
Accepted 13 February 2017
Available online 17 February 2017

Keywords:
Structure-rate constant relationships
QSPR
Genetic algorithm-decision tree
Reactants
Solvents
Hybrid algorithm
Divide and conquer

ABSTRACT

In recent years, Computer-Aided Molecular Design (CAMD) has been extensively used for defining and designing reactions at their maximal potential. In all of these contributions, either the structures of reactants/products have been considered to be unchanging or the solvent structure. Developing a QSPR model which not only captures the influence of reactant structures but also the solvent effect on reaction rate, is essential. Since the structures of reactants and products are related, such QSPR models will serve as a prerequisite for the simultaneous CAMD of reactants, products and solvents. They will also provide a useful tool for predicting the rate constant without relying on experiments. To develop such a QSPR, in our work, the Diels-Alder reaction with different sets of reactants and solvents was investigated. Connectivity indices were used to represent the structures of the members of each set. Principal Component Analysis (PCA) was applied to identify principal components (PCs) corresponding to the structures of reactants and solvent of each set. Linear models expressed in terms of PCs were then generated using a Decision Tree (DT) algorithm such that the R^2 value was maximized. These models formed the initial population on which the GA performed operations such as crossover and mutation to obtain model(s) with best rate constant prediction. Thus, the novelty of our approach is that after feature extraction using PCA, a DT algorithm generates an ensemble of linear models, which through the GA is transformed into a model with best fit. Our approach required much lesser generations to provide a model with highest $R^2{}_{ext}$ value as compared to the case where the DT did not initialize the population of models.

1. Introduction

The search for QSPR models to predict influence of structures of reactants and solvents on reaction rate constant has long been a challenge. According to Roy et al. (2015), QSPR (Quantitative Structure Property Relationship) models are generally linear or non-linear mathematical relationships that correlate a particular property or activity of a chemical species with their structure. Such structures are generally represented numerically by descriptors, which can be determined experimentally or theoretically as per the definition. Early attempts to develop QSPR models for the prediction of rate constant of a reaction have been restricted. Either the rate constant was studied as a function of structures of reactants while keeping the solvent structure constant or the solvent structures were varied but the reactants' structures were kept constant. With regards to the study of the influence of reactants' structures, Chaudry and Popelier (2003) developed a property model to predict the rate constant of hydrolysis of esters by utilizing quantum chemical descriptors. Estrada and Matamala (2007) have used generalized topological indices to predict the gas phase reaction rate constants of alkanes and cycloalkanes with OH, Cl and NO_3 radicals. With regards to studying the effect of solvent structures, not only have property models been derived but CAMD of solvents to enhance reaction kinetics has been pursued by researchers. Recently, Struebing et al. (2013) developed a methodology to design solvents by utilizing surrogate models and quantum chemical calculations. In the past, Stanescu and Achenie (2006) have presented a theoretical study of solvent effects on Kolbe-Schmitt reaction rates. There is a need for QSPR models that capture the combined influence of reactants' structures and solvent structures. Such models will serve two purposes: The first would be to quickly predict the rate constant without relying on experiments, while the second purpose will be the simultaneous design of reactants, products and solvents. With regards to CAMD of reactants/products, Dev et al.

* Corresponding author.
E-mail address: edenmar@auburn.edu (M.R. Eden).

http://dx.doi.org/10.1016/j.compchemeng.2017.02.022
0098-1354/© 2017 Elsevier Ltd. All rights reserved.

Nomenclature

CAMD	Computer-aided molecule design
DT	Decision tree
GA	Genetic algorithm
QSPR/QSAR	Quantitative sturcture-property/activity relationship
RBF	Radial basis function
SVR	Support vector regression

(2015) have proposed a methodology to design reactions based on thermodynamic properties of reactions. Thus, there is scope for a methodology which designs reactants, products and solvents such that the rate of reaction is optimized. With respect to QSPRs that capture reactant and solvent influence, Nandi et al. (2013) developed a quantitative structure-activation barrier relationship for Diels-Alder reaction that utilizes quantum chemical descriptors. Their aim was to construct a relationship between the activation energy and the structures of the utilized reactants and solvent. However, their data set lacked solvent variety. Recently, Zhou et al. (2014) have studied a variety of solvents for the Diels-Alder reaction in their search for new solvent descriptors though they only used one set of reactants. Thus, we have combined the data sets utilized by Nandi et al. (2013) and Zhou et al. (2014) and created a set which offers more diversity in terms of the solvents utilized. We have utilized this more diverse data set to develop a rate constant model in terms of connectivity indices. It is worth noting that Nandi et al. (2013) relied on the data set utilized in the work of Tang et al. (2012). In addition to improvement in the data set, we have also proposed an efficient hybrid GA-DT algorithm for model development which utilizes the "divide and conquer" strategy in combination with Principal Component Analysis. The GA-DT hybrid algorithms hold promise as they incur reduced computational and manual expense. Both internal and external validations were performed separately to determine model confidence. R^2 values in case of both external and internal validation were calculated as they describe model fitness of data.

1.1. Diels-Alder reaction

As presented in Fig. 1, Diels-Adler reaction is a famous and well-studied organic chemical reaction involving a conjugate diene and an alkene, which is also termed as dienophile. Evans and Johnson (1999) have provided a comprehensive review of this reaction in their work. This reaction involves cycloaddition of two reactants in the presence of a solvent. Rideout and Breslow (1980) have presented the hydrophobic acceleration of the Diels-Alder reaction. Their work focused on discussing the influence of hydrophobic cavity in organic structures for acceleration of the reaction rate. In both of the aforementioned works, the impact of the solvent on the rate constant of the reaction has been observed. This feature of the Diels-Alder reaction makes it an appealing choice for our study.

1.2. Hybrid algorithm

According to Grosan and Abraham (2007), hybridization of evolutionary algorithms produces a different algorithm which can improve the algorithm performance and overall result. A hybrid algorithm is generated by combining two or more algorithms. The hybridization of the algorithms is performed such that two or more separate algorithms systematically perform their desired tasks. In this process, the hybridization is meaningful only if the work-load of the algorithms is distributed in an efficient manner. This suggests that the level of hybridization is to be carefully determined for generating more efficient algorithms that produce better results than the singular algorithms. In this respect, Loukas (2001) has developed an adaptive Neuro-Fuzzy inference algorithm to develop better QSAR models. The aim was to develop a QSAR model to calculate the apparent inhibition constant. This project used Gaussian Member Functions of the Fuzzy system trained with hybrid back-propagation for property model development. Later, Goodarzi et al. (2009) developed a hybrid GA based Support Vector Regression (SVR) method. In their work, they used GA to optimize parameter values received from SVR to improve the prediction efficiency of the initial model. On the other hand, Jun et al. (2010) developed two separate hybrid algorithms combining GA with Support Vector Machine learning and RBF Neural Networks respectively. Their generated models also provided better prediction for developing a QSAR describing aqueous solubility of polycyclic aromatic hydrocarbons. In the aforementioned works hybrid algorithms have produced better QSAR models than the case where no hybridization was involved. These works also draw attention to the possibility of model generation requiring less computational time if the hybridization is performed properly.

2. Methodology

From the work of Tang et al. (2012) and Zhou et al. (2014, 2015), we generated a diverse data set of 72 reactions that consisted of 38 different dienophiles, 19 dienes and 10 solvents. These reactions, along with respective experimental and predicted reaction rates, are presented in Appendix A. All chemical species were designed using Avogadro software. The structures were optimized using MMFF94s, a built-in geometry optimization algorithm of Avogadro software, as suggested by Datta et al. (2015). The optimized geometries were saved as mol files. These files were then used as input for the Dragon 6 software to calculate descriptors. Following convention, one sixth of the reactions were separated for external validation and all other reactions were used to train the model. Keeping reaction design simplicity in mind, only connectivity descriptors were used for model development.

2.1. Divide and conquer approach

This method has been previously utilized in many studies that required the handling of complicated datasets. Bently (1980) has expressed the significance of multidimensional divide and conquer algorithm in his work. He has presented some classic point-based problems, where the main goal is to determine the domination of a point over other points present in dataset. This domination is determined by ranking the points, which refers to the influence or significance of the point in dataset. Fig. 2 shows a simple representation of the algorithm. Zhang (2004) has shown application of this algorithm in his peptide sequencing algorithm. Cheng et al. (2012) used this technique to improve their template-based modeling for their protein modeling project. The concept has also been discussed in detail in the work of Hemmateenejad et al. (2004). They used PCA on different classes of descriptors to minimize the workload of the GA. However, in this work, subsets of chemical species were subject to this strategy. The three subsets used were dienophiles, dienes and solvents. Application of PCA over these sets reduced the number of meaningful descriptors to 32. These descriptors were then used as input to the DT algorithm for generating the initial populations in the GA.

2.2. Modification in DT algorithm

Decision tree is an algorithm that represents possible choices and their probable outcomes. This unique algorithm maps the best possible path to follow to reach the desired result. As presented in