

# 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<sup>a</sup>, Esteban G. Vega-Hissi<sup>a,b</sup>, Mario R. Estrada<sup>a</sup>, Juan C. Garro Martínez<sup>a,b,\*</sup>

<sup>a</sup> Area de Química Física, Facultad de Química, Bioquímica y Farmacia, Universidad Nacional de San Luis, Chacabuco 917, San Luis, 5700, Argentina

<sup>b</sup> 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 *piID* (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  $R_{\text{train}} = 0.811$  and  $R_{\text{test}} = 0.801$ . 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.

© 2015 Elsevier B.V. All rights reserved.

## 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](mailto:jcgarro@unsl.edu.ar) (J.C. Garro Martínez).

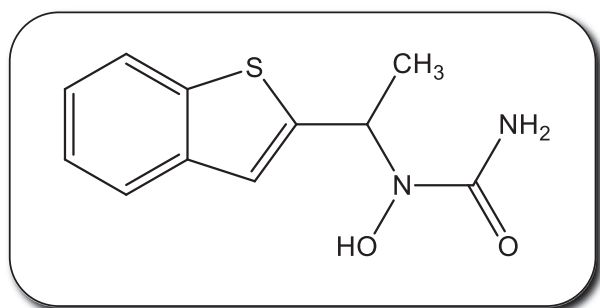


Fig. 1. Molecular structure of Zileuton.

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  $0.01 \text{ kcal } \text{Å}^{-1}$  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 ( $d \ll D$ ) 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].

**Table 1**  
Structure and biological activity (nM) of the total set of compounds.

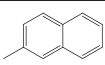
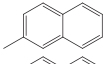
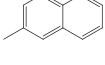
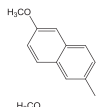

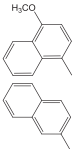








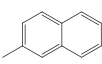

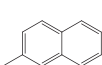
ID. mol	R1	R2	IC <sub>50</sub> (nM) <sup>a</sup>	log (1/IC <sub>50</sub> )	Predicted <sup>b</sup> log (1/IC <sub>50</sub> )
1	4-CH <sub>3</sub>	4-OCH <sub>3</sub>	210	-2.322	-2.611
2	4-Cl	4-OCH <sub>3</sub>	90	-1.954	-2.106
3	4-OCH <sub>3</sub>	4-OCH <sub>2</sub> CH <sub>2</sub> CH <sub>3</sub>	210	-2.322	-2.596
4	4-OCH <sub>3</sub>	4-Br	90	-1.954	-2.196
5	4-OCH <sub>3</sub>	4-F	550	-2.740	-2.397
6	4-OCH <sub>3</sub>	3-Cl	100	-2.000	-1.633
7	4-OCH <sub>3</sub>	2-Cl	180	-2.255	-2.438
8	4-OCH <sub>3</sub>	4-COPh	150	-2.176	-1.860
9*	4-OCH <sub>3</sub>	3-OH, 4-NO <sub>2</sub>	4660	-3.668	-
10	4-OCH <sub>3</sub>	4-NHCOCH <sub>2</sub> CH <sub>3</sub>	760	-2.881	-2.885
11*	4-OCH <sub>3</sub>	4-NHCOCH <sub>2</sub> CH <sub>2</sub> Ph	11050	-4.043	-
12	4-OCH <sub>2</sub> CH <sub>2</sub> CH <sub>3</sub>	4-Cl	80	-1.903	-2.692
13	3-OPh	4-Cl	90	-1.954	-1.786
14	4-OCH <sub>2</sub> Ph	4-Cl	80	-1.903	-1.961
15	4-OCH <sub>2</sub> CH <sub>2</sub> CH <sub>3</sub>	4-OCH <sub>3</sub>	2140	-3.330	-2.859
16	4-OPh	4-OPh	90	-1.954	-2.137
17*	4-OCH <sub>3</sub>	4-OCH <sub>3</sub>	4000	-3.602	-
18	4-OCH <sub>3</sub>	4-NH <sub>2</sub>	630	-2.799	-2.683
19	4-OCH <sub>2</sub> CH <sub>3</sub>	H	500	-2.699	-2.726
20	3-Cl	4-CH <sub>3</sub>	300	-2.477	-2.584
21	2-OH, 3-OCH <sub>3</sub> , 5-Cl	H	300	-2.477	-2.771
22	2-OH, 3-OCH <sub>3</sub>	H	540	-2.732	-2.943
23	3-OCH <sub>3</sub> , 4-OH, 5-Cl	H	3000	-3.477	-3.063
24	3-OCH <sub>3</sub> , 4-OH, 5-NO <sub>2</sub>	H	3000	-3.477	-3.306
25	4-OCH <sub>3</sub>	4-CH <sub>3</sub>	300	-2.477	-2.477
26	2-OCH <sub>3</sub> , 5-OCH <sub>3</sub>	4-CH <sub>3</sub>	130	-2.114	-2.383
27	2-OCH <sub>3</sub> , 3-OCH <sub>3</sub> , 4-OCH <sub>3</sub>	4-CH <sub>3</sub>	400	-2.602	-2.793
28	2-OCH <sub>3</sub> , 4-OCH <sub>3</sub>	4-CH <sub>3</sub>	980	-2.991	-2.803
29	2-OH, 3-OCH <sub>3</sub>	4-CH <sub>3</sub>	1300	-3.114	-2.964
30	3-OCH <sub>3</sub> , 4-OH, 5-Cl	4-CH <sub>3</sub>	2700	-3.431	-3.001
31	3-OCH <sub>2</sub> CH <sub>3</sub> , 4-OH, 5-Cl	4-CH <sub>3</sub>	1250	-3.097	-3.018
32	H	4-CH <sub>3</sub>	350	-2.544	-2.582
33	t-Bu	H	300	-2.477	-2.435
34	4-OCH <sub>3</sub>	H	90	-1.954	-1.959
35	4-OCH <sub>3</sub>	H	150	-2.176	-2.164
36	4-OCH <sub>2</sub> (O)OCH <sub>3</sub>	4-CH <sub>3</sub>	190	-2.279	-2.730
37	4-OCH <sub>3</sub>	4-OCH <sub>2</sub> C(O)OCH <sub>3</sub>	580	-2.763	-2.510
38	4-OCH <sub>3</sub>	4-OH	650	-2.813	-2.551
39	4-OCH <sub>3</sub>	4-OCH <sub>3</sub>	110	-2.041	-2.498
40	4-OCH <sub>3</sub>	4-NH <sub>2</sub>	130	-2.114	-2.588
41	4-OCH <sub>3</sub>	4-C(O)CH <sub>3</sub>	120	-2.079	-2.462
42	4-OCH <sub>3</sub>	3-C(O)CH <sub>3</sub>	110	-2.041	-2.150
43	4-OCH <sub>3</sub>		80	-1.903	-2.016
44	3-OH, 4-OCH <sub>3</sub>		80	-1.903	-2.180
45	4-OPh		60	-1.778	-1.892
46		4-Cl	70	-1.845	-1.961
47		4-Cl	130	-2.114	-2.237
48*		4-OCH <sub>3</sub>	4310	-3.634	-

Table 1 (continued)

ID. mol	R1	R2	IC <sub>50</sub> (nM) <sup>a</sup>	log (1/IC <sub>50</sub> )	Predicted <sup>b</sup> log (1/IC <sub>50</sub> )
49*		4-OCH <sub>3</sub>	5840	-3.766	-
50		4-OCH <sub>3</sub>	570	-2.756	-2.591
51		4-OPh	130	-2.114	-2.294
52		4-OPh	140	-2.146	-2.294
53		4-CH <sub>3</sub>	230	-2.362	-1.921
54		H	300	-2.477	-2.102
55		4-OCH <sub>3</sub>	2080	-3.318	-2.302
56*		4-ONH <sub>2</sub>	6260	-3.797	-
57			170	-2.230	-1.821
58			790	-2.898	-3.002

\*Compounds considered outliers.

<sup>a</sup> Experimental IC<sub>50</sub> obtained by the S100 assay [9,10].<sup>b</sup> log(1/IC<sub>50</sub>) predicted from Eq. (2).

### 3. Results and discussion

#### 3.1. Outliers compounds

One point to consider before carrying out a QSAR study is that the range of biological activity covered should be as large as possible and symmetrically distributed around its mean [40]. To address this issue an analysis of the data dispersion was performed. As can be seen in the plot of log(1/IC<sub>50</sub>) presented in Fig. 2, the distribution of the data is not quite symmetrical around the mean (-2.58). Therefore, considering a limit value of ± 1.5 S (1.5 times the standard deviation) around the mean, six values lied out of the range (compounds id. 9, 11, 17, 48, 49, and 56, see Fig. 2). These compounds were considered outliers and the total set was reduced to 52 compounds.

#### 3.2. k-Means clustering analysis

The k-means clustering method, incorporated in the software package Matlab 7.0, was used to analyze the possibility to split the data set into clusters. The total set was separated into two, three and four clusters

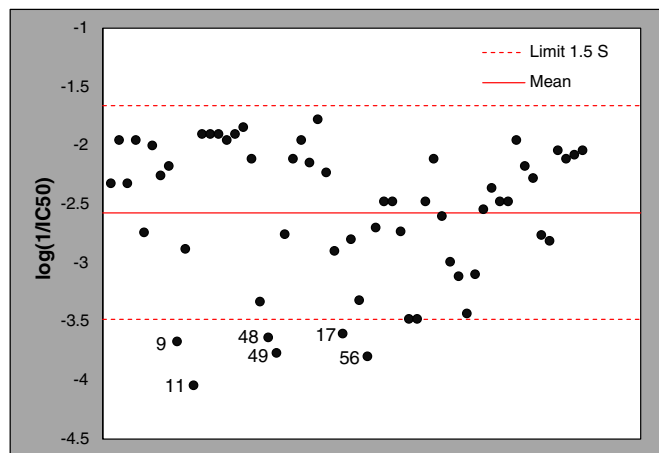


Fig. 2. Dispersion plot of log(1/IC<sub>50</sub>) values. The numbered points are the outlier compounds.

( $k = 2, 3$  and  $4$ ) and the results are displayed as silhouette plots (numbers of cluster versus silhouette values) in Fig. 3. The silhouette plot represents a measure of how close each point that belongs to one cluster is to the points of the neighboring clusters [41]. The silhouette values range from  $+1$  (points that are very distant from neighboring clusters) to  $-1$  (points that are probably assigned to the wrong cluster). A value equal to  $0$  (zero) is assigned to points that cannot be ascribed to any cluster.

The silhouette plot of the data separated into two clusters (Fig. 3, top left panel) shows that almost all points in each cluster have a large silhouette value which indicates that the cluster members are separated from the neighboring clusters. The partitioning into three clusters (Fig. 3, top right panel) leads to one cluster populated with high positive value points (cluster 2) and two clusters (cluster 1 and 3) that contain many points with low silhouette values, and a few points with negative values, indicating that these two clusters are not well separated or classified. The bottom panel of Fig. 3 contains the silhouette plot of the same data, but now partitioned into 4 clusters. The values indicate that this is probably not the right number of clusters since two of the clusters contain points with mostly low silhouette values and others with negative values. So, the results of k-means clustering indicate that the total set has two clearly marked clusters which compounds are tabulated in Table 2.

### 3.3. Linear discriminant function

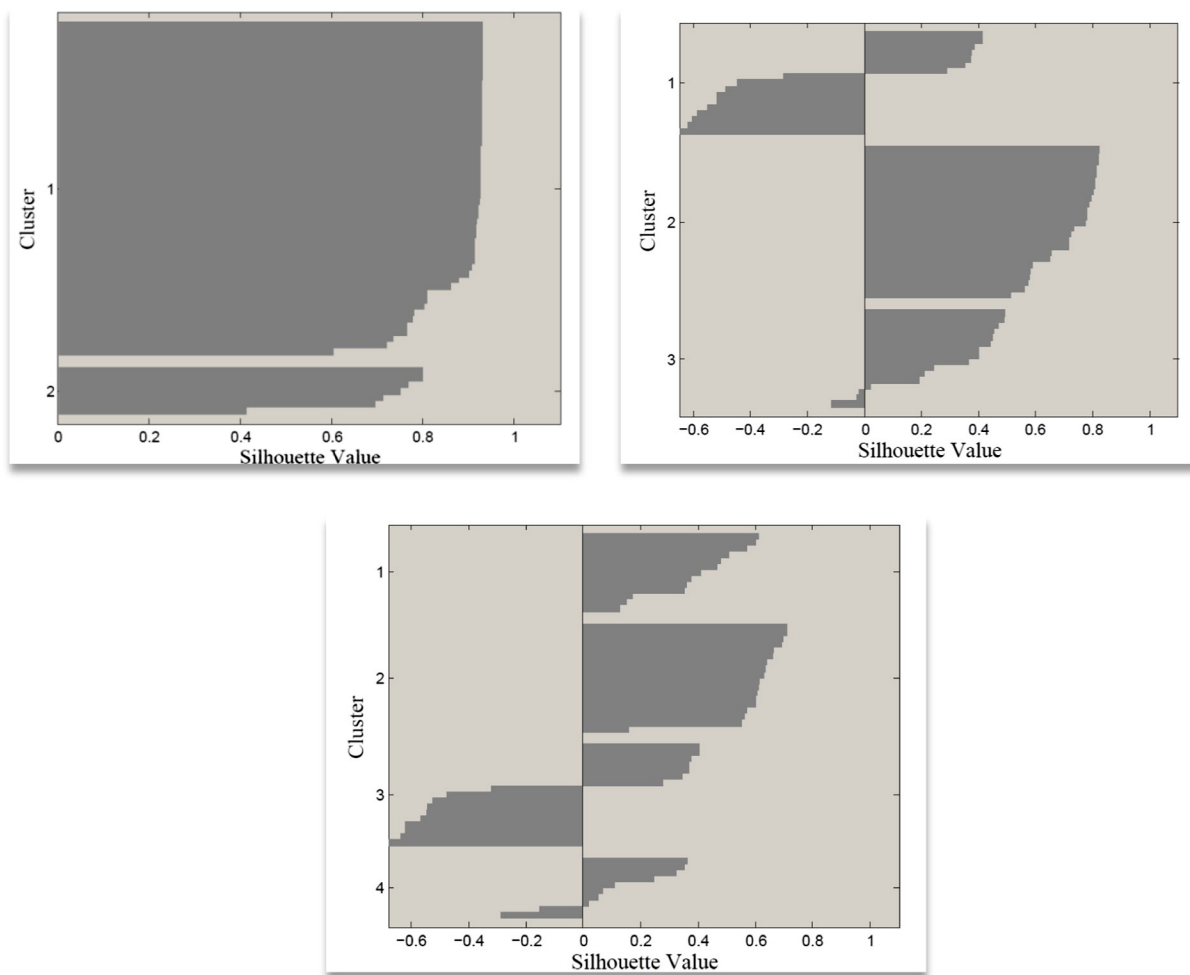
A linear function (DF), which discriminates between the compounds of both clusters, was developed assigning the values  $-1$  and  $1$  to all

**Table 2**  
Compounds of the two clusters obtained by the k-means clustering.

Cluster no.	Id. compounds	Compounds in the cluster
1	1 2 3 4 5 6 7 8 10 12 13 14 15 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 46 47 50 53 55	45
2	16 45 51 52 54 57 58	7

compounds from clusters 1 and 2, respectively. These values were used as dependent variable and the molecular descriptors calculated with Dragon software package as the independent variables. In this way, if the application of the discriminant function to a compound gives a value of  $DF < 0$ , it belongs to cluster 1, and if  $DF > 0$ , the compound belongs to cluster 2. The discriminant ability (expressed as the percentage of reproducibility) was assessed by the percentage of correct classifications attained for each cluster and for a test set of six compounds extracted from the total set.

Five different discriminant functions were analyzed using 1 to 5 independent variables which minimize the standard deviation ( $S$ ) of the function, Table 3. The five discriminant functions present an excellent capacity to discriminate the clusters and separate correctly 100% of the compounds. However, the 100% of reproducibility is achieved using only one molecular descriptor ( $piID$ ) and the DF is



**Fig. 3.** Plots of the k-means clustering for  $k$  ranging from 2 to 4. Two clusters (top left panel), three clusters (top right panel) and four clusters (bottom panel).

**Table 3**  
Results of linear discriminant analysis.

DM no.	Molecular descriptors	R	S	% Reproducibility	
				Cluster 1	Cluster 2
1	<i>piID</i>	0.971	0.163	100	100
2	<i>PCD D/Dr10</i>	0.980	0.137	100	100
3	<i>piID RDF045u G1p</i>	0.988	0.109	100	100
4	<i>nR06 PCD MWC10 BELe2</i>	0.992	0.088	100	100
5	<i>piID D/Dr10 MWC10 HOMT BELe2</i>	0.996	0.062	100	100

found to be a simple equation with one independent variable. The DF equation (Eq. (1)) is:

$$DF = -1.0959 + 2.72 \times 10^{-6} piID$$

$$R = 0.971 \quad R^2 = 0.944 \quad S = 0.163 \quad (1)$$

Eq. (1) was validated applying it to a test set (compounds Id. 6, 29, 30, 32, 47, and 57) showing a 100% reproducibility and an excellent discriminant capacity.

The topological descriptor *piID* or *piD* (conventional bond order *ID* number) is a molecular weighted path number obtained by weighting graph edges with conventional bond order [42]. This *ID* number accounts for multiple bonds in the molecule; for saturated molecules each bond weight is equal to one, therefore the *ID* number coincides with the total path count. The literature indicates that this descriptor was mainly proposed to univocally identify a molecule by a single real number, the aim being to obtain a highly discriminatory power suitable for chemical documentation [43,44].

### 3.4. Development of QSAR model

The search of a predictive QSAR model was carried out by performing 31 different combinations of training and test sets. In all cases the training sets were comprised of 80% of the total set and the test sets by 20% (ten compounds).

The first ten models were analyzed using a test set with all compounds from cluster 1 and the training set includes all compounds of cluster 2 (series 1), Table 4. The next ten models were tested with a test set including all the compounds from cluster 2 and three compounds from cluster 1, and the training set was formed only with compounds of cluster 1 (series 2). The test set used in the next ten models was built utilizing 20% of the compounds from clusters 1 and 2, and 80% of the compounds were assigned to the training set (series 3). Finally, the test set of model 31 was selected using the Kennard–Stone algorithm [45]. The results are listed in Table 4.

The average values of the statistical parameters indicate that series 3 presents the highest predictive power. In series 1, all of the compounds of cluster 2 are included in the training set and the test set only has compounds from cluster 1. This distribution causes that any sets are representative of the whole. Series 2 presents an acceptable calibration ( $R_{\text{average}} = 0.823$ ) due to all compounds of training set belonging to cluster 1. However, the validation of the models is poor because the test set with the seven compounds of cluster 2 and three of cluster 1 is not representative of the training set. The best cases were found using series 3 and the Kennard–Stone algorithm because the training and test sets are representative of both clusters. However series 3 presents better statistical parameters.

All the developed models comply with the classic semiempirical “rule of thumb”, which indicates that at least six or seven data points (i.e. compounds) should be present by descriptor [43]. According to the number of compounds of the training set ( $N = 42$ ), a linear regression model containing from 1 to 6 descriptors (selected from a total of 1497) would provide sufficient information about the relationship between the biological activity and the structure of the compounds. Series

**Table 4**

Results of QSAR models and the test set used in each one. The molecular descriptor number of model is named DM, the regression coefficient and standard deviations are identified as  $R$  and  $S$ , respectively. The train subscripts correspond to training set and the test subscripts to test set.

Series 1																
Models	DM	Test set <sup>a</sup>						$R_{\text{train}}$	$S_{\text{train}}$	$R_{\text{test}}$	$S_{\text{test}}$					
1	3	5	6	13	15	24	28	37	42	50	55	0.805	0.264	0.730	0.692	
2	4	4	7	18	19	22	30	35	40	44	50	0.807	0.299	0.772	0.448	
3	4	2	3	6	10	17	25	27	28	33	46	50	0.813	0.304	0.687	0.418
4	3	2	15	18	20	21	31	40	43	53	55	0.803	0.286	0.518	0.567	
5	4	7	10	18	24	25	26	28	30	40	46	0.803	0.280	0.785	0.491	
6	3	2	6	15	27	31	33	40	50	53	55	0.800	0.289	0.525	0.558	
7	4	10	18	24	25	26	28	30	42	43	46	0.791	0.282	0.786	0.524	
8	4	7	13	15	18	20	27	30	40	44	55	0.863	0.236	0.549	0.665	
9	4	1	8	13	14	19	33	36	40	43	53	0.803	0.315	0.639	0.380	
10	4	2	13	19	20	22	34	35	38	43	50	0.788	0.321	0.781	0.329	
		Means values						0.807	0.287	0.677	0.507					
Series 2																
Models	DM	Test set <sup>b</sup>						$R_{\text{train}}$	$S_{\text{train}}$	$R_{\text{test}}$	$S_{\text{test}}$					
11	4	5	10	<b>16</b>	<b>45</b>	<b>47</b>	<b>51</b>	<b>52</b>	<b>55</b>	<b>57</b>	<b>58</b>	0.828	0.292	0.737	0.520	
12	4	3	15	<b>16</b>	<b>40</b>	<b>45</b>	<b>51</b>	<b>52</b>	<b>55</b>	<b>57</b>	<b>58</b>	0.872	0.248	0.404	0.597	
13	5	13	<b>16</b>	30	42	<b>45</b>	<b>51</b>	<b>52</b>	<b>55</b>	<b>57</b>	<b>58</b>	0.856	0.259	0.797	0.552	
14	4	2	<b>16</b>	19	23	<b>45</b>	<b>51</b>	<b>52</b>	<b>55</b>	<b>57</b>	<b>58</b>	0.811	0.289	0.702	0.530	
15	4	<b>16</b>	36	38	<b>45</b>	46	<b>51</b>	<b>52</b>	<b>55</b>	<b>57</b>	<b>58</b>	0.821	0.294	0.653	0.478	
16	3	4	<b>5</b>	<b>16</b>	<b>45</b>	46	<b>51</b>	<b>52</b>	<b>55</b>	<b>57</b>	<b>58</b>	0.795	0.305	0.668	0.424	
17	4	1	<b>8</b>	<b>16</b>	39	<b>45</b>	<b>51</b>	<b>52</b>	<b>55</b>	<b>57</b>	<b>58</b>	0.833	0.289	0.415	0.747	
18	4	<b>16</b>	26	30	37	<b>45</b>	<b>51</b>	<b>52</b>	<b>55</b>	<b>57</b>	<b>58</b>	0.841	0.270	0.551	0.568	
19	4	5	<b>7</b>	<b>16</b>	42	<b>45</b>	<b>51</b>	<b>52</b>	<b>55</b>	<b>57</b>	<b>58</b>	0.809	0.307	0.687	0.362	
20	3	<b>16</b>	33	35	37	<b>45</b>	<b>51</b>	<b>52</b>	<b>55</b>	<b>57</b>	<b>58</b>	0.765	0.334	0.523	0.561	
		Means values						0.823	0.288	0.613	0.533					
Series 3																
Models	DM	Test set <sup>c</sup>						$R_{\text{train}}$	$S_{\text{train}}$	$R_{\text{test}}$	$S_{\text{test}}$					
21	4	1	6	23	29	30	32	41	47	<b>51</b>	<b>57</b>	0.822	0.272	0.722	0.637	
22	4	6	<b>8</b>	<b>16</b>	19	33	37	40	46	<b>52</b>		0.800	0.300	0.765	0.431	
<b>23</b>	<b>4</b>	7	10	14	18	32	38	39	41	<b>52</b>	<b>58</b>	<b>0.811</b>	<b>0.307</b>	<b>0.801</b>	<b>0.333</b>	
24	4	3	10	25	26	30	34	39	43	<b>51</b>	<b>54</b>	0.816	0.292	0.784	0.411	
25	4	2	<b>6</b>	<b>16</b>	27	28	33	37	<b>45</b>	46	50	0.801	0.304	0.777	0.454	
26	4	8	10	15	<b>16</b>	18	19	20	23	35	<b>57</b>	0.856	0.252	0.711	0.619	
27	4	3	5	8	12	<b>16</b>	37	39	41	47	<b>52</b>	0.818	0.302	0.507	0.521	
28	3	7	13	<b>16</b>	19	26	32	39	47	<b>50</b>	<b>51</b>	0.765	0.335	0.561	0.391	
29	3	1	8	12	19	20	39	40	41	<b>45</b>	<b>54</b>	0.858	0.269	0.567	0.549	
30	4	3	10	19	24	29	30	34	47	<b>54</b>	<b>57</b>	0.791	0.289	0.756	0.535	
		Means values						0.813	0.292	0.695	0.488					
Kennard–Stones																
31	3	8	20	22	23	24	25	27	28	<b>51</b>	53	0.748	0.408	0.712	0.573	

<sup>a</sup> Test set including only compounds from cluster 1.

<sup>b</sup> Test set including all the compounds from cluster 2 (in bold font) and three compounds from cluster 1.

<sup>c</sup> Test set including eight compounds from cluster 1 and two from cluster 2 (in bold font), according to Table 2.

2 has models with higher number of descriptors than the rest of the series. Therefore, the models of series 1 and 3 are simpler QSAR models with high predictive capability.

Model 23 from series 3 (highlighted in bold in Table 4) was selected as the most predictive QSAR model. This model shows high calibration and validation parameter values ( $R_{\text{train}} = 0.811$  and  $R_{\text{test}} = 0.801$ , respectively). The mathematical equation (Eq. (2)) and the statistical parameters are:

$$\log(1/IC_{50}) = -0.612 - 1.184 IC_1 + 0.143 RDF100m$$

$$+ 0.625 Mor11p + 15.901 R5e^+$$

$$R_{\text{train}} = 0.811 \quad R^2_{\text{train}} = 0.658 \quad S_{\text{train}} = 0.307 \quad R_{\text{loo}} = 0.746 \quad (2)$$

$$S_{\text{loo}} = 0.352 \quad R_{\text{test}} = 0.801 \quad R^2_{\text{test}} = 0.643 \quad S_{\text{test}} = 0.333$$

$$R_{120\%} = 0.645 \quad S_{120\%} = 0.441 \quad S_{\text{rand}} = 0.400$$

The validation of the selected model was carried out through four different methods: leave-one-out, leave-more-out, employing a test set and y-randomization. The regression coefficients  $R_{100}$ ,  $R_{1\%0}$  and  $R_{\text{test}}$  exceed the accepted value of 0.50. In addition, the smallest  $S_{\text{rand}}$  value ( $S_{\text{rand}} = 0.400$ ) achieved through the analysis of 10,000 cases of y-randomization was greater than the value found ( $S = 0.307$ ) when true calibration was considered, showing that the developed QSAR model is predictive [44].

The values of  $\log(1/IC_{50})$  predicted by Eq. (2) are listed in Table 1 and are graphically depicted in Fig. 4. The plot in the top panel of this figure shows the good correlation that exists between the predicted and the experimental activity values of the compounds of training and test sets. The plot in the bottom panel represents the quality of the leave-one-out validation.

### 3.5. Molecular descriptors and the 5-LOX inhibitory activity

We performed an exhaustive search of a QSAR model to predict the 5-LOX inhibitory activity, expressed as  $\log(1/IC_{50})$ , of a series of 5-Benzylidene-2-phenylthiazolinones. According to the obtained model (model number 23 of Table 4), the most relevant molecular descriptors related to the 5-LOX inhibitory activity are IC1, RDF100m, Mor11p and R5e+. The correlation matrix, given in the Table 5, indicates that the information provided by each descriptor to the model is not redundant, showing a maximum correlation of 0.476 between IC1 and R5e+. A brief description of the four descriptors is shown in Table 6.

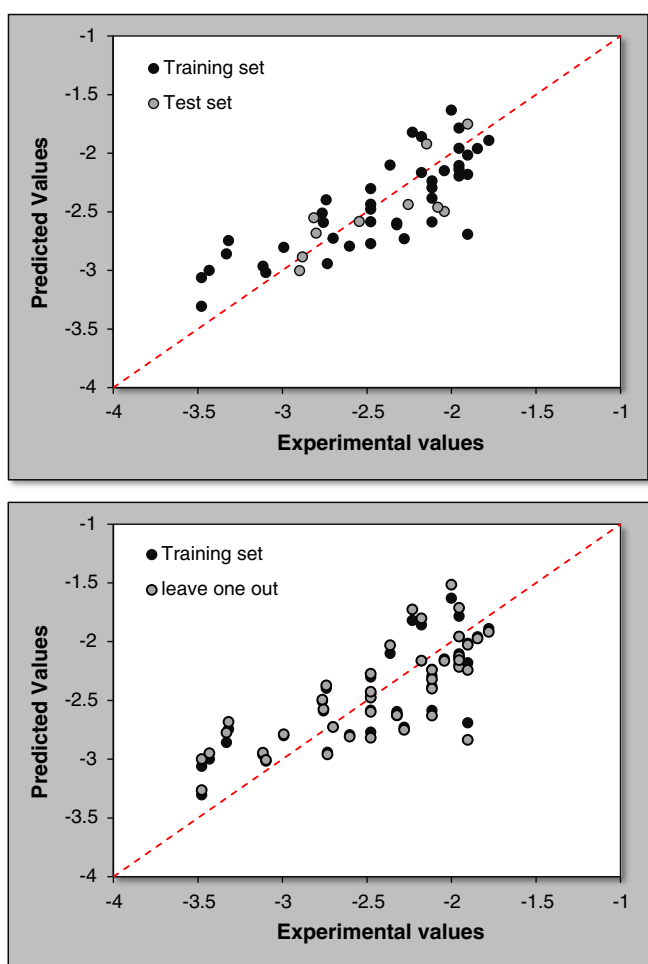


Fig. 4. Top: Experimental  $\log(1/IC_{50})$  versus predicted  $\log(1/IC_{50})$ . Bottom: Experimental  $\log(1/IC_{50})$  versus predicted  $\log(1/IC_{50})$  for the leave one out validation.

Table 5  
Correlation matrix of the molecular descriptor of model 23, Eq. (2).

	IC1	RDF100m	Mor11p	R5e+
IC1	1.000	0.172	0.225	0.476
RDF100m	0.172	1.000	0.385	0.113
Mor11p	0.225	0.385	1.000	0.475
R5e+	0.476	0.113	0.475	1.000

The standardization of the regression coefficients of Eq. (2), allows assigning a greater importance to the molecular descriptors with larger absolute standardized coefficient values [46]. The most important descriptor in the selected model is IC1, which is an Information Content descriptor. The negative sign in Eq. (2) indicates that  $\log(1/IC_{50})$  values are indirectly related to this descriptor. The second most important descriptors are R5e+ and RDF100m, which are an R index weighted by atomic Sanderson electronegativity and radial distribution function (RDF), respectively. The weighting of the R indexes encode information about substituents differently from unweighted indexes. In this case, the largest values of this descriptor can be expected when high electronegative atoms are situated far from the center of the molecule at a topological distance of 5 bonds. The radial distribution function (RDF) is a kind of molecular descriptor defined for an ensemble of atoms, and may be interpreted as the probability distribution for finding an atom in a spherical volume of certain radius, incorporating different types of atomic properties in order to differentiate the nature and contribution of atoms to the property being modeled. This descriptor also reveals an enthalpic contribution on activity (related to the interactions of hydrogen bond and van der Waals types) and it is important for hydrophobic interactions with an enzyme [47,48]. Since these descriptors have a positive contribution to the model, it is expected that the  $\log(1/IC_{50})$  increases with the increase of their values. The Mor11p descriptor has less influence on the activity. This descriptor belongs to the 3D-MorSE descriptors and is weighted by atomic polarizabilities.

## 4. Conclusion

In this work, we have developed a QSAR model with high predictive capacity which can be used to predict the 5-LOX inhibition activity of new 5-Benzylidene-2-phenylthiazolinones derivatives. We have found that biological activity is related to the structural information provided by IC1, RDF100m, Mor11p and R5e+ molecular descriptors. In the search of this predictive model, we have used the k-means clustering and LDA approach to find the possible clusters and perform the most representative selection of training and test sets. The statistical parameter of the model ( $R_{\text{train}} = 0.811$ ,  $R_{\text{test}} = 0.801$ ) shows the great stability that exists between the results obtained in the calibration and the validation when a rational selection of the training and test sets is performed, proving that the random selection is sometimes inappropriate.

We consider that the information provided in this article can be used to develop the most representative QSAR model and for future investigations and development of new potential 5-LOX inhibitors.

Table 6  
Description of the molecular descriptors of the QSAR model.

Name	Description	Block
IC1	Information content index (neighborhood symmetry of 1-order)	Information indices
RDF100m	Radial distribution function - 100/weighted by mass	RDF descriptors
Mor11p	Signal 11 / weighted by polarizability	3D-MorSE descriptors
R5e+	R maximal autocorrelation of lag 5/weighted by Sanderson electronegativity	GETAWAY descriptors

## References

- [1] K. Goodarzi, M. Goodarzi, A.M. Tager, A.D. Luster, U.H. von Andrian, Leukotriene B<sub>4</sub> and BLT1 control cytotoxic effector T cell recruitment to inflamed tissues, *Nat. Immunol.* 4 (2003) 965–973.
- [2] P.J. Barnes, Future Advances in COPD Therapy, *Respiration* 68 (2001) 441–450.
- [3] M. Back, D.X. Bu, R. Branstrom, Y. Sheikine, Z.Q. Yan, G.K. Hansson, Leukotriene B<sub>4</sub> signaling through NF-kappaB-dependent BLT1 receptors on vascular smooth muscle cells in atherosclerosis and intimal hiperplasia, *Proc. Natl. Acad. Sci. U. S. A.* 102 (2005) 17501–17506.
- [4] V. Sandanayaka, B. Mamat, R.K. Mishra, J. Winger, M. Krohn, L.M. Zhou, M. Keyvan, L. Enache, D. Sullins, E. Onua, J. Zhang, G. Halldorsdottir, H. Sighthorsdottir, A. Thorlaksdottir, G. Sighthorsson, M. Thorsteinnsdottir, D.R. Davies, L.J. Stewart, D.E. Zembower, T. Andersson, A.S. Kiselyov, J. Singh, M.E. Gurney, Discovery of 4-[(2S)-2-[(4-(4-chlorophenoxy)phenoxy)methyl]-1-pyrrolidinyl]butanoic acid (DG-051) as a novel leukotriene A<sub>4</sub> hydrolase inhibitor of leukotriene B<sub>4</sub> biosynthesis, *J. Med. Chem.* 53 (2010) 573–585.
- [5] X. Chen, S. Wang, N. Wu, C.S. Yang, Leukotriene A<sub>4</sub> hydrolase as a target for cancer prevention and therapy, *Curr. Cancer Drug Targets* 4 (2004) 267–283.
- [6] Y. Kawasaki, M. Tanji, K. Takano, Y. Fukuda, M. Isome, R. Nozawa, H. Suzuki, M. Hosoya, The leukotriene B<sub>4</sub> receptor antagonist ONO-4057 inhibits mesangioproliferative changes in anti-Thy-1 nephritis, *Nephrol. Dial. Transplant.* 20 (2005) 2697–2703.
- [7] T. Uz, N. Dimitrijevic, M. Imbesi, H. Manev, R. Manev, Effects of MK-886, a 5-lipoxygenase activating protein (FLAP) inhibitor, and 5-lipoxygenase deficiency on the forced swimming behavior of mice, *Neurosci. Lett.* 436 (2008) 269–272.
- [8] S.E. Wenzel, A.K. Kamada, Zileuton: the first 5-lipoxygenase inhibitor for the treatment of asthma, *Ann. Pharmacother.* 30 (1996) 858–864.
- [9] S. Barzen, C.B. Rödl, A. Lill, D. Steinhilber, H. Stark, B. Hofmann, Synthesis and biological evaluation of a class of 5-benzylidene-2-phenyl-thiazolinones as potent 5-lipoxygenase inhibitors, *Bioorg. Med. Chem.* 20 (2012) 3575–3583.
- [10] B. Hofmann, S. Barzen, C.B. Rödl, A. Kiehl, J. Borig, A. Zivkovic, H. Stark, G. Schneider, D. Steinhilber, A Class of 5-Benzylidene-2-phenylthiazolinones with high potency as direct 5-lipoxygenase inhibitors, *J. Med. Chem.* 54 (2011) 1943–1947.
- [11] G. Yu, P.N. Praveen Rao, M.A. Chowdhury, K.R.A. Abdellatif, Y. Dong, D. Das, C.A. Velázquez, M.R. Suresh, E.E. Knaus, Synthesis and biological evaluation of N-difluoromethyl-1,2-dihydropyrid-2-one acetic acid regioisomers: dual inhibitors of cyclooxygenases and 5-lipoxygenase, *Bioorg. Med. Chem. Lett.* 20 (2010) 2168–2173.
- [12] J. Suh, E.K. Yum, H.G. Cheon, Y.S. Cho, Synthesis and biological evaluation of N-aryl-4-aryl-1,3-thiazole-2-Amine derivatives as direct 5-lipoxygenase inhibitors, *Chem. Biol. Drug Des.* 80 (2012) 89–98.
- [13] A. Peduto, F. Bruno, F. Dehm, V. Krauth, P. De Caprariis, C. Weinigel, D. Barz, A. Massa, M. De Rosa, O. Wertz, R. Filosa, Further studies on ethyl 5-hydroxy-indole-3-carboxylate scaffold: design, synthesis and evaluation of 2-phenylthiomethyl-indole derivatives as efficient inhibitors of human 5-lipoxygenase, *Eur. J. Med. Chem.* 81 (2014) 492–498.
- [14] E. Shang, Y. Wu, P. Liu, Y. Liu, W. Zhu, X. Deng, C. He, S. He, C. Li, L. Lai, Benzol[d]isothiazole 1,1-dioxide derivatives as dual functional inhibitors of 5-lipoxygenase and microsomal prostaglandin E<sub>2</sub> synthase-1, *Bioorg. Med. Chem. Lett.* 24 (2014) 2764–2767.
- [15] B. Jiang, X. Huang, H. Yao, J. Jiang, X. Wu, S. Jiang, Q. Wang, T. Lu, J. Xu, Discovery of potential anti-inflammatory drugs: diaryl-1,2,4-triazoles bearing N-hydroxyurea moiety as dual inhibitors of cyclooxygenase-2 and 5-lipoxygenase, *Org. Biomol. Chem.* 12 (2014) 2114–2127.
- [16] E. Buscató, J.M. Wisniewska, C.B. Rödl, A. Brüggerhoff, A. Kaiser, F. Rörsch, E. Kostewicz, M. Wurglics, M. Schubert-Zsilavecz, S. Grösch, D. Steinhilber, B. Hofmann, E. Proschak, Structure–activity relationship and in vitro pharmacological evaluation of imidazo[1,2-a]pyridine-based inhibitors of 5-LO, *Future Med. Chem.* 5 (2013) 865–880.
- [17] M.C. Sharma, Molecular modeling studies of substituted 3,4-dihydroxycalcone derivatives as 5-lipoxygenase and cyclooxygenase inhibitors, *Med. Chem. Res.* 23 (2014) 1797–1818.
- [18] M.C. Sharma, S. Sharma, P. Sharma, A. Kumar, Molecular modeling and pharmacophore approach for structural requirements of some 2-substituted-1-naphthols derivatives as potent 5-lipoxygenase inhibitors, *Med. Chem. Res.* 22 (2013) 5390–5407.
- [19] B. Niu, Q. Su, X. Yuan, W. Lu, J. Ding, QSAR study on 5-lipoxygenase inhibitors based on support vector machine, *Med. Chem.* 8 (2012) 1108–1116.
- [20] G. Eren, A. MacChiarulo, E. Banoglu, From molecular docking to 3D-quantitative structure–activity relationships (3D-QSAR): insights into the binding mode of 5-lipoxygenase inhibitors, *Mol. Inf.* 31 (2012) 123–134.
- [21] J. Zheng, G. Xiao, J. Guo, Y. Zheng, H. Gao, S. Zhao, K. Zhang, P. Sun, Exploring QSARs for 5-lipoxygenase (5-LO) inhibitory activity of 2-substituted 5-hydroxyindole-3-carboxylates by CoMFA and CoMSIA, *Chem. Biol. Drug Des.* 78 (2011) 314–321.
- [22] B.K. Sharma, P. Pilania, P. Singh, QSAR study on 5-lipoxygenase activating protein (FLAP) inhibitors: the derivatives of 2,2-bisaryl-bicycloheptane, *Lett. Drug Des. Discov.* 8 (2011) 32–43.
- [23] P. Aparoy, G.K. Suresh, K. Kumar Reddy, P. Reddanna, CoMFA and CoMSIA studies on 5-hydroxyindole-3-carboxylate derivatives as 5-lipoxygenase inhibitors: generation of homology model and docking studies, *Bioorg. Med. Chem. Lett.* 21 (2011) 456–462.
- [24] I.G. Tsygankova, Variable selection in QSAR models for drug design, *Curr. Comput. Aided Drug Des.* 4 (2008) 132–142.
- [25] B.S. Everitt, S. Landau, M. Leese, Cluster Analysis, Edward Arnold, London, 2001.
- [26] VCCLAB, Virtual Computational Chemistry Laboratory E-Dragon, Version 1.0, <http://michem.disat.unimib.it/chm2005>.
- [27] J.B. MacQueen, Some Methods for Classification and Analysis of Multivariate Observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, 1, Univ. of California Press, 1967, pp. 281–297.
- [28] A. Malek-Khatibi, M. Kompany-Zareh, S. Gholami, S. Bagheri, Replacement based non-linear data reduction in radial basis function networks QSAR modeling, *Chemom. Intell. Lab. Syst.* 135 (2014) 157–165.
- [29] S. Sardari, H. Kohanzad, G. Ghavami, Artificial neural network modeling of antimycobacterial chemical space to introduce efficient descriptors employed for drug design, *Chemom. Intell. Lab. Syst.* 130 (2014) 151–158.
- [30] E. Pourbasheer, R. Aalizadeh, M.R. Ganjali, P. Norouzi, QSAR study of IKKβ inhibitors by the genetic algorithm: multiple linear regressions, *Med. Chem. Res.* 23 (2014) 57–66.
- [31] Matlab 7.0, The MathWorks Inc., 2004
- [32] T. Collus, Discriminant Analysis and Applications, Academic Press, New York, 1973.
- [33] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, New Jersey, 1988.
- [34] N. Mahmoudi, J.V. de Julián-Ortiz, L. Ciceron, J. Gálvez, D. Mazier, M. Danis, F. Derouin, R. García-Domenech, Identification of new antimalarial drugs by linear discriminant analysis and topological virtual screening, *J. Antimicrob. Chemother.* 57 (2006) 489–497.
- [35] P.R. Duchowicz, E.A. Castro, F.M. Fernández, Alternative Algorithm for the Search of an Optimal Set of Descriptors in QSAR-QSPR Studies, *MATCH Commun. Math. Comput. Chem.* 55 (2006) 179–192.
- [36] P.B. Paz, E.G. Vega-Hissi, M.F. Andrada, M.R. Estrada, J.C. Garro Martinez, Quantitative structure activity relationship and binding investigation of N-alkyl glycine amides as inhibitors of leukotriene A<sub>4</sub> hydrolase, *Med. Chem. Res.* DOI <http://dx.doi.org/10.1007/s00044-014-1121>.
- [37] J.C. Garro Martinez, E.G. Vega-Hissi, M.F. Andrada, P.R. Duchowicz, F. Torrens, M.R. Estrada, Lacosamide derivatives with anticonvulsant activity as carbonic anhydrase inhibitors. molecular modeling, docking and QSAR analysis, *Curr. Comput. Aided Drug Des.* 10 (2014) 160–167.
- [38] A.G. Mercader, P.R. Duchowicz, F.M. Fernandez, E.A. Castro, Replacement method and enhanced replacement method versus the genetic algorithm approach for the selection of molecular descriptors in QSPR/QSAR theories, *J. Chem. Inf. Model.* 50 (2010) 1542–1548.
- [39] A.G. Mercader, P.R. Duchowicz, F.M. Fernández, E.A. Castro, Advances in the replacement and enhanced replacement method in QSAR and QSPR theories, *J. Chem. Inf. Model.* 51 (2011) 1575–1581.
- [40] J. Verma, V.M. Khedkar, E.C. Coutinho, 3D-QSAR in drug design—a review, *Curr. Top. Med. Chem.* 10 (2010) 95–115.
- [41] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.* 20 (1987) 53–65.
- [42] M. Randic, P.C. Jurs, On a fragment approach to structure–activity correlations, *Quant. Struct. Act. Relat.* 8 (1989) 39–48.
- [43] C. Hansch, Comprehensive Drug Design, Pergamon Press, New York, 1990.
- [44] S. Wold, L. Eriksson, Chemometrics Methods in Molecular Design, Wiley-VCH, Weinheim, Germany, 1995.
- [45] R.W. Kennard, L.A. Stone, Computer aided design of experiments, *Technometrics* 11 (1969) 137–148.
- [46] N.R. Draper, H. Smith, Applied Regression Analysis, John Wiley & Sons, New York, 1981.
- [47] A. Jain, S.C. Chaturvedi, R. Sharma, Structural insight for benzimidazole as angiotensin II AT1 receptor antagonist by using molecular property and biological activity correlation: QSAR approach, *Int. J. Pharm. Pharm. Sci.* 3 (2011) 541–546.
- [48] E. Vicente, P.R. Duchowicz, D. Benitez, E.A. Castro, H. Cerecetto, M. González, A. Monge, Anti-T cruzi activities and QSAR studies of 3-arylquinoxaline-2-carbonitrile di-N-oxides, *Bioorg. Med. Chem. Lett.* 20 (2010) 4831–4835.