Research in Pharmaceutical Sciences, October 2011; 6(2): 71-80
Received: Jun 2011
Accepted: Aug 2011

School of Pharmacy & Pharmaceutical Sciences
Isfahan University of Medical Sciences

Original Article

A modeling study of aldehyde inhibitors of human cathepsin K using partial least squares method

M. Shahlaei[1,2], A. Fassihi[2,*], L. Saghaie[2], E. Arkan[3] and A. Pourhossein[4]

[1]*Department of Medicinal Chemistry, Faculty of Pharmacy, Kermanshah University of Medical Sciences, Kermanshah, I.R.Iran.*
[2]*Department of Medicinal Chemistry, School of Pharmacy and Isfahan Pharmaceutical Sciences Research Center, Isfahan University of Medical Sciences, Isfahan, I.R.Iran.*
[3]*Department of Medical Nanotechnology, School of Advanced Medical Technologies, Tehran University of Medical Sciences, Tehran, I.R.Iran.*
[4]*Young Researchers Club, Kermanshah branch, Islamic Azad University, Kermanshah, I.R.Iran.*

Abstract

Quantitative relationships between molecular structure of forty eight aldehyde compounds with their known Cathepsin K inhibitory effects were discovered by partial least squares (PLS) method. Evaluation of a test set of 10 compounds with the developed PLS model revealed that this model is reliable with a good predictability. Since the QSAR study was performed on the basis of theoretical descriptors calculated completely from the molecular structures, the proposed model could potentially provide useful information about the activity of the studied compounds. Various tests and criteria such as leave-one-out cross validation, leave-many-out cross validation, and also criteria suggested by Tropsha were employed to examine the predictability and robustness of the developed model.

Keywords: QSAR**;** Partial Least Squares**;** Cathepsin K inhibitory activity

INTRODUCTION

Design, development, and introduction of new drugs to the market are difficult, time consuming and cost-intensive procedures. Furthermore, during the procedure, limited number of candidates will be tested in the clinic and even smaller number will be introduced to the market. Any process or tool that can accelerate the effectiveness of any step in the drug discovery procedure seems to be very attractive. Quantitative structure activity relationship (QSAR) studies have been proved as a new possibility to facilitate drug discovery procedures (1-6). The key point is that in medicinal chemistry the activity of each ligand depends on its molecular structure. In QSAR models, mathematical equations are constructed and used to make a connection between the activity and the structure of the compounds. In a typical QSAR study, numerous descriptors are calculated. These descriptors have been classified into different categories, including constitutional, geometrical, topological, quantum chemical and so on. After calculation of the descriptors, one needs to find a set of molecular descriptors with the higher impact on the biological activity of the interest.

Cathepsin K (catK), one of the most important members of the group of lysosomal cysteine proteases, is mainly expressed in ovary and in osteoclasts or osteoclastomas (7). Various studies have shown that catK is a cysteine protease with a predominant if not exclusive function in degradation of the bone matrix. This proposal was indeed supported by the identification of catK as the target gene in the human disorder Pycnodysostosis, where functional mutations in the catK gene cause severe bone malformation (8). Further proof of the function of catK in osteoclast-mediated bone degradation came from catK-deficient mice that demonstrated an osteopetrotic phenotype (9). Since then, most of the investigations have concentrated on this intriguing

*****Corresponding author: Afshin Fassihi
Tel. 0098 311 7922562, Fax. 0098 311 6680011
E-mail:fassihi @pharm.mui.ac.ir

function of catK, because it represents an excellent target for the development of therapeutic strategies in the treatment of skeletal disorders such as Pycnodysostosis or osteoporosis.

Therfore, catK is a key protease in osteoclast-mediated bone resorption and it highlights the attractiveness of this cysteine protease as a target for inhibition in diseases characterized by elevated level of bone turnover such as osteoporosis (10,11). Currently, many kinds of inhibitors against catK have been designed which include nonpeptidic biaryl compounds, aldehydes and their derivatives, acyclic and cyclic ketones, nonpeptidyl nitriles, epoxysuccinyl analogues, β-lactams, vinyl sulfones, and so on. Some of them inhibited bone resorption well *in vivo* (10,11). In this study, a QSAR model is developed from the calculated descriptors derived from semi-empirical (AM1) quantum chemical calculations for predicting the activity values of some of aldehyde compounds as human catK. Main objective of this study is to develop an accurate, simple, reliable, and less expensive technique for calculation of bioactivity values. The PLS method was used in QSAR for modeling the relationship between activities of 48 aldehyde compounds and their structural descriptors.

A training set (38 aldehydes) of compounds was employed to refine the generated model and a testing set (10 aldehydes) of appropriately selected chemicals was chosen to test the model.

Multiple linear regression (MLR) is an approach commonly employed in QSAR studies. The multicolinearity problem of the MLR technique has been overcome by using the development of the partial least squares (PLS) approach, which plays a significant role in various QSAR studies. PLS is a helpful method for relating a set of activities to many explanatory variables such as theoretical descriptors. It can be regarded as a general dimension reduction method which takes into account the linear relationship between the dependent and independent variables.

MATERIALS AND METHODS

Preparation of data set and calculation of the descriptors

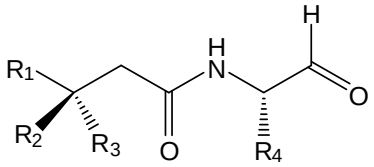
The studied compounds and their biological activities were taken from the literature (12,13) which are listed in Table 1 The biological activity was expressed by IC_{50} (the molar concentration of aldehyde compounds required to inhibit 50% of catK). In our study, $-\log(IC_{50})$ values were employed as the dependent variables which are given in Table 1.

All molecules were drawn by Hyperchem and preoptimized using the MM+ molecular mechanic force field and then a more precise optimization was performed with the semiempirical AM1 method (14). The molecular structures were optimized using the Polak–Ribiere algorithm until the root mean square gradient reached 0.01. The Hyperchem output files (.hin files) were introduced to DRAGON program (15) to calculate four classes of the descriptors: constitutional (number of various types of atoms and bonds, number of rings, molecular weight, etc.), topological (Wiener index, Randic indices, Kier–Hall shape indices, etc.), geometrical (moments of inertia, molecular volume, molecular surface area, etc.), and functional group (number of total tertiary carbons (nCt), number of H-bond acceptor atoms (nHAcc), number of total hydroxyl groups (nOH), number of unsubstituted aromatic C (nCaH), number of ethers (aromatic) (nRORPh), etc.) (15,16).

Kennard and Stone algorithm

After building new X matrix including latent variables for evaluation of performance of generated regression methods, about 20% of the molecules were selected as test set molecules. It is well known that for building of any QSAR model in general, the selection of the molecules is the important step in building or training of the model. In order to apply the standard QSAR modeling method, the studied data set should be split into the training (learning) and the testing sets. The best situation of this stage of model building is dividing data set to guarantee that both training and testing sets individually cover the total space occupied by original data set. Then

Table 1. Structural details of investigated compounds used in this study



Compound	R1	R2	R3	R4	pIC_{50}	Predicted pIC_{50}
1[a]	Me	Me	Me	H	2.24	3.80
2	Me	Me	Me	Me	3.7	3.75
3	Me	Me	Me	Et	4.24	3.94
4	Me	Me	Me	i-Pr	3.8	4.23
5[a]	Me	Me	Me	Pr	4.42	4.25
6	Me	Me	Me	CH(Me)Et(S)	3.6	4.07
7[a]	Me	Me	Me	CH(Me)Et(R)	3.89	4.13
8	Me	Me	Me	$CH_2C(CH_3{=}CH2)$	4.1	3.81
9	Me	Me	Me	CH_2-i-Pr	4.51	4.19
10	Me	Me	Me	CH_2tBu	3.6	3.67
11	Me	Me	Me	Bu	4.29	3.98
12	Me	Me	Me	$CH_2C{\equiv}CCH_3$	3.39	3.41
13[a]	Me	Me	Me	n-Pentyl	3.96	4.27
14	Me	Me	Me	n-Hexyl	4.64	4.63
15	Me	Me	Me	CH_2SEt	4.08	4.09
16	Me	Me	Me	$(CH_2)_2OMe$	3.17	3.62
17[a]	Me	Me	Me	$(CH_2)_2SMe$	4.11	4.16
18	Me	Me	Me	Ph	4.29	4.14
19	Me	Me	Me	Benzyl	3.96	4.18
20	Me	Me	Me	(CH_2)cyclohexyl	4.44	4.37
21[a]	Me	Me	Me	$(CH_2)_2$cyclohexyl	4.59	4.69
22	Me	Me	Me	$(CH_2)_3Ph$	4.82	4.84
23	Me	Me	Me	$(CH_2)_3$cyclohexyl	4.8	4.92
24	Me	Me	Me	CH_2SCH_2	4.4	4.32
25	Me	Me	Me	$CH_2NHCOMe$	3.43	3.37
26	Me	Me	Me	$(CH_2)_4NHCO_2Me$	4.82	4.81
27	Me	Me	Me	$(CH_2)_4N(H)COCF_3$	4.39	4.24
28	Me	Me	Me	$(CH_2)_4N(Me)COCF_3$	4.3	4.44
29	$PhCH_2$	Me	Me	n-Bu	4.92	5.26
30	$PhCH_2$	$(CH_2)_3$		n-Bu	5.62	5.68
31	$PhCH_2$	Et	Et	n-Bu	5.68	5.62
32	$PhCH_2$	(CH2)4		n-Bu	6.46	6.68
33	$PhCH_2$	(CH2)5		n-Bu	5.7	5.62
34	$PhCH_2$	Me	H	n-Bu	5.74	5.76
35[a]	$C_6H_{11}CH_2$	Me	H	n-Bu	5.57	5.89
36	$PhCH_2$	Et	H	n-Bu	6.89	5.83
37[a]	$PhCH_2$	n-Pr	H	n-Bu	5.82	5.97
38[a]	$PhCH_2$	i-Pr	H	n-Bu	6.3	5.85
39	$PhCH_2$	i-Bu	H	n-Bu	5.19	5.72
40	H	Et	Et	n-Bu	5.4	5.58
41	H	n-Pr	n-Pr	n-Bu	6.06	6.29
42	H	i-Pr	i-Pr	n-Bu	6.25	6.11
43	H	i-Bu	i-Bu	n-Bu	5.96	6.02

Table 1. (Continued)

Compound	R1	R2	R3	R4	pIC_{50}	Predicted pIC_{50}
44	Me	i-Pr	i-Pr	n-Bu	5.33	5.19
45	3-MeO-Ph-CH_2	Me	Me	n-Bu	5.27	5.12
46 [a]	2-Cl-Ph-CH_2	Me	Me	n-Bu	5.51	5.26
47	4-Cl-Ph-CH_2	Me	Me	n-Bu	5.34	5.12
48	3-Thiophenyl-CH_2	Me	Me	n-Bu	5.12	5.09

[a]molecules selected as the test set

ideal splitting of data set is carried out such that each of objects in the testing set be close to at least one of the objects in the training set. Various methods were used as tools for splitting the whole original data set to the training and testing sets. According to Tropsha, the best models would be built when Kennard and Stone algorithm is used (17). This algorithm was applied in the current study (18). This method has some advantages: the training set molecules map the measured region of the input variable space completely with respect to the induced metric. The other advantage is that all of the testing molecules fall inside the measured region.

Partial Least Squares (PLS)

PLS is a regression approach which is used to build a predictive model between two matrices of variables: the X matrix of predictor variables and the Y matrix of dependent variables. In its simplest type of model building, a linear model indicates the relation-ship between dependent (bioactivity) varia-bles and independent (descriptors) variables by means of latent variables (LVs).

In the PLS regression, it is assumed that X matrix (I × J) contains the descriptors that can be used for predicting the matrix of activities that is Y (I × M). Here the dependent variables are represented by an (I × 1) column vector. PLS decomposes these matrices into a two-matrix product plus residual.

$$X = TP^T + E = \sum t_f p'_f + E$$

$$Y = UQ^T + F = \sum u_f q'_f + E$$

where, T and U are the matrices of score for X and Y; P and Q are the matrices of loadings for X, Y; E and F are the matrices residual, respectively, for a model with *f* latent variables.

Above equations are solved in a way to maximize the covariance between T and U. These two matrices are related by the following inner relationship:

$$U=TB+H$$

where, B is a diagonal matrix and H is a residual matrix. This allows PLS to be expressed as a predictive model. The matrix Y can be calculated from U as follows:

$$Y=TBQ^T + F$$

The activity of the new compounds can be approximated from the new scores T^*, which are substituted in the above equation, leading to the following equation:

$$Y_{pred.}=T \times BQ^T$$

In order to find the optimum number of latent variables to be used in model building, a leave-one-out cross validation was carried out (19).

RESULTS

Numerous descriptors were calculated for each studied molecule using Dragon. In order to get the linear relationship with independent variables, logarithms of the inverse of biological activity (Log 1/IC_{50}) data of 48 molecules were used.

PLS modeling

PLS generated eleven significant LVs (the percent of variance explaind > 1) which can explain around 95% of the variances in the original descriptors data matrices. eleven LVs are reported in Table 2 In this table the percent of variances was explained by each LVs and the cumulative percent of variances are represented. Therefore, we restricted the next studies to the selection of best subset of these LVs to perform regression between

Table 2. The results of PLS from the total calculated descriptors

Latent variable	% of Variance explained	Cumulative %
1	60.410	60.410
2	11.828	72.238
3	7.732	79.971
4	4.355	84.326
5	2.380	86.707
6	1.405	88.112
7	1.404	89.517
8	0.748	90.266
9	2.369	92.635
10	1.372	94.007
11	1.065	95.073

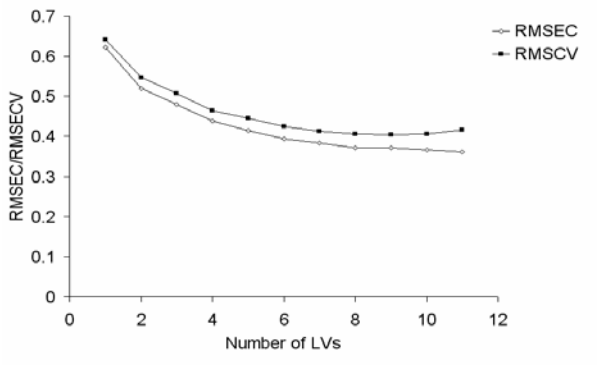


Fig. 1. Optimization of the number of LVs

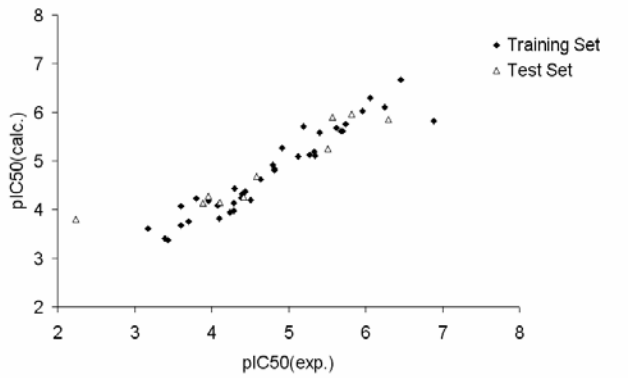


Fig. 2. The calculated pIC_{50} of studied compounds vs experimental pIC_{50}

descriptors and activity. After dividing the molecules into two parts, calibration and validation sets, based on Kennard and Stone algorithm, building of regression model using calibration set was performed. The training and validation compounds are clearly indicated in Table 1.

Two quantities including root mean square error of calibration (*RMSEC*) and root mean square error of cross validation (*RMSECV*) were used to optimize the number of the latent variables in model development. As it is shown in Fig. 1, the best PLS model contained nine latent variables. The predicted pIC_{50}s by using PLS regression technique are listed in Table 1 and are plotted in Fig. 1 The plot of Fig. 2 shows that the data are distributed around a straight line with the respective slope equal to 0.907.

As it can be seen from Table 3, the QSAR model based on PLS possess a high statistical quality. It could respectively explain and predict 90% and 83% of variances in the human catK inhibitory activity of the investigated compounds. The predictability of the generated PLS-based QSAR model was estimated according to Tropsha, Roy and coworkers recommended criteria (17,20). The results of LOO-CV technique applied on the training set are reported in Table 3. This results showed that generated PLS model is a reasonable QSAR model. These results confirm the success of calculated descriptors in modeling of the human catK inhibitory activity of the studied compounds. The value of R^2 for test set is reported in Table 3 The data revealed that the proposed model has high prediction ability for the prediction set.

The proposed regression models passed all the Tropsha tests for the predictive ability. Values of these quantities are shown in Table 3 In order to avoid chance correlations which

Table 3. Statistic parameters and figures of merits of developed GA-ANFIS model

Statistics parameter	PLS model	
	Training set	Test set
N	38	10
R^2	0.907	0.838
$RMSE$	0.323	0.549
$PRESS$	2.922	3.032
R^2_{LOOCV}	0.847	
$RMSE_{LOOCV}$	0.363	
R^2_{L5OCV}	0.812	
$RMSE_{L5OCV}$	0.403	
R^2-R_0^2/R^2	-0.101	-0.153
R^2-$R'_0{}^2/R^2$	-0.101	-0.186
k	1.000	0.970
k'	0.996	1.017
R_m^2	0.632	0.537

are possible because of a large number of generated columns (independent variables), and to examine the robustness of developed models, Y randomization test was applied to the models. The dependent variable vector is randomly permuted and a new QSAR model was constructed using the original independent variable matrix. The new modeling was expected to have low R^2 values. For sureness, some iteration was carried out. If the results show a high R^2, it implies that an acceptable QSAR model can not be obtained. The low $R^2$and R^2_{CV} values show that the good results in our original model are not due to a chance correlation or structural dependence of the training set.

DISCUSSION

To solve the problem of multicollinearity in the generated descriptors, PLS regression as a linear method was used to model structure-activity relationships quantitatively. All the calculated descriptors were used in the modeling procedure.

In multivariate data analysis, a representative training set must be extracted from a pool of real objects. Moreover, test objects should also be chosen to assess the quality of the developed model and to determine model parameters such as the number of latent variables in PLS regression. Several studies have addressed the problem of choosing a representative subgroup from a pool of objects.

In this context, random sampling is a well-liked method because of its straight forward-ness and also because a set of objects randomly selected from a larger set follows the statistical distribution of the entire data set. However, random sampling does not assure the representativity of the total data set, nor does it avoid extrapolation problems. Actually, random selection does not guarantee that the objects on the boundaries of the total data set are included in the training set. An alternative approach to random selection method that is frequently used is the Kennard and Stone algorithm. Kennard and Stone is aimed at covering the multidimensional space in a uniform manner by maximizing the Euclidean distances between the calculated descriptors X matrix of the studied molecules.

There are several tools to estimate and calculate the accuracy,the validity of the proposed QSAR model and the impacts of the preprocessing steps. Here, we have employed several techniques to ensure the effectiveness of the PLS in the modeling of catK inhibitory activity of studied aldehydes. Some of the common parameters used for checking the predictability of proposed PLS model are root

mean square error (*RMSE*), square of the correlation coefficient (R^2), and predictive residual error sum of squares (PRESS). These parameters were calculated as follows:

$$RMSE = [1/n\sum_{i=1}^{n}(\hat{y} - y_i)^2]^{1/2}$$

$$R^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 / \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$$

$$PRESS = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

where, y_i is the measured bioactivity of the investigated compound i, $\hat{y}_i$ represents the calculated bioactivity of the compound i, $\bar{x}$ is the mean of true activity in the studied set, and n is the total number of molecules used in the studied sets.

The efficacy of QSAR models is not just their capability to regenerate known data, but also they should have talent to generate a good estimation for any external data (21). The predictabilities of developed models are powerfully influenced by the overfitting problem. Overfitting problem is occurred when uninformative regressions enter to the developed QSAR model. Another reason of overfitting problem is the use of exceeded number of LVs in PLS model. There are several techniques to approximate the quality and accuracy of the QSAR models (22). Cross-validation is the most regularly employed validation techniques (23). Consequently, to examine the predictability and to check overfitting problem in the resulting PLS model, the leave-one-out cross validation procedure was employed. The squared correlation coefficient for cross-validation (R^2_{CV}) was then calculated by the following equation:

$$R^2_{CV} = 1 - (PRESS / SSD)$$

where, PRESS and SSD are the predicted residual sum of squares and the sum of the

squared deviation from the mean, respectively.

For a generated QSAR model, internal validation (including leave-one-out cross validation), although significant and essential, does not adequately assure the predictability of a developed model. In fact, it is insisted that models with high apparent predictive ability which is highlighted only by internal validation methods cannot be predictive when applied on new compounds not employed in developing the model. Thus, for a stronger estimation of the application of developed model for prediction on new chemicals, external validation of the models should always be carried out (17). To complete the study with regards to the predictability of the generated model, the proposed PLS must be used to predict the activity of ten molecules that did not employ in the modeling step (the testing set compounds). This predictive ability is estimated by the external R^2_p (R^2 for test set) that is defined as follows (24):

$$R^2_p = 1 - \frac{\sum_{i=1}^{test}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{test}(y_i - \bar{y}_{tr})^2}$$

where, $\bar{y}_t$ is the average value of the bioactivity for the training set. The summations cover all the molecules in the testing set.

Some criteria are suggested by Tropsha (17). If these criteria were satisfied then it could be concluded that the model is predictive (17). These criteria include:

$R^2_{LOO} > 0.5$

$R^2 > 0.6$

$$\frac{R^2 - R_0^2}{R^2} < 0.1 \qquad \frac{R^2 - R_0'^2}{R^2} < 0.1$$

$0.85 < k < 1.15$ or $0.85 < k' < 1.15$

R^2 is the correlation coefficient of regression between the predicted and observed activities of the compounds in training and test sets. R_0^2 is the correlation coefficients for regressions between predicted versus observed activities through the origin, $R_0'^2$ is the correlation coefficients for regressions between observed versus predicted activities through the origin, and the slope of the regression lines through the origin are assigned by k and k', respectively. Details of definitions of parameters such as R_0^2, $R_0'^2$, k and k' are presented in the literature (17).

In addition, according to Roy and coworkers (20) the difference between values

of R_0^2 and $R_0'^2$ must be studied and given importance. They suggested following modified R^2 form:

$$R_m^2 = R^2(1-\left|\sqrt{R^2-R_0^2}\right|)$$

If R_m^2 value for given model is >0.5, indicates good external predictability of the developed model.

QSAR applicability domain

The applicability of domain (AD) was explained by the Williams plot of standardized residuals versus leverage (Hat diagonal) values (h_i). The leverage method for defining the AD has been explained in details in the literature (17). The leverage (h) value of a compound in the original independent variable space is defined as below:

$$h_i = x_i^T (X^T X)^{-1} x_i (i = 1, \ldots, n)$$

where, x_i is the LV vector of the investigated compound and X is the model matrix derived from the training set LV values.

The warning leverage value (h^*) is defined as $3(K + 1)/n$, where, K is the number of independent variables. When h value of a molecule is lower than h^*, the probability of accordance between calculated and experimental values is as high as that of the molecules in the training set (4). A compound with $h_i > h^*$ will reinforce the model if the compound is in the training set. But such a compound in the testing set implies that it is structurally distant from chemicals in the calibration set and its predicted data may be unreliable. However, this compound may not appear to be an outlier because its residuals may be low. Thus the leverage and the standardized residual should be used simultaneously for the description of the AD of the expanded model. It must be noted that the outliers are objects that emerge to break the pattern or grouping shown by the majority of the objects. Presence of outliers in the studied data set is more the rule than the exception for real world data. The reasons for outliers are different, such as instrument failure, non-representative sampling, formatting errors and observations stemming from other populations. Most usual multivariate regression techniques are sensitive to outliers because of the fact that they are based on least squares or similar criteria where even one outlier can have an illogically large effect on the accuracy of developed model and decline the model. Therefore, it is essential to (a) recognize outliers and (b) make a decision whether the outliers should be included or omitted in the modeling step.

Applicability of domain for the developed PLS model is shown in Fig. 3 Response outliers are compounds that have standard residual points greater than the two standard deviation units. Influential compounds are points with leverage value higher than the warning leverage limit. As can be seen in Fig. 3 all studied molecules in training and test sets lie in application domain of developed model.

Suggestion of potent compounds

As a final point, one could dispute that how researchers can interpret the developed PLS model or how developed model can be used to propose novel aldehyde derivatives with improved activity. In other words, what does the developed QSAR model mean to medicinal chemists? As discussed above, the calculated latent variables do not mean physico-chemically, but they may be employed for building statistical models which help the medicinal chemist limit the number of compounds to be synthesized. For instance, medicinal chemist can propose a training set comprised of molecules which have the characters of two or more chemical classes with the smallest amount of similarity. Then one can use the developed models to predict the activity of the proposed molecules. This practice may lead to the introduction of biologically active molecules.

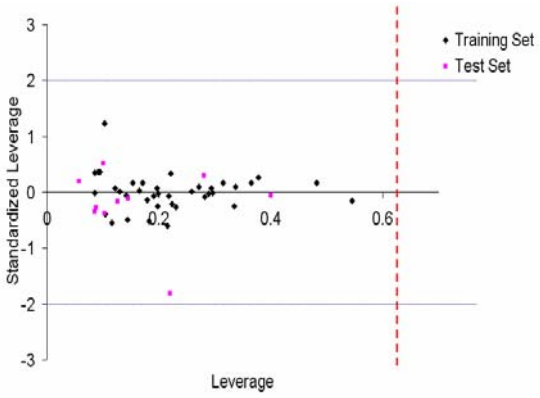
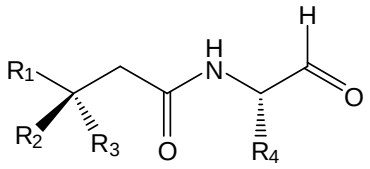


Fig. 3. William's plot of generated PLS-based QSAR model

Table 4. Structure and details of suggested antagonists



Compound	R_1	R_2	R_3	R_4	Activity
S1	$PhCH_2$	Me	i-Bu	Et	6.12
S2	$PhCH_2$	n-Pr	i-Bu	CH_2-i-Pr	6.13
S3	$PhCH_2$	Me	i-Bu	CH_2-i-Pr	6.24
S4	$PhCH_2$	n-Pr	i-Pr	Me	6.31
S5	$PhCH_2$	Me	i-Pr	n-Bu	6.11
S6	$PhCH_2$	n-Pr	i-Pr	n-Bu	6.22

Since experimental and computed activities of compounds used in the model development step showed good correlation, developed QSAR model was employed to calculate inhibitory activities of suggested compounds. Structures of novel antagonists of catK may then be suggested and their activities could be evaluated by using the developed model. Compounds owning the general structure similar to the investigated compounds containing various substituents may give rise to the novel compounds. Structures of these novel ligands as well as their LVs were generated. Consequentlye, using calculated LVs and developed model, activities of proposed ligands were calculated.

The general structures of six suggested compounds and details of their calculated activities are reported in Table 4 The suggested compounds are combination of the most potent compounds of Table 1 All suggested compounds were submitted to activity evaluation using developed QSAR model. The relative high predicted activity of suggested compounds could be further confirmed by synthesising their chemical entities.

CONCLUSION

Quantitative relationship between molecular structure and human catK inhibitory activity of a series of aldehyde derivatives was discovered by one of the most commonly used regression methods, PLS. Evaluation of a test set of ten compounds with the developed PLS model revealed that this model is reliable and has a good predictability. Since the QSAR study was performed on the basis of theoretical descriptors calculated completely from molecular structure, the proposed model could potentially provide useful information about the activity of the studied compounds. Various tests and criteria such as leave-one-out cross validation, leave-many-out cross validation, and also criteria suggested by Tropsha were employed to examine the predictability and robustness of the developed model. This model could explain and predict 90 % and 83 % of variances in the pIC_{50} data, respectively.

REFERENCES

1. Arkan E, Shahlaei M, Pourhossein A, Fakhri K, Fassihi A. Validated QSAR analysis of some diaryl substituted pyrazoles as CCR2 inhibitors by various linear and nonlinear multivariate chemometrics methods. Eur J Med Chem. 2010;45:3394-3406.
2. Saghaie L, Shahlaei M, Fassihi A, Madadkar-Sobhani A, Gholivand M, Pourhossein A. QSAR Analysis for Some Diaryl-substituted Pyrazoles as CCR2 Inhibitors by GA-Stepwise MLR. Chem Biol Drug Des. 2011;77:75-85.
3. Saghaie L, Shahlaei M, Madadkar-Sobhani A, Fassihi A Application of partial least squares and radial basis function neural networks in multivariate imaging analysis-quantitative structure activity relationship: Study of cyclin dependent kinase 4 inhibitors. J Mol Graph Model. 2010;29:518-528.
4. Shahlaei M, Fassihi A, Nezami A. QSAR Study of some 5-methyl/trifluoromethoxy-1H-indole-2,3-dione-3-thiosemicarbazone derivatives as anti-tubercular agents. Res Pharm Sci. 2009;4:123-131.
5. Shahlaei M, Fassihi A, Saghaie L. Application of

PC-ANN and PC-LS-SVM in QSAR of CCR1 antagonist compounds: A comparative study. Eur J Med Chem. 2010;45:1572-1582.
6. Shahlaei M, Sabet R, Ziari MB, Moeinifard B, Fassihi A, Karbakhsh R. QSAR study of anthranilic acid sulfonamides as inhibitors of methionine aminopeptidase-2 using LS-SVM and GRNN based on principal components. Eur J Med Chem. 2010;45:4499-4508.
7. Bromme D, Okamoto K. Human cathepsin O2, a novel cysteine protease highly expressed in osteoclastomas. Biol Chem Hoppe Seyler. 1995;376:379-384.
8. Gelb BD, Shi GP, Chapman HA, Desnick RJ. Pycnodysostosis, a lysosomal disease caused by cathepsin K deficiency. Science. 1996;273:1236-1238.
9. Saftig P, Hunziker E, Wehmeyer O, Jones S, Boyde A, Rommerskirch W, et al Impaired osteoclastic bone resorption leads to osteopetrosis in cathepsin-K-deficient mice. Proc Natl Acad Sci USA. 1998;95:13453-13458.
10. Alves MFM, Puzer L, Cotrin SS, Juliano MA, Juliano L, Brömme D, et al S3 to S3′ subsite specificity of recombinant human cathepsin K and development of selective internally quenched fluorescent substrates. Biochem J. 2003;373:981-986.
11. Robichaud J, Oballa R, Prasit P, Falgueyret JP, Percival MD, Wesolowski G, et al A novel class of nonpeptidic biaryl inhibitors of human cathepsin K. J Med Chem. 2003;46:3709-3727.
12. Boros EE, Deaton DN, Hassell AM, McFadyen RB, Miller AB, Miller LR, et al Exploration of the P2-P3 SAR of aldehyde cathepsin K inhibitors. Bioorg Med Chem Lett. 2004;14:3425-3429.
13. Catalano JG, Deaton DN, Furfine ES, Hassell AM, McFadyen RB, Miller AB, et al Exploration of the P1 SAR of aldehyde cathepsin K inhibitors. Bioorg Med Chem Lett. 2004;14:275-278.
14. Hyperchem. Molecular Modeling System. In: Developed by Hyper Cube Inc. and Auto Desk, Inc.
15. Todeschini R, Consonni V, Mauri A, Pavan M. DRAGON software. Milano, Italy: 2002.
16. Todeschini R, Consonni V. Handbook of Molecular Descriptors. Weinheim, Germany: Wiley-VCH; 2000.
17. Tropsha A, Gramatica P, Gombar V, The importance of being Eearnest: Validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci. 2003;22:69-77.
18. Kennard R, Stone L. Computer Aided Design of Experiments. Technometrics. 1969;11:137-148.
19. Wold H. Estimation of Principal Components and Related Methods by Iterative Least Squares. In: Krishnaiah PR, Editor. Multivariate Analysis. New York: Academic Press, 1966. p. 391-420.
20. Roy PP, Roy K. On some aspects of variable selection for partial least squares regression models. QSAR Comb. Sci. 2008;27:302-313.
21. Gramatica P, Papa E. QSAR modeling of bioconcentration factor by theoretical molecular descriptors. QSAR Comb Sci. 2003;22:374-385.
22. Wold S. Validation of QSARs. Quant Struct-Act Relat. 1991;10:191-193.
23. Zhang W, Tropsha A. Novel variable selection quantitative structure–property relationship approach based on the k-nearest-neighbor principle. J Chem Inf Comput Sci. 2000;40:185-194.
24. Atkinson AC. Plots, Transformations and Regression. UK: Clarendon Press; 1985.