Chemometrics and Intelligent Laboratory Systems 118 (2012) 79–87

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Modeling physical and toxicity endpoints of alkyl (1-phenylsulfonyl) cycloalkane-carboxylates using the Ordered Predictors Selection (OPS) for variable selection and descriptors derived with SMILES

Eduardo Borges de Melo *

Theoretical Medicinal and Environmental Chemistry Laboratory (LQMAT), Department of Pharmacy, Western Paraná State University (Unioeste), 2069 Universitária St, Cascavel, Paraná, 85819-110, Brazil

ARTICLE INFO

Article history:
Received 30 January 2012
Received in revised form 10 August 2012
Accepted 16 August 2012
Available online 24 August 2012

Keywords:
QSAR
Alkyl(1-phenylsulfonyl) cycloalkane-carboxylate
Toxicity
LogK_{ow}
LogK_{oc}
LogS_w

ABSTRACT

Among the methods of variable selection for Quantitative Structure-Property Relationship (QSPR) studies, one of the currently available alternatives is the Ordered Predictors Selection (OPS). Using this algorithm and descriptors obtained using only Simplified Molecular Input Line Entry System (SMILES) strings in the free web server Parameter Client, a QSPR study with a data set of 28 alkyl (1-phenylsulfonyl) cycloalkane-carboxylates and six different endpoints of environmental importance were developed and compared with other works. The comparison with models previously published was performed only with the internal validation, and four of the six new models proved to be superior. However, the six new models also presented high quality for external predictions, were robust and showed no chance correlation. The predicted endpoints of the six models were within the applicability domain. Thus, it can be concluded that the OPS algorithm was able to generate QSA(P)R models with high statistical quality for predicting of physicochemical and toxicological endpoints, thus showing its potential for development of predictive models of environmental interest.

1. Introduction

Quantitative structure-activity and structure-property relationship (QSAR and QSPR, or QSA(P)R) studies, which correlate and predict physical–chemical and biological properties of environmental, industrial, and medicinal importance, from molecular descriptors experimentally or theoretically derived, currently play an important role in the effective assessment of organic compounds [1–4]. The ultimate role of the QSA(P)R theory is to suggest mathematical models for estimating endpoints of interest, especially when the experimental values are not available for some reason [1,5]. Currently, this approach is an important support tool for aiding the rational drug discovery [6–8] as well as for environmental and regulatory purposes [9–13]. In June 1st 2007 it came into force in the European Community the new legislation of Registration, Evaluation, Authorization, and Restriction of Chemicals (REACH), which considers, among other topics, the utility of validated QSA(P)R models as a preliminary tool for calculating the properties of chemicals [14,15].

The determination of experimental properties (for instance, LogK_{ow}) are normally time consuming, expensive, and labor intensive [16]. In some cases, they are not available because the compound does not yet exist (for instance, for a new drug or pesticide in development) [17]. Thus, theoretical molecular descriptors are usually necessary [1,5,18]. But, nowadays, it is possible to calculate thousands of molecular descriptors for a single molecule that may be highly correlated to each other or irrelevant to the target endpoint. Therefore, the problem in selecting those which are the most representatives for the endpoint under study should be seriously considered.

The step of variable selection in a QSA(P)R study is a way to identify reduced subsets of molecular descriptors that in fact are useful to reproduce the observed values of an endpoint by an accurate regression model. This point has become important for several researches of interest for areas that manipulate data sets with a large number of independent variables, just as currently are the QSA(P)R studies. The use of a good method for variable selection helps obtaining the necessary subset to build the optimal mathematical model for the prediction of a particular endpoint and, therefore, simple, robust, and more easily interpretable models [19,20].

Several variable selection methods have been developed since the early years of chemometrics. Some of the most used are the stepwise regression method (SWR), genetic algorithms (GA), cluster significant analysis (CSA), *K*-nearest neighbor (KNN), among others [18,19]. Recently, Teófilo et al. [20] developed the Ordered Predictors Selection (OPS), a new algorithm of variable selection, useful for constructing multivariate mathematical models. This method has produced good results in QSA(P)R [21–24] and other multivariate studies [25,26] involving large amounts of data.

* Tel.: +55 45 3220 3256; fax: +55 45 3220 3280.
E-mail address: eduardo.melo@unioeste.br.

0169-7439/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2012.08.006

The aromatic sulfones constitute a class of chemicals extensively used by the pharmaceuticals, agrochemicals, petrochemical and metallurgical industries [27]. Therefore, the introduction of these compounds into the environment as well as its ecological and toxicological consequences have been increasing interest from research groups. Over the last years, several QSA(P)R works with these compounds were developed. These studies were based in some physical and toxicological endpoints, and were developed using, in most cases, the Linear Solvation Energy Relationships (LSER) solvatochromic parameters or related theoretical descriptors (TSLER) [2,3,28–35]. Yin et al. [36] also developed a QSAR study using quantum-mechanical descriptors calculated by semi-empirical approaches. Other studies indicate the use of E-state indices [37–39], molecular shape index [38], and molecular connectivity indices [37,40].

In this study, the objective is to obtain QSA(P)R models with high internal and external statistical quality for the most studied endpoints (three physical and three toxicological) in the previously cited papers (and available in databases or printed copies) for the 28 alkyl(1-phenylsulphonyl)-cycloalkane-carboxilates (Fig. 1 and Supplementary Material, Fig. S1), using the OPS algorithm [20] and only descriptors derived by Simplified Molecular Input Line Entry System (SMILES) strings [41].

2. Methodology

2.1. Data set, training set, test set, and obtainment of descriptors

In this study, six endpoints were used. The values of LogK_{ow} (logarithm of 1-octanol/water partition coefficient), LogK_{oc} (logarithm of adsorption coefficient for soil and sediments) and Log$S$ (logarithm of water solubility) were obtained in Chen et al. [30]; the toxicity against *Daphnia magna straus* (inverse logarithm of medium effective immobilization concentration, $-\text{Log}EC_{50\_Daphnia}$, and inverse logarithm of median lethal concentration, $-\text{Log}LC_{50\_Daphnia}$) in Wang et al. [2]; and the toxicity against *Photobacterium phosphoreum* (inverse logarithm of concentration that causes a 50% inhibition of bioluminescence after a 15-minute exposure, $-\text{Log}EC_{50\_Photobact}$) in Chen and Wang [34]. The values of each endpoint and the SMILES strings of each compound (28) of the data set are presented in Table 1.

The structures of all compounds were built in the JME Editor (http://www.molinspiration.com/jme) and the SMILES strings were obtained. These strings were used to generate 717 descriptors, divided into eleven categories: (i) walk and path counts; (ii) topological charge index; (iii) 2D autocorrelations; (iv) Burden eigenvalues; (v) edge adjacency indices; (vi) topological descriptors; (vii) molecular properties; (viii) information indices; (ix) connectivity indices; (x) eigenvalue-based indices; and (xi) E-state indices. All descriptors were obtained through the online programs E-Dragon 1.0 and ETState, using the interface Parameter Client (http://www.vcclab.org/lab/pclient). Only these descriptors were used in order to obtain them quickly and easily. Furthermore, the JME Editor and Parameter Client are free and available on the Internet. Its functioning on any operating system depends only on the availability of the Java Runtime Environment (http://www.java.com) and an internet browser that supports this programming language. These characteristics facilitate the reproduction of the results and the effective use of the proposed models by researchers and companies that develop and use alkyl (1-phenylsulfonyl) cycloalkane-carboxylates derivatives.

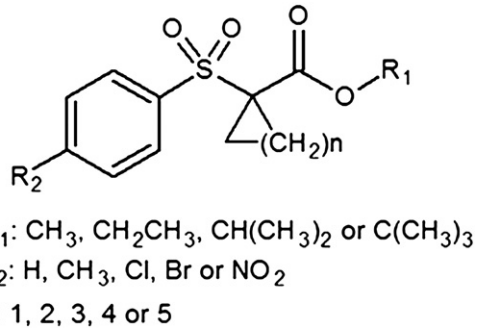

Fig. 1. Basic structure of the alkyl(1-phenylsulphonyl)-cycloalkane-carboxilates studied.

The invariant and quasi invariant descriptors, the descriptors with absent values (represented by the code "999"), and the atom type count descriptors used in the calculations of E-state indices were manually removed. The descriptors ALog$P$ (Ghose-Crippen octanol-water partition coefficient), AlogP^2 (Squared Ghose-Crippen octanol-water partition coefficient), Mlog$P$ (Moriguchi octanol-water partition coefficient) and MLogP^2 (Squared Moriguchi octanol-water partition coefficient), calculated in the "molecular properties" option, were also removed in the study with LogK_{ow}. For the other endpoints, the LogK_{ow} was utilized as the molecular descriptor related with hydrophobicity. The next step was to eliminate the descriptors with absolute Pearson correlation coefficient ($|r|$) lower than 0.3, leaving 433 descriptors for LogK_{ow}, 435 for LogK_{oc}, 452 for Log$S$, 443 for $-\text{Log}EC_{50\_Daphnia}$, 444 for $-\text{Log}LC_{50\_Daphnia}$, and 442 for $-\text{Log}EC_{50\_Photobac}$. These matrices were subjected to variable selection algorithm with the OPS.

As the data set is relatively small (twenty-two derivatives for LogK_{oc} and twenty-eight for the other), it used the approach suggested by Ferreira and Kiralj [43] to ensure that the division of the data set into training and test sets did not undermine the study, leading to non-representative models of the structures under study. Initially, an auxiliary model with the complete dataset was obtained. After internal validation, the dataset was divided into training set, generating the real model, and test set, always formed by five compounds, which is the minimum recommended in the literature. Then, the real model was used to predict the endpoint of the test set. For its use to be possible, it had to present an internal statistic as close as possible to the auxiliary model [44].

Subsequently, the data set was manually split in a training set (17 compounds for LogK_{oc} study and 23 for the others) and in test sets (five compounds for all endpoints). This step was performed with the aid of a Principal Component Analysis (PCA), a classification method that is able to represent the pattern of similarity between data [42,45]. In this case, if the dependent variables of the twenty-two compounds common to each endpoint (Table 1) present similar behavior. As the first Principal Component (PC1) cumulated the greater amount of information (86.232%), the similarity between the values of five endpoints (LogK_{ow}, LogK_{oc}, $-\text{Log}EC_{50\_Daphnia}$, $-\text{Log}LC_{50\_Daphnia}$ and $-\text{Log}EC_{50\_Photobact}$) is evidenced. This result may be visualized in the loading plot (Fig. 2). Thus, the same test set (compounds **3**, **9**, **16**, **22** and **25**) was chosen for these five endpoints. On the other hand, as the LogS_w showed different information, other data set (**6**, **12**, **13**, **20** and **27**) was selected. Both test sets were selected to cover, in the most appropriate way, the range of variation of each endpoint. The PCA analysis was performed in the Pirouette 4 (http://www.infometrix.com).

2.2. Chemometric analysis

2.2.1. OPS algorithm and PLS regression method

OPS [20] is an iterative algorithm for variable selection, that uses Partial Least Squares (PLS) [46] for building models. For the algorithm to work, it is initially necessary to determine the maximum number of latent variables (LV) that the user wishes to be explored to obtain the models (the maximum value is equal to the number of columns of the data matrix) and the number of LV which the user wishes to be used for building the models (the maximum value is equal to the number of LV which will be explored). After this, the models are