European Journal of Medicinal Chemistry 45 (2010) 2753–2760

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

QSAR analysis of diaryl COX-2 inhibitors: Comparison of feature selection and train-test data selection methods

Somaieh Soltani [a], Hoda Abolhasani [b], Afshin Zarghi [c], Abolghasem Jouyban [d,*]

[a] Gifted and Talented Students Center, Tabriz University of Medical Sciences, Tabriz, Iran
[b] Liver and Gastrointestinal Diseases Research Center, Tabriz University of Medical Sciences, Tabriz 51664, Iran
[c] Faculty of Pharmacy, Shahid Beheshti University (M.C), Tehran, Iran
[d] Faculty of Pharmacy and Drug Applied Research Center, Tabriz University of Medical Sciences, Tabriz, Iran

ARTICLE INFO

Article history:
Received 6 April 2009
Received in revised form
18 February 2010
Accepted 23 February 2010
Available online 1 March 2010

Keywords:
COX-2 selective inhibition
QSAR
Feature selection
Test-train selection

ABSTRACT

QSAR analyses were performed on a series of trans-stilbenoid diaryl compounds for modeling their COX-2 inhibitory activities. The multivariate regression equations were developed with the selected independent variables using various feature selection methods. In addition, model training was done using different test-train data selection methods. The applicability of each variable and the test-train selection methods was investigated through the type and number of the selected significant descriptors as well as the statistical criteria of the developed model for each pair of feature and test-train selection methods. The goodness of fit and the statistical significance of 15 developed equations were evaluated using the correlation coefficient (R), the variance ratio (F), and the standard error of estimate (S.E.). The models were validated using the leave many out and the leave one out cross-validation methods. The mean percentage deviation (MPD (±SD)) was used as an accuracy criterion for checking the predicted activities. It was found that the developed models could predict the COX-2 and COX-1 inhibitory activities as well as the COX-2/COX-1 selectivity ratios producing the MPD values of 1.6 (± 0.8)%, 7.7 (± 5.6)%, and 16.9 (± 9.6)%, respectively.

1. Introduction

Non-steroidal anti-inflammatory drugs (NSAIDs) act as anti-inflammatory agents through the inhibition of cyclooxygenase (COX) which catalyzes the conversion of arachidonic acid (AA) to prostaglandins (PGs). Cyclooxygenase is known to occur in at least two isoforms: COX-1 which is a constitutive enzyme responsible for the maintenance of physiologic homeostasis, and COX-2 which is an inducible isoform that leads to inflammation. This discovery has led to the theory that the inhibition of COX-1 causes some side effects of NSAIDs such as gastric ulceration, bleeding, and renal function suppression, whereas the inhibition of COX-2 accounts for the therapeutic effects of NSAIDs [1].

All classical NSAIDs, such as aspirin, ibuprofen, and indomethacin, are capable of inhibiting both COX-1 and COX-2; however, they are found to bind more tightly to COX-1. As a result, these drugs are associated with a high risk of gastrointestinal effects as well as adverse effects of the renal function. Owing to these problems, researchers have made attempts to increase the gastrointestinal safety of new entities by increasing the selectivity of the COX-2, thereby reducing the COX-1 inhibitory effect. The major finding in this area is attributed to the initial evidences of anti-inflammatory effects without ulcerogenic effects in DuP-697 (a diaryl heterocyclic). Subsequently, a number of selective COX-2 inhibitors with proven therapeutic utility for the treatment of inflammation such as celecoxib, rofecoxib, valdecoxib, and etoricoxib, have been developed [2].

However, the use of specific COX-2 selective drugs was a failure owing to their cardiovascular side effects that occurred because of the inhibition of prostaglandins secretion which is required for the normal functioning of the cardiovascular system and is produced by COX-1 isoform [3].

In search for selective COX-2 inhibitors using conventional drug development and synthesis methods, the concept of QSAR was exploited in modifying conventionally available NSAIDs in the hope of reducing the development time and cost. It is now well known that the QSAR modeling method is capable of identifying failures in the early stage of drug development, which helps to resume the sources.

The majority of selective COX-2 inhibitors belong to a class of tricyclic compounds possessing 1,2-diaryle substitution on a central heterocyclic, or carbocyclic ring. Recently, a number of compounds that occur in nature [1] and are synthetic trans stilbenoid [4, 5]

* Corresponding author. Tel.: +0098 411 3379323; fax: +0098 411 3363231.
E-mail addresses: Somaieh.s@gmail.com (S. Soltani), residenthoda@gmail.com (H. Abolhasani), azarghi@yahoo.com (A. Zarghi), ajouyban@hotmail.com (A. Jouyban).

0223-5234/$ – see front matter © 2010 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2010.02.055

were evaluated as COX-2 selective inhibitors and their structure – activity relationships were investigated [6].

Since the QSAR analysis of these newly developed drug candidates is abandoned, there is only a study of resveratrol analogues [1]; also, there has been no further study of the inhibitory activity or the selectivity of COX-2. Such studies might help in the design and synthesis of better selective COX-2/COX-1 inhibitors with reduced side effects. Thus, the aim of the present study is to present the QSAR of these compounds.

The purpose of developing a QSAR model is to reduce the cost of the target designing by modifying the molecular structures for achieving the desired molecule with the proposed property, without experimental measurement [7]. Subsequently, an ideal QSAR model should be capable of accurately predicting the desired property of a newly synthesized or a hypothetical molecule. The main steps for the development of a QSAR model could be summarized as: data preparation, data analysis, and model validation. Firstly, to obtain a valid model with high predictive ability, the data used for developing the model (training data set) should be carefully selected to cover all spaces of the row data set, and subsequently, the relevant descriptors should be selected using the appropriate descriptor selection method which can identify relevant variables. In this study, three frequently used test-train selection methods were investigated along with five feature selection procedures to determine the most appropriate data splitting and descriptor selection methods. The developed models were used for predicting the COX-2, COX-1 and COX-2/COX-1 inhibitory activity of resveratrol analogues.

2. QSAR model development

2.1. Data preparation

2.1.1. Experimental data

Four data sets of COX-2 selective inhibitors (a total of 54 data points) were collected from the literature [1, 4, 5] (Table 1). The selected structures were diaryl entities that were linked together through a diatomic (C=C, N=N, C=N, N=C) linker (Fig. 1). The experimental IC_{50} values (50% inhibitory concentration of the enzyme) were evaluated in a cellular assay using human whole blood. The enzyme inhibition data were converted to negative logarithmic value (concentration expressed in mole per liter) and subsequently used as the response variable for QSAR analyses. In order to minimize the inter laboratory differences between evaluations, all pIC_{50} values were divided to celecoxib's pIC_{50} value which was evaluated and reported for each data set. The normalized experimental values (Y_1 for COX-1 inhibition, Y_2 for COX-2 inhibition, and Y_s for the ratio Y_2/Y_1) were then used for further analyses. In order to identify the outliers we used two different methods (i.e. PCA map and standard scores). There is a general agreement that there is not a single approach to identify all outliers, then we tried to use two approaches one in descriptor space (PCA plot of scores) and the other in response space (standard score) prior to the numerical analysis. The PCA plot of scores (Fig. 2) identified two outliers (compounds 11w, 12w). Further investigations of these two molecules showed that they have very high molecular weights (531.59 and 615.77, respectively) comparing with the mean value of all data (296.02). Also the logP values of these molecules are significantly more than the mean logP of other molecules in the data set. These characteristics are because of the large functional groups on R_2 and R_3 situations (3,4-($OCONHC_6H_5$)$_2$ and 3,4-($OCONHC_6H_4$-*p*-*i*-C_3H_7)$_2$ respectively for 11w and 12w) which also could lead to different mechanism of interaction with the receptor. Because of these characteristics and low potency (high drug concentrations) which could be a result of their chemical characteristics, these two molecules excluded as outliers from the analysis. In order to identify the outliers in response space we conducted a standard score analysis and the results showed that there are four molecules (Res 9–12) with standard score near or more than 2.5 which could be expected as outlier (Fig. 3). The further studies showed that these molecules are very potent molecules with low concentrations (at least ten

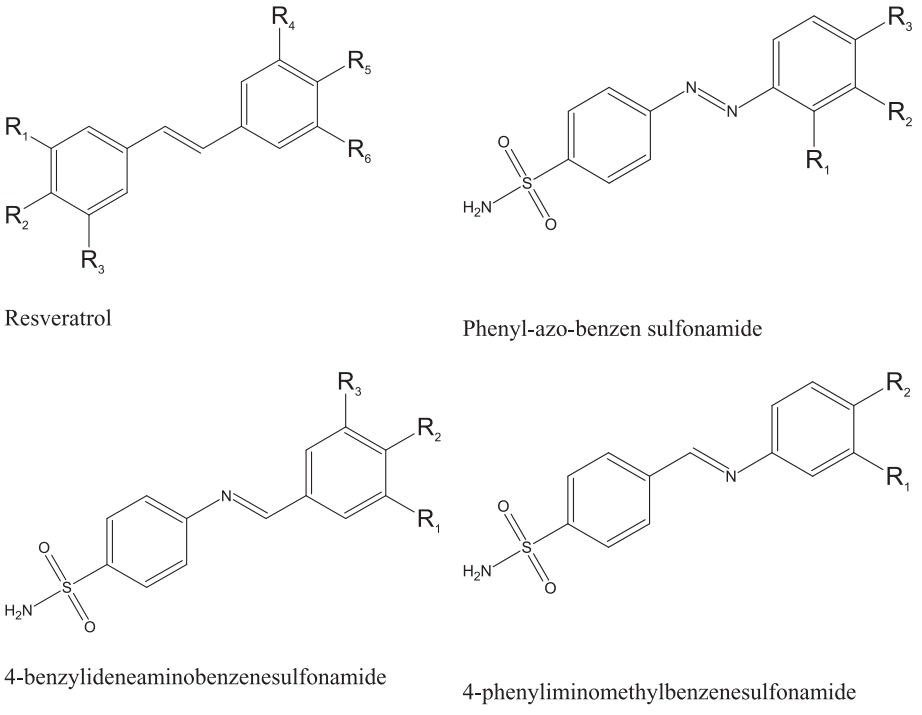

Fig. 1. Structures of selected stilbenoid diaryl compounds.

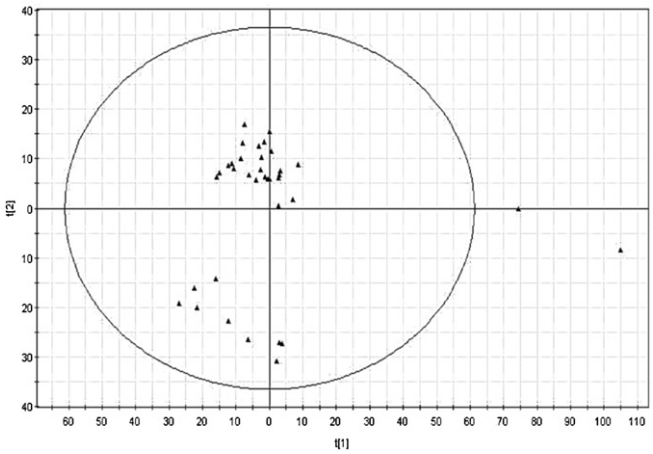

Fig. 2. PCA scores plot based on calculated variables where the first two components were autofitted. Outliers are outside circle.

times less than the mean concentration of other molecules) and we decided to omit these four molecules as outliers too.

2.1.2. Molecular descriptors

The 2D structures of all molecules were drawn and converted to 3D structures using the Hyper-Chem 7 software. The model built and the molecular mechanics energy minimized molecules were used as inputs for the Dragon 5.4 software. The software was used for calculating 20 subsets of molecular descriptors including: 2D autocorrelation, 3D-Morse descriptors, Atom-centered fragments, Burden eigenvalues, Connectivity indices, Constitutional descriptors, Edge adjacency indices, Eigenvalue-based indices, Functional group counts, Geometrical descriptors, GETAWAY descriptors, Information indices, Molecular properties, Randic molecular profiles, RDF descriptors, Topological charge indices, Topological descriptors, Walk and path counts, and WHIM descriptors. The structural parameters calculated after discarding the constant and near constant values (1217 descriptors) were saved and further analyzed using the SPSS and MATLAB software.

2.2. Data analysis

2.2.1. Selection of the training and test sets

The training set plays an important role in developing the properties of the model since the more similar molecules for training the model, the more accurate are the expected results. Thus, the selection of the training and test sets is one of the most important steps in QSAR model development. An ideal division of a training and test set will lead to data sets with the following criteria: (i) similarity of all representative compounds of the test set in multidimensional descriptor space to the training set; (ii) similarity of all representative compounds of the training set to the test set; and (iii) distribution of the training set representative points within the whole area occupied by the entire data set [8]. In other words, an ideal splitting leads to a test set in which each of its members is close to at least one member of the training set [7]. Several attempts were made to develop rational approaches for selecting the training and test data sets. The frequently used methods are straightforward random selection through activity sampling, systematic clustering techniques such as K means algorithm and hierarchical clustering, self organizing maps such as Kohonen's maps, formal statistical experimental design such as fractional and factorial designs, and sphere exclusion algorithms [7]. Among these, the activity sampling and clustering techniques are the most frequently used methods. In this study, both methods were used and the results were compared to determine their possible advantages and limitations (Table 2).

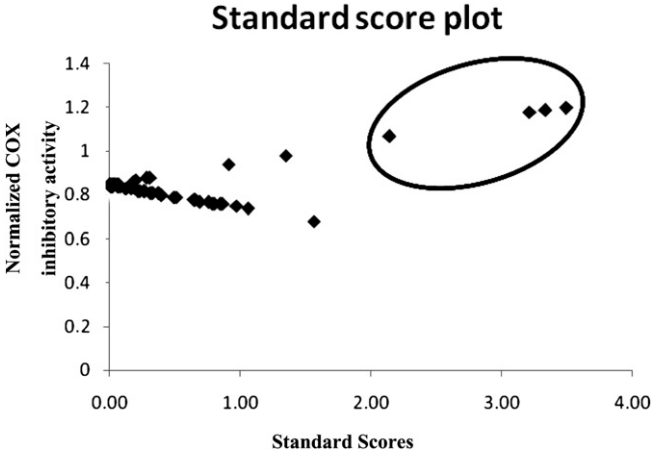

Fig. 3. Standard score plot of normalized COX inhibitory activity of the investigated compounds. Outliers are in the circle.

2.2.1.1. Activity sampling (AS). Activity sampling or activity splitting is one of the most frequently used methods for selecting test and training data sets. In this method, the activity or property to be modeled is used to split the data set into bins and the test and training data sets are selected randomly from each bin. In this study, the data was sorted owing to the inhibitory activity (pIC_{50}), and then the data set was randomly divided into training ($^3/_4$ data points) and test (¼ data points) sets (Table 2).

2.2.1.2. K-means clustering (KMC). K means clustering is the other most frequently used method of data set splitting. This procedure identifies relatively homogeneous groups of molecules (each observation has the nearest distance to the mean of cluster) based on selected properties (biologic activities (AKMC), structural variables (VKMC), or both (AVKMC)). In this study, the VKMC and AVKMC methods were used. It is obvious that the AVKMC method uses all the available information for clustering. The clusters were divided into training ($^3/_4$ data points) and test (¼ data points) sets. The details of the test and training data sets are summarized in Table 2.

2.2.2. Descriptor selection

Various feature selection methods have been used for selecting reliable descriptors in QSAR studies. In this study, five frequently used methods were used to check their ability for selecting relevant descriptors. The training sets selected using data splitting methods were used for selecting descriptors by employing five different methods as explained below.

2.2.2.1. Stepwise regression (SR). In this method, all variables that passed the tolerance criterion of 0.05 were entered in a single equation, regardless of the stepwise entry method. However, a variable was not entered if it would cause the tolerance of another variable already present in the model to drop below the tolerance criterion. The significance values are based on fitting the experimental data to a single model. Therefore, they are generally invalid and some excluded variables are relevant and valid parameters.

2.2.2.2. Variable blocked (200 variables in each block) stepwise regression (VBSR). All independent variables that were selected in the stepwise method were added to a single regression model. However, different subsets of variables could be specified as entry

Table 1
Details of studied structures.

Data set	SN[a]	N[b]	IC_{50} range (μM)	IC_{50} (μM)[c]	Reference
Resveratrol	1	14	0.00104–2.21	0.034	[1]
4-benzylideneaminobenzenesulfonamide	2	21	0.74–6.75	0.30	[4]
4-phenyliminomethylbenzenesulfonamide	3	10	1.95–6.54	0.30	[4]
Phenylazobenzen sulfonamide	4	9	2.04–16.85	0.33	[5]

[a] Set number.
[b] Number of data points.
[c] IC_{50} value of celecoxib.

Table 2
Details of data sets and test-train data points.

SN[a]	Molecule ID	IC_{50} (COX-1)	IC_{50} (COX-2)	Y_1	Y_2	AS	VKMC	AVKMC
4	3w	59.96	12.42	0.90	0.76	2	1	1
4	4w	N.D[b]	16.85	*	0.74	1	1	2
4	6w	N.D	14.63	*	0.75	1	2	1
4	7w	61.73	4.28	0.90	0.83	2	1	2
4	8w	23.28	2.04	0.99	0.88	1	1	1
4	9w	31.57	8.89	**0.96**[c]	0.78	1	1	1
4	10w	33.32	11.34	**0.95**	0.76	1	1	1
4	11w	35.07	11.07	1.75	0.76	Outlier	Outlier	Outlier
4	12w	26.26	12.37	1.31	0.76	Outlier	Outlier	Outlier
2	benz6	108.34	2.87	0.86	0.85	1	2	1
2	benz7	159.07	2.22	0.82	0.87	2	1	2
2	benz8	183.50	2.73	0.81	0.85	1	1	1
2	benz9	141.25	3.00	0.83	0.85	1	1	1
2	benz10	105.63	6.75	**0.86**	0.79	1	1	1
2	benz11	195.74	3.36	0.80	0.84	1	1	1
2	benz12	148.20	3.42	0.83	0.84	2	1	1
2	benz13	182.19	4.60	0.81	0.82	2	1	2
2	benz14	190.47	4.94	0.80	0.81	1	1	2
2	benz15	313.83	9.88	0.76	0.77	1	1	2
2	benz16	383.36	5.45	0.74	0.81	1	1	1
2	benz17	38.20	2.78	0.95	0.85	2	1	1
2	benz18	23.15	2.85	1.00	0.85	1	1	2
2	benz19	78.20	2.95	0.89	0.85	2	1	2
2	benz20	85.13	0.74	0.88	0.94	1	1	2
2	benz21	47.91	3.69	**0.93**	0.83	1	1	1
2	benz22	43.80	3.50	0.94	0.84	1	2	1
2	benz23	110.27	3.09	0.85	0.84	1	2	2
2	benz24	109.69	3.40	0.86	0.84	1	1	1
2	benz25	167.09	2.93	0.82	0.85	1	1	1
2	benz26	155.95	2.71	0.82	0.85	1	2	1
3	phi27	63.59	3.11	0.91	0.84	1	1	1
3	phi28	80.20	4.38	**0.88**	0.82	1	1	1
3	phi29	63.22	4.62	0.91	0.82	1	2	1
3	phi30	43.89	6.54	0.94	0.79	1	2	1
3	phi31	64.42	1.95	0.91	0.88	1	2	1
3	phi32	51.83	5.09	0.93	0.81	2	1	2
3	phi33	60.74	4.14	0.91	0.83	1	1	1
3	phi34	31.27	4.28	0.97	0.82	1	1	1
3	phi35	23.99	3.13	**1.00**	0.84	2	1	1
3	phi36	56.73	3.72	0.92	0.83	1	2	1
1	res1	1.23	1.67	**1.25**	0.77	2	1	1
1	res2	9.10	7.80	1.07	0.68	1	1	1
1	res3	27.78	1.58	0.97	0.78	1	2	1
1	res4	2.83	0.80	**1.18**	0.82	1	1	1
1	res5	7.25	0.51	1.09	0.84	1	1	1
1	res6	11.35	0.36	1.05	0.86	1	2	1
1	res7	0.54	1.00	1.33	0.80	1	1	1
1	res8	2.07	0.05	1.20	0.98	2	1	2
1	res9	0.01	0.00	1.69	1.18	Outlier	Outlier	Outlier
1	res10	4.71	0.01	1.13	1.07	Outlier	Outlier	Outlier
1	res11	0.01	0.00	1.69	1.19	Outlier	Outlier	Outlier
1	res12	0.75	0.00	1.30	1.20	Outlier	Outlier	Outlier
1	res13	4.92	2.21	**1.12**	0.76	1	1	1
1	res14	4.84	1.19	1.13	0.79	2	2	1

[a] Set number.
[b] No data.
[c] The bolded data were selected as test set.

into the stepwise regression using the SPSS software. In this method, the selected parameters for the first subset were added to the next subset and the stepwise regression was continued up to the last subset.

2.2.2.3. Variable subsets (families) stepwise regression (VSSR). In another attempt, 20 calculated subsets were used and the stepwise regression was carried out separately for each subset. The validity problem of significant values was the same in each subset; however, as the parameters were selected separately for each subset, the exclusion of significant parameters was lower than that of the previous methods.

2.2.2.4. Partial least squares (PLS). Since multi-co-linearity among the variables might affect the regression analysis, PLS is frequently used as the variable redundant method. PLS analysis was carried out using the Simca 9.0 software. The PLS method was used for both feature selection as well as PLS model development. During PLS development, the non-significant parameters were eliminated according to the numerical values of the variable coefficients and their importance, and the remaining parameters were used for both the variable selection step for MLR equation as well as PLS model development. A stepwise regression was run through the selected descriptors using the PLS method and the final selected descriptors were used for developing the MLR equations.

2.2.2.5. Genetic algorithm–partial least square (GA–PLS). Genetic algorithms (GA) have been developed to mimic some of the processes observed in natural evolution, which is an efficient strategy to search for the global optima of solutions, and have been successfully applied for feature selection in regression analysis [9]. Moreover, an approach combining GA with PLS (GA–PLS) has also been proposed for variable selection in QSAR and QSPR studies [10,11]. The MATLab 7.1 software was used to run the GA–PLS method developed by Leardi [12]. In this method, the variables were divided into 6 subgroups and each subgroup with the corresponding normalized pIC_{50} values (Y_2) was introduced to the algorithm as input. The output was produced after 20 runs as scored variables. The top 10% scores of each subgroup were used for final feature selection.

In each feature selection method, the variables remaining after exclusion of non-significant parameters were cross correlated in order to select the most relevant parameters concerning the following criteria: (1) $p < 0.05$; (2) having the highest correlation with experimental data; and (3) having the lowest correlation with each other.

2.2.3. MLR model development

The parameters selected using each method were used for developing QSAR equations, and the goodness of fit and statistical significance of the models were evaluated using *R* (correlation coefficient), *F* (variance ratio), and the MPD (Mean Percentage Deviation) values. The MPDs were calculated as follows:

$$MPD = \frac{100}{N}\sum\left|\frac{Y_{pred.} - Y_{\exp.}}{Y_{\exp.}}\right| \quad (1)$$

where, *N* denotes the number of data points, and $Y_{pred.}$ and $Y_{exp.}$ are the predicted and experimental normalized pIC_{50} values. The developed models were evaluated using the leave many out (LMO) and the leave one out (LOO) cross-validation methods. The LOO cross-validation results can be used for estimating the overall mean of squared prediction errors; however, there might be high prediction errors for a subset (or subsets) which could not be reflected in the overall errors. Therefore, in the latter case and especially for model selection purposes, the LMO cross-validation is preferred. In this study, both LOO and LMO methods were used to investigate the robustness and prediction capability of the developed models. The prediction error sum of squares (PRESS) and the cumulative prediction error sum of squares (CumPRESS) were reported for the LMO and LOO validation methods, respectively.

3. Results and discussion

3.1. QSAR analyses of COX-2 inhibition

3.1.1. MLR models

Fifteen MLR equations were developed using 15 pairs of feature selection and test-train selection methods. The developed equations and the details of *R*, *F*, S.E., and MPD (back-calculated and predicted) values are summarized in Table 3. It was found that the best equations based on the *R* and *F* values (goodness of fit), regardless of the feature selection methods, had emerged from the AVKMC test-train selection method. The *R*, *F*, and MPD (±SD) values of test sets for AVKMC series equations were found to be in the range of 0.84–0.91, 25.91–29.53, and 4.8 (±4.5)–5.5 (±4.7), respectively. It should be noted that the higher values of MPD for AVKMC and AS series when compared with the VKMC method is owing to the high error of res8 (in the test set and showing outlier behavior); therefore, when this compound is deleted from the test set, the MPD (±SD) values are in the reduced range of 3.8 (±3.1)–4.5 (±3.3). The results indicated that all 15 equations are applicable to the model COX-2 inhibitory activity of these diaryl compounds and the differences in the model properties are owing to the entering of some compounds (i.e. res8 in test or train set). Subsequently, the equations could be used to accurately predict the activity of these compounds. However, in order to avoid confusion for the user, Eq. (16) which produces less prediction error of 1.6 (±0.8) and contains two frequently selected variables ($R6m^+$ and EEIG07D), was selected as the best equation. This equation could predict 100% of the desired activities with prediction error less than 3.0%, which is comparable with the reported experimental RSD values of 3.0% for similar compounds [1].

3.1.2. PLS models

PLS models were developed using three sets of training data. The details of the developed models are summarized in Table 4. This method was found to produce higher correlation coefficients; however, the other results showed no superiority when compared to the MLR models.

3.2. Comparison of the five feature selection methods

As discussed in Section 3.1.1, the loss of latent variables in the stepwise regression method is one of the problems faced in feature selection. It was found that the number of selected parameters increased (Fig. 4) when using the VSSR method, and it was observed that the use of VBSR without any logical classification of descriptors could not improve the selection procedure. It can be seen in the figure that the GA–PLS and VSSR methods could select the highest number of significant parameters whereas the other methods lost several significant parameters. The most significant and relevant parameters selected by the last three methods (VSSR, PLS, GA–PLS) were found to be similar, while the parameters selected by the first two methods, mainly VBSR, were found to be different for the same data set. This was probably owing to the losses of parameters during the selection process. Although the VSSR method could select the most relevant parameters as PLS and GA–PLS, its main disadvantage was the same as that for the SR method for each family as subsets and it is possible to loose the

Table 3
Details of developed equations along with test-train and feature selection methods and statistical criteria.

Developed equations for COX-2 inhibitory activity								
Equation number	Test-train selection	Feature selection	Equations	R	F	S.E.	MPD (±SD)[a]	MPD (±SD)[b]
2	AS	SR	$Y_2 = -14.81(2.66) + 4.14(0.68)$BEHv1 $- 5.14(0.96)$R8m$^+$ $- 0.06(0.02)$ms $- 1.20(0.43)$G1v	0.83	17.49	0.04	2.5 (±1.9)	5.7 (±4.5)
3	VKMC	SR	$Y_2 = 1.20(0.11) + 0.04(0.02)$Mor27u $+ 0.16(0.03)$EEIG08D $+ 0.09(0.03)$Mor16m $- 1.63(0.63)$G1v $- 0.05(0.01)$RDF120p $- 1.24(0.25)$HATS1u	0.83	10.96	0.05	2.9 (±2.1)	2.4 (±1.9)
4	AVKMC	SR	$Y_2 = 0.61(0.02) + 5.08(0.57)$R6m$^+$ $+ 0.09(0.02)$EEIG07D $+ 0.52(0.15)$RBF $- 0.01(0.00)$RDF035p	0.89	29.53	0.05	1. 9 (±1.4)	5.5 (±4.7)
5	AS	VBSR	$Y_2 = -16.47(2.64) + 4.44(0.68)$BEHv1 $- 0.05(0.02)$Ms $- 0.04(0.01)$RDF120p $+ 0.03(0.01)$RDF010e	0.83	16.98	0.03	2. 8 (±2.0)	4.7 (±4.6)
6	VKMC	VBSR	$Y_2 = 0.81(0.04) + 0.12(0.02)$Mor27u $- 0.22(0.05)$Mor14v $- 0.78(0.19)$R2e$^+$	0.76	14.44	0.04	3.3 (±2.8)	3.0 (±3.3)
7	AVKMC	VBSR	$Y_2 = 0.72(0.03) + 3.96(0.60)$R6m$^+$ $+ 1.15(0.21)$RBF $- 0.0004(0.0000)$MW	0.84	26.43	0.02	2.3 (±2.8)	4.8 (±4.5)
8	AS	VSSR	$Y_2 = 0.74(0.04) + 4.33(0.61)$R6M$^+$ $+ 0.09(0.02)$EEIG07d $- 0.07(0.02)$R2e	0.84	25.02	0.02	2.6 (±2.8)	2.8 (±4.6)
9	VKMC	VSSR	$Y_2 = 0.74(0.09) + 0.05(0.02)$Mor27u $- 0.28(0.12)$HATS2u $+ 0.04(0.02)$L3u $+ 0.03(0.01)$RDF010e	0.73	8.60	0.04	3.5 (±3.9)	2.0 (±1.8)
10	AVKMC	VSSR	$Y_2 = 0.40(0.08) + 4.01(0.55)$R6m$^+$ $+ 0.06(0.01)$EEIG07D $+ 0.11(0.04)$JHETP $+ 0.50(0.16)$RBF	0.88	25.91	0.02	1.9 (±1.7)	5.1 (±4.5)
11	AS	PLS	$Y_2 = 0.64(0.03) + 4.15(0.8)$R6m$^+$ $+ 0.08(0.02)$EEIG07D	0.74	19.95	0.03	2.8 (±2.8)	2.1 (±2.0)
12	VKMC	PLS	$Y_2 = 0.73(0.05) + 2.87(0.86)$R6M$^+$ $+ 0.07(0.02)$EEIG07D $- 0.02(0.01)$ALOGP	0.76	14.19	0.03	3.3 (±3.9)	2.1 (±1.5)
13	AVKMC	PLS	$Y_2 = 0.64(0.02) + 5.55(0.79)$R6m$^+$ $+ 0.06(0.01)$EEIG07D $+ 0.50(0.14)$RBF $- 1.89(0.5)$R7m$^+$ $- 0.12(0.04)$H8e	0.91	27.40	0.02	1.6 (±1.7)	5.2 (±4.8)
14	AS	GA–PLS	$Y_2 = 0.33(0.11) + 3.67(0.74)$R6M$^+$ $+ 0.08(\pm0.02)$EEIG07D $+ 0.17(0.06)$JHETP	0.80	18.86	0.03	2.5 (±2.7)	3.1 (±5.5)
15	VKMC	GA–PLS	$Y_2 = 0.90(0.04) + 0.04(0.02)$Mor27u $- 0.45(0.11)$HATS2E $+ 0.50(0.16)$GGI8	0.72	11.25	0.04	3.5 (±2.9)	3.2 (±2.5)
16	AVKMC	GA–PLS	$Y_2 = 0.40(0.08) + 4.01(0.55)$R6m$^+$ $+ 0.06(0.01)$EEIG07D $+ 0.11(0.04)$JHETP $+ 0.50(0.16)$RBF	0.88	25.91	0.02	3.1 (±3.3)	1.6 (±0.8)
Developed equation for COX-1 inhibitory activity								
17	AVKMC	GA–PLS	$Y_1 = 1.79(0.10) - 0.26(0.05)$GATS1M $- 0.43(0.12)$H0p	0.89	60.61	0.06	5.0 (±3.9)	7.7 (±5.6)

[a] MPD for back-calculation (train set).
[b] MPD for prediction (test set).

Table 4
Details of developed PLS model.

Training selection	R	Number of PCs	Number of parameters	MPD (±SD) for train set	MPD (±SD) for test set
AS	0.91	1	61	3.4 (±2.9)	4.3 (±5.9)
VKMC	0.88	2	49	3.5 (±3.9)	2.4 (±2.4)
AVKMC	0.93	4	111	1.4 (±1.5)	5.9 (±4.5)

relevant parameters from each subset. In order to overcome this drawback, it has been recommended that the number of variables in each variable subset should be decreased. In that case, the probability of the latent variable loss decreases, and the method could be introduced as an alternative for the classic stepwise methods. Alternatively, GA–PLS could be used in order to ensure that most of the relevant parameters are selected. Regarding the PLS method, its disadvantage was the high cross correlation between the selected parameters wherein all selected parameters could evaluate the response well.

3.3. Comparison of the three test-train selection methods

It should be noted as shown in Section 3.1 that the best equation emerged from the AVKMC training set, and the use of all information space of data sets (response space and descriptor space) leads to more general training sets. Owing to the generality of the training set selected using this method, it can be indicated that almost all feature selection methods would lead to acceptable equations with the lowest prediction errors. Since the VKMC method uses only descriptors' information, it produced less significant equations, where in some cases (Eqs. (9) and (12)) low prediction errors were produced. Also, the variation in the feature selection methods leads to different equations and it can be seen that there is least similarity between selected descriptors in this method. Similar to the AVKMC method, the AS method was found to be less sensitive to the feature selection method and the most similar equations were found in this method; however, the equations developed using this method were found to produce the highest prediction errors.

3.4. QSAR analysis of COX-1 inhibition

The training and test sets were selected using the AVKMC method and the details of selection are summarized in Table 2. The GA–PLS feature selection method was used for selecting relevant

Fig. 4. Comparison of the number of significant parameters selected using different feature selection methods, SR: stepwise regression, VBSR: Variable blocked (200

Table 5
The results for leave one out and 9 fold leave many out cross validations.

Equation number	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
LMO (PRESS)	0.03	0.04	0.02	0.03	0.02	0.04	0.01	0.02	0.02	0.03	0.02	0.02	0.02	0.03	0.02	0.02
	0.01	0.01	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.00	0.02
	0.01	0.03	0.00	0.01	0.01	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.10
	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0.01	0.02
	0.00[a]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
	0.04	0.06	0.02	0.03	0.02	0.06	0.02	0.03	0.02	0.04	0.04	0.02	0.03	0.02	0.02	0.09
LOO (CumPRESS)	0.1	0.16	0.07	0.08	0.07	0.12	0.06	0.08	0.08	0.08	0.08	0.07	0.08	0.1	0.08	0.25

[a] 0.00 means less than 0.005.

parameters with regard to the selection criteria explained in Section 2.2.2. The QSAR equation was developed as follows:

$$Y_1 = 1.79(0.10) - 0.26(0.05)\mathrm{GATS1M} - 0.43(0.12)\mathrm{H0p} \quad (17)$$

The MPD (±SD) values obtained using this equation for the training and prediction set were 5.0 (±3.9) and 7.7 (±5.6), respectively. The significance criterion was $p < 0.05$ for both descriptors. It was found that the model could predict 88.8% of data with prediction error less than 10.0%. However, the developed model could not predict the COX-2 inhibition activity, and the parameters were not significant.

3.5. Selectivity prediction

The proposed QSAR model for COX-2 inhibition (Eq. (16) of Table 3) and COX-1 inhibition (Eq. (17)) was used for predicting the selectivity of entire structures, and the MPD (±SD) values obtained were 12.7 (±10.4) and 16.9 (±9.6), respectively, for the training and test data sets. It was found that the proposed method could predict about 50% of selectivity data with prediction error less than 10.0%.

3.6. Method validation

The leave one out and leave many out cross-validation methods were used for evaluating the robustness of the developed models, and the PRESS and CumPRESS values obtained are reported in Table 5. It can be seen in the table that there is no significant difference between the errors produced by the different models and the acceptable range of prediction (0.00–0.04 and 0.06–0.16 for the LMO and LOO methods, respectively), which indicates that all the developed models are valid and applicable for different sets of data points.

Table 6
Selected descriptors for COX-2 and COX-1 inhibitory activities.

COX-2		
Variable name	Explanation	Descriptor family
Mor08m	3D-MoRSE – signal 08/weighted by atomic masses	3D-MoRSE descriptors
Mor14v	3D-MoRSE – signal 14/weighted by atomic van der Waals volumes	3D-MoRSE descriptors
Mor16m	3D-MoRSE – signal 16/weighted by atomic masses	3D-MoRSE descriptors
Mor17u	3D-MoRSE – signal 17/unweighted	3D-MoRSE descriptors
Mor27u	**3D-MoRSE – signal 27/unweighted**	**3D-MoRSE descriptors**
BEHv1	Highest eigenvalue n. 1 of Burden matrix /weighted by atomic van der Waals volumes	Burden eigenvalues
BEHv1	Highest eigenvalue n. 1 of Burden matrix /weighted by atomic van der Waals volumes	Burden eigenvalues
Ms	Mean electrotopological state	Constitutional descriptors
RBF	**Rotatable bond fraction**	**Constitutional descriptors**
EEig07d	**Eigenvalue 07 from edge adj. matrix weighted by dipole moments**	**Edge adjacency indices**
EEig08d	Eigenvalue 08 from edge adj. matrix weighted by dipole moments	Edge adjacency indices
H8e	H autocorrelation of lag 8/weighted by atomic Sanderson electronegativities	GETAWAY descriptors
HATS1u	Leverage-weighted autocorrelation of lag 1/unweighted	GETAWAY descriptors
HATS2e	Leverage-weighted autocorrelation of lag 2/weighted by atomic Sanderson electronegativities	GETAWAY descriptors
HATS2u	Leverage-weighted autocorrelation of lag 2/unweighted	GETAWAY descriptors
R2e	*R* autocorrelation of lag 2/weighted by atomic Sanderson electronegativities	GETAWAY descriptors
R2e+	*R* maximal autocorrelation of lag 2/weighted by atomic Sanderson electronegativities	GETAWAY descriptors
R6m+	**R maximal autocorrelation of lag 6/weighted by atomic masses**	**GETAWAY descriptors**
R7m+	*R* maximal autocorrelation of lag 7/weighted by atomic masses	GETAWAY descriptors
R8u+	*R* maximal autocorrelation of lag 8/unweighted	GETAWAY descriptors
ALOGP	Ghose-Crippen octanol-water partition coeff. (logP)	Molecular properties
RDF010e	Radial Distribution Function – 1.0/weighted by atomic Sanderson electronegativities	RDF descriptors
RDF035p	Radial Distribution Function – 3.5/weighted by atomic polarizabilities	RDF descriptors
RDF120p	Radial Distribution Function – 12.0/weighted by atomic polarizabilities	RDF descriptors
GGI8	Topological charge index of order 8	Topological charge indices
Jhetp	**Balaban-type index from polarizability weighted distance matrix**	**Topological descriptors**
G1v	1st component symmetry directional WHIM index /weighted by atomic van der Waals volumes	WHIM descriptors
L3u	3rd component size directional WHIM index/unweighted	WHIM descriptors
GATS1m	Geary autocorrelation – lag 1 /weighted by atomic masses	2D autocorrelations
H0p	H autocorrelation of lag 0 /weighted by atomic polarizabilities	GETAWAY descriptors

The frequently selected parameters are bolded.

3.7. Descriptor evaluation

The parameters selected using different methods and their definitions are shown in Table 6. The selected descriptors belong to Constitutional, 3D-MoRSE, Edge adjacency indices, Topological and GETAWAY descriptors. The results indicate that in most cases a combination of one autocorrelation ($R6m^{+}$) and one edge adjacency index (EEig07d) could represent the COX-2 inhibition of the drugs investigated. There is at least one descriptor weighted by atomic mass or volume and one descriptor by dipole moment or electronegativity in all models. In addition, the RBF parameter, which is a measure of molecule flexibility, was selected frequently. It was found that the structures with the highest efficiency were those with the lowest flexibility, atomic distribution, and the maximum negative electronegativity. These findings are in agreement with the previous findings regarding a limited space of active site and its hydrophobic characteristics [10].

The low electronegativity is favorable for COX-1 inhibition while this activity is reduced by an increase in the size of the molecules. In summary, steric effects are important for both COX-1 and COX-2 inhibitions, and in order to design an effective COX-2 inhibitor with a moderate COX-1 inhibitory effect, an optimum conformation should be found. Also, the electronegativity of the substitute should be optimized.

4. Conclusion

The results suggest that the selective inhibition is dependent on the size and polarizability, which is in agreement with the previous findings of the limited space of active site and its hydrophobic characteristics. Similar results of the PLS, GA–PLS, and the VSSR methods indicated that these methods could be used for the feature selection whereas the highest number of significant and relevant variables could be selected using the GA–PLS method. Also despite of PLS, cross correlation between selected parameters are less in GA–PLS. Owing to the low sensitivity to the feature selection methods, the AVKMC was selected as an appropriate test-train selection method. The experiments were performed on one data set and further investigations should be carried out on a number of data sets to determine the most appropriate method of feature selection and test-train selection. To conclude, the findings would facilitate the development of COX-2 inhibitors with a mild inhibitory activity of COX-1 isoform.

Acknowledgments

We thank Dr Dastmalchi and Dr Ghafourian to their kind permission of using Dragon and Simca software, and Drug Applied Research Centre of Tabriz University of Medical Sciences for the partial financial support under grant No. 88/11.

References

[1] M. Murias, N. Handler, T. Erker, K. Pleban, G. Ecker, P. Saiko, T. Szekeres, W. Jäger, Resveratrol analogues as selective cyclooxygenase-2 inhibitors: synthesis and structure-activity relationship. Bioorg. Med. Chem. 12 (2004) 5571–5578.
[2] S. Prasanna, E. Manivannan, S.C. Chaturvedi, QSAR analyses of conformationally restricted 1,5-diaryl pyrazoles as selective COX-2 inhibitors: application of connection table representation of ligands. Bioorg. Med. Chem. Lett. 15 (2005) 2097–2102.
[3] M. Khoshneviszadeh, N. Edraki, R. Miri, B. Hemmateenejad, Exploring QSAR for substituted 2-sulfonyl-phenyl-indol derivatives as potent and selective COX-2 inhibitors using different chemometrics tools. Chem. Biol. Drug. Des. 72 (2008) 564–574.
[4] W.J. Tsai, Y.J. Shiao, S.J. Lin, W.F. Chiou, L.C. Lin, T.H. Yang, C.M. Teng, T.S. Wu, L. M. Yang, Selective COX-2 inhibitors. Part 1: synthesis and biological evaluation of phenylazobenzenesulfonamides. Bioorg. Med. Chem. Lett. 16 (2006) 4440–4443.
[5] S.J. Lin, W.J. Tsai, W.F. Chiou, T.H. Yang, L.M. Yang, Selective COX-2 inhibitors. Part 2: synthesis and biological evaluation of 4-benzylideneamino- and 4-phenyliminomethyl-benzenesulfonamides. Bioorg. Med. Chem. 16 (2008) 2697–2706.
[6] Ž. Debeljak, V. Marohnić, G. Srečnik, M. Medić-Šarić, Novel approach to evolutionary neural network based descriptor selection and QSAR model development. J. Comput. Aided. Mol. Des. 19 (2005) 835–855.
[7] Organazation for Economic Co-operation and Development, Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship ((Q) SAR) Models OECD series on testing and assessment 69. OECD document ENV/JM/MONO. Organazation for Economic Co-operation and Development, 2007, pp 55–65.
[8] A. Golbraikh, M. Shen, Z. Xiao, Y.D. Xiao, K.H. Lee, A. Tropsha, Rational selection of training and test sets for the development of validated QSAR models. J. Comput. Aided. Mol. Des. 17 (2003) 241–253.
[9] R. Leardi, M.B. Seasholtz, R.J. Pell, Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data. Anal. Chim. Acta. 461 (2002) 189–200.
[10] H. Kubinyi, Evolutionary variable selection in regression and PLS analyses. J. Chemom 10 (1996) 119–133.
[11] R. Leardi, A.L. GonzÃlez, Genetic algorithms applied to feature selection in PLS regression: how and when to use them. Chemom. Intel. Lab. Sys. 41 (1998) 195–207.
[12] P. Silakari, S.D. Shrivastava, G. Silakari, D.V. Kohli, G. Rambabu, S. Srivastava, S. K. Shrivastava, O. Silakari, QSAR analysis of 1,3-diaryl-4,5,6,7-tetrahydro-2H-isoindole derivatives as selective COX-2 inhibitors. Eur. J. Med. Chem. 43 (2008) 1559–1569.