Chemometrics and Intelligent Laboratory Systems 143 (2015) 122–129

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Application of k-means clustering, linear discriminant analysis and multivariate linear regression for the development of a predictive QSAR model on 5-lipoxygenase inhibitors

Matías F. Andrada [a], Esteban G. Vega-Hissi [a,b], Mario R. Estrada [a], Juan C. Garro Martinez [a,b,*]

[a] Area de Química Física, Facultad de Química, Bioquímica y Farmacia, Universidad Nacional de San Luis, Chacabuco 917, San Luis, 5700, Argentina
[b] Centro Científico Tecnológico San Luis (CCT-CONICET), Chacabuco 917, San Luis, 5700, Argentina

ARTICLE INFO

Article history:
Received 19 December 2014
Received in revised form 2 March 2015
Accepted 4 March 2015
Available online 12 March 2015

Keywords:
QSAR
5-Lipoxygenase inhibitors
k-Means clustering
Linear discriminant analysis
Multivariate linear regression

ABSTRACT

In this work, we performed a quantitative structure activity relationship (QSAR) model for a family of 5-lipoxygenase (5-LOX) inhibitors using k-means clustering and linear discriminant analysis (LDA) for the selection of training and test sets and multivariate linear regression (MLR) for the independent variable selection. With the k-means clustering method, the total set of compounds (58 derivatives of 5-Benzylidene-2-phenylthiazolinones) was divided in two clusters according to a simple discriminant function. We found that *pilD* (conventional bond order *ID* number) molecular descriptor discriminates correctly 100% of the compounds of each clusters. Thirty different models divided in three series were analyzed and the series with representative training and test sets (series 3) had the most predictive models. The statistical parameters of the best model are $R_{train} = 0.811$ and $R_{test} = 0.801$. We found that a rational selection in the setting-up of training and test sets allows to obtain the most predictive models and the random selection is sometimes unsuitable, especially, when the total set of compounds can be classified in different clusters according to structural features.

1. Introduction

The 5-lipoxygenase (5-LOX) is a key enzyme involved in the first step of the synthesis of leukotrienes (LTs), a type of eicosanoid inflammatory mediators. The dysregulation of this enzyme causes various inflammatory diseases such as asthma, inflammatory bowel disease (IBD), chronic obstructive pulmonary disease (COPD), arthritis, psoriasis, and atherosclerosis [1–3]. It has been recently reported that increased production of LTs is associated with the increased risk for myocardial infarction, stroke [4] and cancer [5]. Most of the drugs that inhibit LT production are based on the suppression of the ligand–receptor interaction, inhibition of leukotriene A4 hydrolase or indirect interference in the activation of 5-LOX [6,7]. At the moment, the only drug approved as a direct 5-LOX inhibitor is Zileuton (N-[1-(1-benzothien-2-yl)ethyl]-N-hydroxyurea), Fig. 1 [8]. With the aim of finding new drugs that present fewer adverse effects than Zileuton, many 5-LOX inhibitors have been designed and synthesized in the recent years [9–15].

One of the most used tools in drug design aided by computers is the quantitative activity–structure relationship (QSAR). This methodology is a mathematical hypothesis based on the assumption that the molecular structure is responsible for the biological activity of a compound. Thus, entities with similar molecular structure would present the same biological activity. Since 2011, eight QSAR studies specifically targeted to 5-LOX inhibitors have been performed, showing the current interest in the development of new QSAR models specific for 5-LOX inhibitors which serve to elucidate the key structural features for the inhibition [16–23].

In QSAR, the relationship between the molecular structure and the biological activity is quantified by means of a mathematical equation using the activity as the dependent variable and the structural parameters (called molecular descriptors) as independent variables. The search and development of an optimal QSAR model that relates the dependent and independent variables can be generally divided into three stages: data preparation, data analysis, and model validation. These stages are carried out using several mathematical techniques. The last step, model validation, is a crucial aspect which is performed once the model has been built. The most commonly used criteria for validation are the leave-one-out (loo) and leave-more-out (l%o) cross-validations, external validation (using a test set) and y-randomization approach. A high value of the statistical feature ($R^2 > 0.5$) in the cross-validations is considered proof of the high predictive ability of a model. Within the data analysis stage, the partial least squares (PLS), the multivariate linear regression (MLR), and the artificial neural network (ANN) are the techniques used for the selection of a subset of the most relevant molecular descriptors [24].

* Corresponding author at: Area de Química Física, Facultad de Química, Bioquímica y Farmacia, Universidad Nacional de San Luis, Chacabuco 917, San Luis, 5700, Argentina.
E-mail address: jcgarro@unsl.edu.ar (J.C. Garro Martinez).

http://dx.doi.org/10.1016/j.chemolab.2015.03.001
0169-7439/

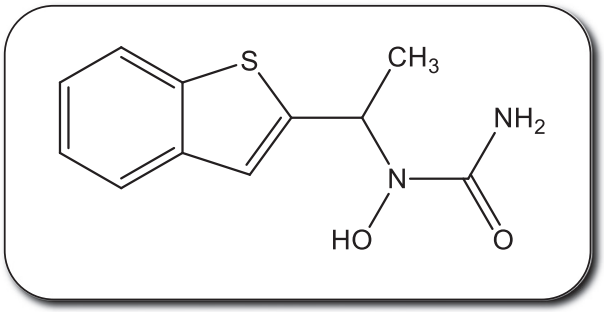

Fig. 1. Molecular structure of Zileuton.

Other interesting approaches not so commonly used in this type of studies are the k-means clustering and linear discriminant analysis (LDA). The selection of the sets (training and test sets) during the data preparation stage is generally performed using random selection. However, this may be inappropriate when the data set can be divided in different clusters according to the structural characteristics. In these cases, if random selection is applied, all members of the validation set can belong to the same group yielding a set unrepresentative of the whole. The k-means clustering is a statistical method that is used to assign groups (clusters) according to certain properties that the elements have in common (molecular descriptors) [25]. So, aided by this method, the members of training and test sets can be selected so as to be representative of the existing clusters and the total data set. LDA is the other statistical method used to characterize or separate two or more classes of objects and in the dimensionality reduction. This method allows obtaining a linear regression which discriminates the objects in each group and thus, it is able to find features (descriptors) responsible for such discrimination. Unfortunately, few QSAR studies combine these techniques and the selection of the training and test set becomes random.

In the present work we have developed a QSAR analysis for a series of 5-Benzylidene-2-phenylthiazolinones with 5-LOX inhibitory activity [9,10]. In contrast with other papers, we have employed the goodness of k-means clustering, linear discriminant analysis (LDA) and multivariate linear regression (MLR) to perform a thorough search of a predictive QSAR model.

2. Materials and methods

2.1. Data set

The data set used in this study is composed of 58 derivatives of 5-Benzylidene-2-phenylthiazolinones with known 5-LOX inhibitory activity. This set and the experimental activities were extracted from two studies performed by the same research group [9,10]. The IC_{50} values (concentration of a compound required to inhibit 50% of the 5-LOX activity) exhibit a range of activity from 60 to 11,000 nM. They were converted to the corresponding $\log(1/IC_{50})$ and used as the dependent variable in QSAR investigations. The values of the biological activity as well as the numbering of the compounds included in the data set are presented in Table 1.

2.2. Geometric optimizations and molecular descriptors

The molecular structure of the 58 compounds was optimized at the semiempirical PM3 (parametric method-3) method using the Polak-Ribiere algorithm and a gradient norm limit of 0.01 kcal $Å^{-1}$ with Hyperchem 7.0 package. Then, a set of 1497 molecular descriptors were computed using the Dragon program [26] including all types of descriptors such as Constitutional, Topological, Geometrical, Charge, GETAWAY (Geometry, Topology and Atoms-Weighted AssemblY), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), Molecular Walk Counts, BCUT descriptors, 2D-Autocorrelations, Aromaticity Indices, Randic Molecular Profiles, Radial Distribution Functions, Functional Groups, Atom-Centered Fragments, Empirical and Properties. The descriptors with a correlation higher than 0.9 were removed. Thus, the redundant information was avoided and the full set was reduced to 1195 molecular descriptors.

2.3. The k-means clustering

The k-means clustering is one of the simplest algorithms that solve the clustering problem [27]. This approach follows a simple and easy way to classify a given object through a certain number of fixed clusters (k). In QSAR studies, the results of k-means clustering have been utilized to perform a correct division of data sets into training and test sets using some characteristic information such as the calculated molecular descriptors [28–30]. In the present study, the data set of 58 compounds (objects) was analyzed assigning different values (2, 3 and 4) to the variable k using Matlab 7.0 [31]. Thus, the possibility that the total data set can be divided in 2, 3 and 4 clusters was investigated.

2.4. Linear discriminant analysis

The linear discriminant analysis (LDA) is a method used to find a linear combination of features which characterizes or separates (discriminates) two or more classes of objects (compounds in these study) [32,33]. In some QSAR studies, the LDA was utilized to identify structural features that separate the active and inactive compounds [34]. Here, we use LDA to get a multivariate discriminant function that achieves the separation of compounds of the different clusters obtained from the k-means clustering. Thus, the variables (molecular descriptors) that cause this discrimination can be identified. The calculations of LDA were performed using Matlab 7.0 [31].

2.5. Development and validation of the QSAR model

The data set was divided into training and test set (80% and 20% of the total data set, respectively). A series of 31 different combinations of training and test sets were screened.

All the QSAR models were developed employing the replacement method (RM) as the molecular descriptor selection approach [35]. In earlier reports [36,37], this method has been proven to produce linear QSAR models that are quite close to the full search methods with lower computational cost [38,39]. The RM is an efficient optimization tool which generates multivariate linear QSAR models by searching an optimal subset of **d** descriptors from a set of **D** descriptors ($d \lll D$) with minimum standard deviation (S) of the model. The regression coefficient (R) and the standard deviation (S) were the statistic parameters used for the quantified the models qualify.

The models developed in this study were validated with a test set which does not belong to the training set. In addition, the QSAR selected as the optimal model was also validated using: a) the leave-one-out (loo) and b) the leave-more-out (l%o) cross-validation procedures, generating a million cases of random data removal for l%o, where the % is $\approx$20 (twelve compounds); and c) y-randomization. This last validation consists in the interchange of the experimental property such that the property value and the compound do not match. We carried out 10,000 cases of y-randomization. The algorithms used in this work are included in Matlab 7.0 [31].