Research in Pharmaceutical Sciences, December 2014; 9(6): 471-488
Received: Sep 2013
Accepted: Nov 2013

School of Pharmacy & Pharmaceutical Sciences
Isfahan University of Medical Sciences

Original Article

Prediction of p38 map kinase inhibitory activity of 3, 4-dihydropyrido [3, 2-d] pyrimidone derivatives using an expert system based on principal component analysis and least square support vector machine

M. Shahlaei[1] and L. Saghaie[2,*]

[1]*Nano Drug Delivey Research Center, Faculty of Pharmacy, Kermanshah University of Medical Sciences, Kermanshah, I.R. Iran.*
[2]*Department of Medicinal Chemistry, Bioinformatic Research Center and Isfahan Pharmaceutical Sciences Research Center, Faculty of Pharmacy, Isfahan University of Medical Sciences, Isfahan, I.R. Iran.*

Abstract

A quantitative structure–activity relationship (QSAR) study is suggested for the prediction of biological activity (pIC_{50}) of 3, 4-dihydropyrido [3 ,2-d] pyrimidone derivatives as p38 inhibitors. Modeling of the biological activities of compounds of interest as a function of molecular structures was established by means of principal component analysis (PCA) and least square support vector machine (LS-SVM) methods. The results showed that the pIC_{50} values calculated by LS-SVM are in good agreement with the experimental data, and the performance of the LS-SVM regression model is superior to the PCA-based model. The developed LS-SVM model was applied for the prediction of the biological activities of pyrimidone derivatives, which were not in the modeling procedure. The resulted model showed high prediction ability with root mean square error of prediction of 0.460 for LS-SVM. The study provided a novel and effective approach for predicting biological activities of 3, 4-dihydropyrido [3,2-d] pyrimidone derivatives as p38 inhibitors and disclosed that LS-SVM can be used as a powerful chemometrics tool for QSAR studies.

Keywords: Principal component analysis; Least square support vector machine; p38 inhibitory activity; QSAR

INTRODUCTION

Antagonists of p38 mitogen activated protein (MAP) kinase inhibit the production of proinflammatory cytokines, for instance, tumor necrosis factor-α (TNF-α) and interleukin-1β (IL-1β), whose accumulation initiates a cascade of events leading to inflammation and tissue destruction in diseases such as rheumatoid arthritis (RA) (1), Crohn's disease (2), inflammatory bowel syndrome, and psoriasis. The inhibition of TNF-α and IL-1β presents a useful therapeutic strategy to suppress the inflammation and prevent joint damage caused by RA, as shown by the newer biologic therapies for RA (etanercept, infliximab, adalimumab, and anakinra) that target these cytokines. p38 MAP kinase is a member of a family of serine–threonine kinases that are activated by dual phosphorylation of a threonine glycine and tyrosine (TGY) motif (3). This phosphorylation is performed by dual specificity kinases (MKK3 and MKK6 (4)) in response to extracellular stimuli such as osmotic shock, endotoxins (lipopolysacharide, LPS), UV light or cytokines (5).

Therefore, pharmacological inhibition of p38 kinase is a potential way for treatment of inflammatory conditions due to excessive cytokine production.It has been nearly 40 years since the quantitative structure–activity relationship (QSAR) the quantitative structure–property relationship (QSPR) paradigm first found their way into the practice of medicinal chemistry, analytical chemistry, toxicology, and ultimately most disciplines of chemistry (6). These methods are statistical models of dependent variable (a biological activity or a physical property) in

*Corresponding author: L. Saghaie
Tel: 0098 31 37922565, Fax: 0098 31 36680011
Email: saghaie@pharm.mui.ac.ir

terms of computational descriptors. The developed model is helpful for understanding the factors controlling dependent variable (bioactivity) and for designing new potent and efficient molecules (7). QSAR models could describe large number of relationships between structural descriptors of drug like compounds and their bioactivities (8-20).

Complex physiological molecular processes and systems need a multitude of measured variables and signals for their characterization. With the advent of modern computers, hardware and software, sophisticated molecular descriptor calculation software, efficient instrument and other molecular equipments, large masses of biologically relevant molecular data are gathered at an ever increasing pace. The acquisition of large masses of biologically relevant molecular data results in exploratory and interpretative challenges.

The abundance of biological and physiological molecular data is not in itself a guarantee for obtaining helpful information on major events taking place in active site of target- in system of interest. On the contrary, data from the biological field require to be processed and analyzed, in order to highlight the useful information from the experimental measurements. Since these data regularly are highly multivariate in nature one must employ data analysis methods which are able to handle the challenges inherent in masses of data, notably noise, collinearities, and missing data. Only with a careful data analysis researchers will be able to address central questions such as how to modify molecular structure of investigated compounds interacted with a given protein in order to improve their biological performance, or to understand why a certain protein is particularly sensitive to exposure to a certain group of molecular structures.

With the advancement of the technology and industry, the interest in nonlinear modeling and the development of mathematical tools to determine the behavior of biological phenomena have grown significantly, since the existing techniques for linear modeling cannot reproduce the full range of dynamic behaviors of real systems such as interactions between the ligand and receptor. The problem of identification of important features in biological systems is an area of research interest and has gained increasing significance.

Based on the structural risk minimization principle, an excellent machine learning method of support vector machine (SVM) was first reported by Vapnik and coworkers (21). Compared with other machine learning methods, SVM has many attractive characteristics, including the absence of local minima, its speed and scalability and its ability to condense information contained in the training set (22).

Hence, it can be said that a promising approach in nonlinear identification applications are the support vector machines. As a new and powerful modeling tool, SVM has been extensively used to QSAR research. However, when SVM is utilized to QSAR modeling, one of the most important problems is the selection of optimal features subset. It is well defined that large numbers of input variable vectors fed to SVM can increase computational complexity (23), suffer from the curse of dimensionality and the risk of over-fitting. In contrast, a few input variable vectors that are not relevant to biological activity can result in bad generalization performance and accuracy. Consequently, the selection of optimized input variable vector subset is essential to speed up computation and to improve the generalization performance of SVM. Least square support vector machine (LS-SVM) proposed by Suykens and Vandewalle (24) is a modification of the standard SVM. Unlike artificial neural network based nonlinear models, LS-SVM possesses prominent advantages:

over-fitting is unlikely to occur by adopting the structural risk minimization (SRM) principle, and the global optimal solution can be uniquely obtained by solving a set of linear equations (25). This study focuses on employing LS-SVM based on principal component analysis (PCA) to perform the pattern recognition of a class of potent 3, 4-dihydropyrido [3,2-d] pyrimidone inhibitors of p38a MAP kinase. Although LS-SVM based on PCA has been applied for quantification in

different medicinal and pharmaceutical methods (26,27), to the best of our knowledge, the application of LS-SVM based on PCA in pattern recognition for QSAR is very rare (14,19).

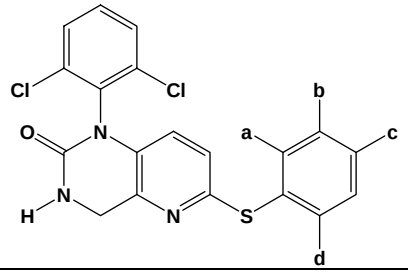
MATERIALS AND METHODS

Data preparation

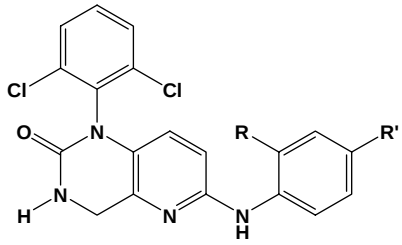
In vitro biological activity data used in this study were p38 inhibitory activity (in terms of-log IC_{50}) of a set of forty 3, 4-dihydropyrido [3,2-d] pyrimidone derivatives selected from literature (28). General chemical structures and the structural details of these compounds and also their activities are reported in Table 1.

The two dimensional structures of studied molecules were built using ChemDraw 7.0 (ChemDraw Ultra, 1985–2001; CambridgeSoft, Cambridge, MA, USA), and then converted to 3D structure using Chem3D Ultera 7.0 (Chem3D Ultra, 1985–2001; CambridgeSoft, Cambridge, MA, USA). Prior to the computation of the various molecular descriptors, all the structures were drawn and pre-optimized using the semi empirical quantum-chemical routine of AM1 implemented in Hyperchem (29).

Table 1. Structures and details of the molecules investigated.



Compound	a	b	c	d	pIC_{50}
1	H	F	H	Cl	7.958
2*	H	F	H	F	7.602
3	H	Cl	H	Cl	7.494
4	Cl	H	H	Cl	7.309
5*	H	H	H	Cl	7.026
6*	H	H	H	Me	6.958
7	H	H	H	Br	6.920
8	H	F	H	NH_2	6.698
9	H	H	H	H	6.568
10	Me	H	H	Me	6.585
11	Cl	H	Cl	H	6.455
12*	F	H	H	H	6.346
13	Cl	H	H	H	6.288
14	H	H	H	NH_2	5.847
15	H	H	H	CF_3	5.732
16	H	H	H	OMe	5.701
17*	H	H	Cl	H	5.659
18	H	H	CF_3	H	5.741



Compound	R	R'	pIC_{50}
19	H	H	6.946
20	F	F	6.468

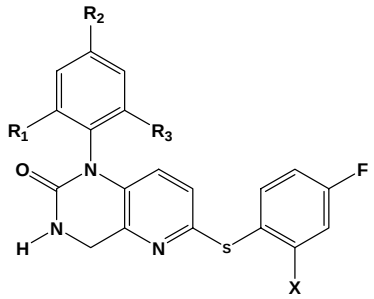


Table 1. (Continued)

Compound	R_1	R_2	R_3	pIC_{50}
21	Me	COOMe	Cl	7.443
22	Et	H	Cl	7.420
23*	Me	COMe	Cl	7.327
24*	F	H	Cl	7.318
25	Me	H	Cl	7.267
26	Me	NO_2	Cl	7.180
27	COOMe	H	Cl	7.096
28	OMe	H	Me	7.000
29	CF3	H	Cl	6.958
30	H	H	Me	6.744
31	Cl	CF_3	Cl	6.638
32	Cl	COOMe	Cl	8.045
33	Cl	COMe	Cl	7.795
34*	Me	H	Cl	7.721
35	Cl	COONH2	Cl	7.638
36	Cl	H	Cl	7.602
37	Cl	CONH–	Cl	7.096
38	Cl	COOH	Cl	7.017
39	H	H	Cl	7.161
40	H	H	H	6.221

* Molecules assigned as test set by Kennard and Stone algorithm

A total of 297 molecular theoretical descriptors of different kinds were used as input to describe compound chemical diversity. Molecular descriptors were computed using the software *DRAGON* (30). The descriptor groups were constitutional, functional groups, topological, and geometrical. Molecular descriptor meanings and their calculation procedure are summarized in the software *DRAGON*, and explained in detail, with related literature references, in the *Handbook of Molecular Descriptors* by Todeschini and coworkers (31).

Kennard and Stone algorithm was used to split the entire dataset of interest into two parts (around 80% as training set and 20% as test set), training set for constructing models and test set for assessing the predictive power of these constructed models.This is a classic technique to extract a representative set of molecules from a given data set. In this technique the molecules are selected consecutively. The first two objects are chosen by selecting the two farthest apart from each other. The third sample chosen is the one farthest from the first two objects, etc. Supposing that m objects have already been selected (m<n), the (m+1)th sample in the calibration set is chosen using the following criterion:

$$\max_{m<r\leq n}(\min(d_{1r}, d_{2r}, ...d_{mr})$$

Where, n stands for the number of samples in the training set, djr, j=1,..., m are the squared Euclidean distances from a candidate sample r, not yet included in the representative set, to the m samples already included in the representative set. One more benefit of the Kennard–Stone method is that it may be used to any matrix of predictors; there are no restrictions regarding the matrix multicollinearity. The other advantage is that the test molecules all fall inside the measured region and the training set molecules map the measured region of the input variable space completely with respect to the induced metric.

Principal component analysis

Principal component analysis is used for reducing the dimensionality of the dataset. The data matrix *X* consists of *N* molecules represented by *M* descriptors (297 columns).

Prior to PCA in a typical QSAR study the matrix of dataset is regularly pre-processed by means of two operations: mean-centering and scaling to unit variance. With PCA, *X* matrix is decomposed into the product of two matrices, the (N × A) score matrix, *T*, times the (A × K) loading matrix, P′, plus an (N × K) "noise" matrix of residuals, E.

$$X = TP^T + E \quad (1)$$

In fact, PCA is based on the concentration of the original data variance into a small number of principal components (PCs) by means of mathematical transformation. As a result, the first PC describes the maximum information from the data; the second PC describes the maximum amount of the residual variance. Each successive PC is an orthogonal combination of the original descriptors such that it covers the maximum of the variance not accounted for by the previous components.

Least square support vector machine

Recently, the SVM, based on statistical learning theory, as a powerful new tool for data classification and function estimation, has been developed (21). SVM maps input data into a high-dimensional feature space where it may become linearly separable. In recent years SVM has been applied to an extensive variety of domains such as pattern recognition and object detection (32), function estimation (33), etc.

One reason that SVM often performs better than earlier methods is that SVM was designed to minimize structural risk whereas previous techniques were usually based on minimization of empirical risk (34). So SVM is usually less vulnerable to the overfitting problem (35). Especially, Suykens and Vandewalle (35) proposed a modified version of SVM called least squares SVM (LS-SVM), which resulted in a set of linear equations instead of a quadratic programming problem, which can extend the application of the SVM. Excellent introductions to SVM appear in Refs. (32,33). The theory of LS-SVM has also been described clearly by Suykens and Vandewalle (35). For this reason, we will only briefly explain the main idea of LS-SVM and the differences between SVM and LS-SVM.

In LS-SVM, as in most linear regression models, linear estimation is carried out between the regressors (x) and the dependent variable (y): y=wTx+b, where w is the regression coefficient. The regression is calculated by minimizing a cost function (C) containing a penalized regression error, as follows:

$$C = \frac{1}{2} w^T w + \frac{1}{2} \sum_{i=1}^{N} e_i^2 \quad (2)$$

Subjected to

$$y_i = w^T X_i + b + e_i \quad \text{i=1 to N} \quad (3)$$

The first part of this cost function is a weight decay that is used to regularize weight sizes and penalize large weights. The second part of Eq. (2) is the regression error for all training data. After analyzing Eq. (2) and its restriction given by Eq. (3), a typical problem of convex optimization is formulated, this can be solved by using the Lagrange multipliers method. The LS-SVM model can be expressed as:

$$y = \sum_{i=1}^{n} a_i k(x_i, x) + b \quad (4)$$

where, k (x, x_i) is the kernel function, x_i is the input vector, a_i is Lagrange multipliers called support value, b is bias term. In this study, the Gaussian kernel was used as kernel function and a cross validation procedure was used to tune the optimized values of the two parameters σ and γ.

Validation of quantitative structure–activity relationship models

There are several tools to estimate and calculate the accuracy and also the validity of the proposed QSAR model and as well the impacts of the preprocessing steps. Here, we have employed several techniques to ensure the effectiveness of the regression methods. Some of the common parameters used for checking the predictability of proposed models 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 for each model as follows:

$$RMSE = [1/n\sum_{i=1}^{n}(y_i - \hat{y}_i)^2]^{\frac{1}{2}} \quad (5)$$

$$R^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad (6)$$

$$PRESS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (7)$$

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

The actual efficacy of the generated QSAR models is not just their capability to reproduce known data, confirmed by their fitting power (R^2), but is mainly their feasibility of predictive application. Hence, the QSAR model estimations were carried out maximizing the explained variance in prediction, assigned by the leave-one-out cross-validated correlation coefficient, Q^2.

Leave-one-out cross-validation (LOOCV) involves using a single molecule from the original samdataset as the validation dset, and the remaining o molecules as the training dset. This is repeated such that each molecule in the original datset is used once as the validation molecule. This is the same as a K-fold cross-validation with K being equal to the number of molecules in the original sadataset.

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 very insists that models with high apparent predictive ability, highlighted only by internal validation methods, can be unpredictive when applied on new compounds not employed in developing the model. Thus, for a stronger estimation of developed model usability for prediction on new chemicals, external validation of the models should always be carried out (36).

Also, the predictive ability of the regression model generated on the training set molecules is estimated on the predictions of the testing set compounds, by the R^2_p (R^2 for test set) that is defined as follows (36):

$$R_p^2 = 1 - \frac{\sum_{i=1}^{test}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{test}(y_i - \bar{y}_{tr})^2} \quad (8)$$

where, $\bar{y}_{tr}$ is the averaged value of the bioactivity for the training set; the summations cover all the molecules in the testing set.

An accepted technique employed by researchers to defend their generated models against the danger of chance correlation between dependent and independent variables has been y-randomization that is, accidental correlation without any predictability for the developed model. y- randomization is a technique that said to be “probably the most powerful validation procedure” (38). Confirmation that a developed model is well established and not just the result of chance correlation is given by the new models obtained on the data set with shuffled bioactivity. If such models present considerably lower R^2 and Q^2 than the original model, it is suggest that QSAR models are not consequence of the chance correlation.

RESULTS

Principal component analysis summarizes the information residing in the initial data, i.e., in our case the theoretical descriptors, into a new variables which may be more easily overviewed and applied. The original multi-dimensional space, defined by the calculated descriptors is contracted into a few descriptive dimensions, represented principal components, which denote the main variation in the data.

Each PC can be displayed graphically and be analyzed separately, and its meaning may often be interpreted according to simple chemical and/or biological fundamental factors, such as, number of carbon atoms, molecular weight, volume, or something else. The greatest amount of variability of the original data set is implied by the first PC, and the second PC describes the maximum variances of the residual data set.

Then, the third one will explain the most important variability of the next residual data set, and so on. According to the theory of least squares, the eigenvectors of all PCs are orthogonal to each other in multi-dimension data space. Generally speaking, only *p* PCs are enough to account for the most variance in an *m*-dimensional data set, where *p* is the number of important PCs of the data set, and m means the number of all the PCs in the data set of interest.

It is obvious that *p* is less than m. So PCA is generally regarded as a data reduction method. That is to say, a multi-dimensional data set can be projected to a lower dimension data space without loss most of the information of the original data set by PCA (39).

To explore the structure of pool of calculated descriptors, PCA was adopted on all the calculated descriptors, then 40 principal components (PCs) were generated. The variances explained by the first fourteen PCs are shown in Fig. 1. It can be found that the PC1 could explain more than 20% variance of all calculated descriptors, and variances explained by the latter PCs gradually decreased.

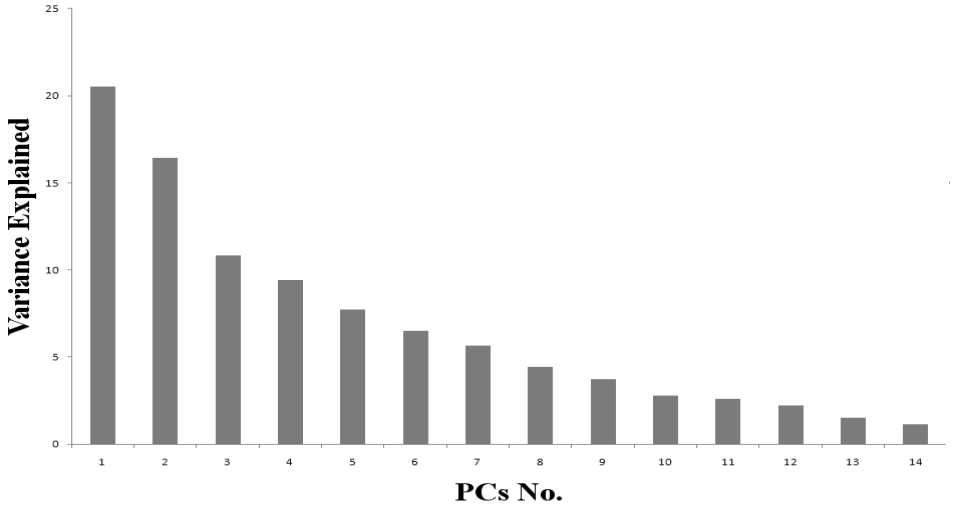


Fig. 1. Variance explained by the first fourteen principal components.

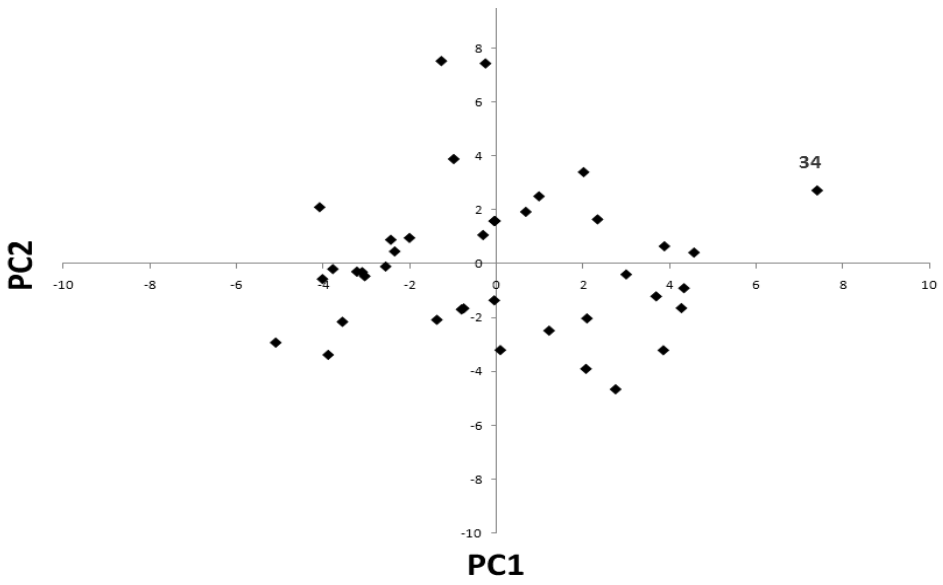


Fig. 2. Score plot of samples based on the first two principal components.

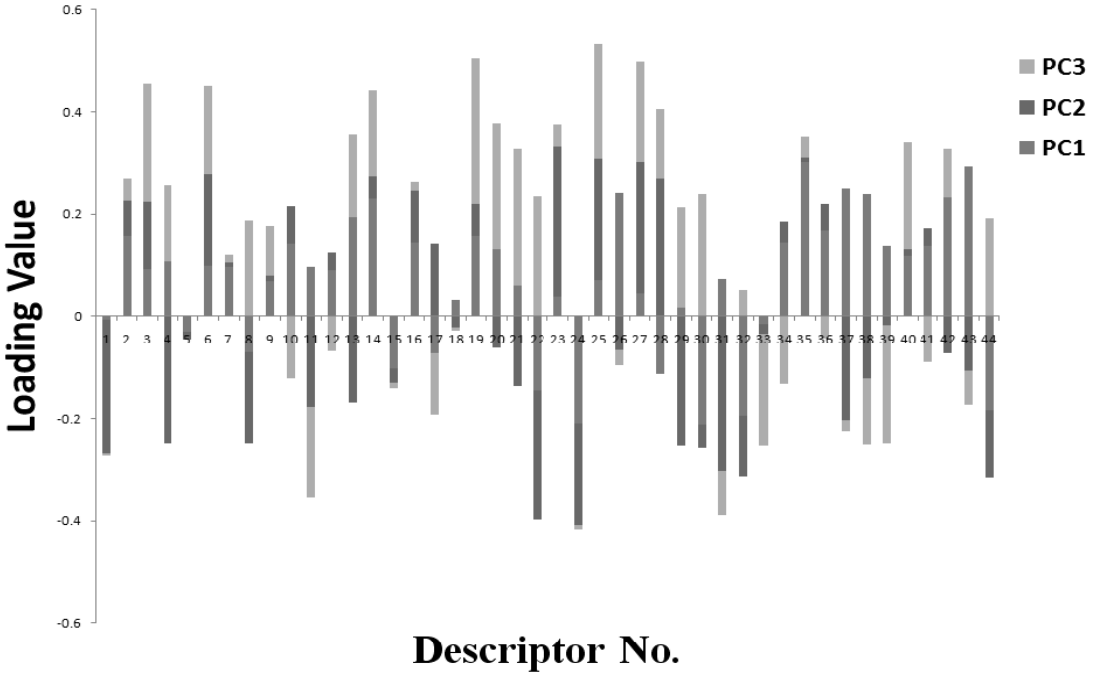


Fig. 3. Loading value analysis of the first 3 principal components.

In total, the accumulative variance of the first fourteen PCs was up to 95%. So, it could be concluded that the first fourteen PCs could explain most of the variance of the calculated descriptors. After PCA procedure, a series of new variables (PCs) were generated, so every molecule could be denoted with the PCs, and the score plot is a description of molecules in the new defined space by PCs. As the PCA had compressed most variance of the calculated descriptors into the first several PCs, the score plot of the first PCs may reveal important information of recognition. Fig. 2 is the score plot of the first two PCs. It could be seen that most molecules were clustered together (the PC1 value of compound 34 is larger than other compounds but in the following we will be seen that this compound is not an outlier). So, it can be concluded that the first PCs contained the main characteristic for recognition of multi-dimensional data set include calculated descriptors, and the major information of descriptors had been compressed into the first PCs by PCA.

As mentioned above, the first PCs were of great importance in explanation of variance in calculated descriptors. But how were these PCs generated from the 297 descriptors? It was very meaningful to explore the structure of the pool of descriptors.

As PCs were constructed with a linear combination of the descriptors, the relationship between PCs and original descriptors could be uncovered by a group of weighting coefficients, which were also named as loading weights. Fig. 3 shows the loading weight analysis of the first 3 PCs. The greater the absolute value of a coefficient (loading weight), leads to the greater the weight of the descriptor in the PCA-based model. On other word, the high values of the coefficients (loading values) show the statistical significance of the descriptors in the final PCA-based model.

It indicated that the 4 descriptors (with the higher absolute loading value in the 3 first PCs) played a very important role for construction of the PCs (Fig. 3). In other words the 3 first very important PCs were greatly affected by the 4 descriptors, include descriptor number 13 (TIE) with an absolute loading value in the first 3 PCs equals to 0.523, descriptor number 22 (G(S,.Cl)) with an absolute loading value in the first 3 PCs equals to 0.632, descriptor number 25 (X1A) with an absolute loading value in the first 3 PCs equals to 0.534, and descriptor number 28 (X4A) with an absolute loading value in the first 3 PCs equals to 0.518.

Table 2. The most important descriptors in the first 3 PCs

Descriptor	Definition	Descriptor class	Absolute loading value
TIE	E-state topological parameter	Geometrical descriptors	0.523
G (S..Cl)	Sum of geometrical distances between S and Cl	Geometrical descriptors	0.632
X1A	average connectivity index chi-1	Topological descriptors	0.534
X4A	average connectivity index chi-2	Topological descriptors	0.518

Table 3. The experimental pIC_{50} and the predicted values of the training and test sets and values of relative error of prediction by each model.

Compound	Experimental pIC_{50}	Predicted pIC_{50} by PCR	RE (PCR)	Predicted pIC_{50} by LS-SVM	RE (LS-SVM)
1	7.959	7.710	0.031	7.246	0.090
2	7.602	7.361	0.032	7.213	0.051
3	7.495	7.402	0.012	6.986	0.068
4	7.310	6.907	0.055	7.044	0.036
5	7.027	6.688	0.048	6.736	0.041
6	6.959	6.431	0.076	6.979	-0.003
7	6.921	6.851	0.010	6.611	0.045
8	6.699	6.600	0.015	7.057	-0.053
9	6.569	6.162	0.062	6.677	-0.017
10	6.585	6.309	0.042	6.722	-0.021
11	6.456	7.151	-0.108	6.910	-0.070
12	6.347	6.504	-0.025	6.789	-0.070
13	6.288	6.210	0.012	6.691	-0.064
14	5.848	5.597	0.043	6.416	-0.097
15	5.733	6.182	-0.078	6.444	-0.124
16	5.701	6.119	-0.073	6.389	-0.121
17	5.660	6.595	-0.165	6.646	-0.174
18	5.741	6.375	-0.110	6.423	-0.119
19	6.947	7.203	-0.037	6.864	0.012
20	6.469	7.261	-0.123	6.824	-0.055
21	7.444	7.569	-0.017	7.147	0.040
22	7.420	7.099	0.043	7.047	0.050
23	7.328	7.345	-0.002	7.122	0.028
24	7.319	7.411	-0.013	7.202	0.016
25	7.268	7.212	0.008	7.066	0.028
26	7.180	7.247	-0.009	7.133	0.007
27	7.097	7.338	-0.034	7.008	0.012
28	7.000	6.850	0.021	6.943	0.008
29	6.959	6.601	0.051	6.913	0.006
30	6.745	6.944	-0.030	6.960	-0.032
31	6.638	7.205	-0.085	6.758	-0.018
32	8.046	7.636	0.051	7.405	0.080
33	7.796	7.401	0.051	7.222	0.074
34	7.721	7.138	0.076	7.275	0.058
35	7.638	7.317	0.042	7.288	0.046
36	7.602	7.329	0.036	7.076	0.069
37	7.097	6.953	0.020	6.983	0.016
38	7.018	6.455	0.080	7.092	-0.011
39	7.161	7.153	0.001	6.998	0.023
40	6.222	6.699	-0.077	6.705	-0.078

Table 2 shows the selected descriptors, the absolute loading value in the first 3 PCs, their definition and class. Because TIE and G(S..Cl) belongs to the geometrical group of descriptors, some geometrical properties including angles between atoms, dihedrals angles, and atomic distances are probably important features in the effectiveness of compounds of interest in this study as p38 inhibitors.

Both X1A and X4A are belonging to topological descriptor class. Presence of these descriptors in the most important descriptors, basically accounts for size, shape, and branching, thus steric contribution to biological activity. With respect to the Fig. 3, the absolute positive sign of both X1A and X4A indicate that biological activity increases with an increase in the magnitude of size as well as branching of molecules.

After acquiring PCs, linear regression with stepwise factor selection was performed. The obtained model consists of 7 terms, with one constant term and 6 terms based on different PCs. The model is given by the following equation:

$$\text{Predicted } pIC_{50} = 6.913 + 0.156*PC_7 - 0.113*PC_5 - 0.118*PC_6 + 0.067*PC_1 + 0.238*PC_{14} - 0.162*PC_{10} \quad (8)$$

The calculated pIC_{50} for each molecule and relative error (RE) of prediction by model are summarized in Table 3. Experimental versus predicted values for pIC_{50}s of training and test sets, obtained by the principal component regression (PCR) modeling, is shown graphically in Fig. 4A.The residuals of the predicted values obtained by PCR are plotted against the experimental values in Fig. 5A. The spread of residuals in both sides of zero line showed that any systematic error doesn't exist in the development of the linear regression method.The cross-validation method employed was eliminating only one molecule at a time and then performing PCR on the remaining of training set (leave-one-out method). The activity of the left-out molecule was predicted by using the developed regression model. This procedure was repeated until each molecule in the training set had been gone out once. Experimental versus predicted values for pIC_{50} values for training and test set, obtained by the PCR modeling, is shown graphically in Fig. 4a.

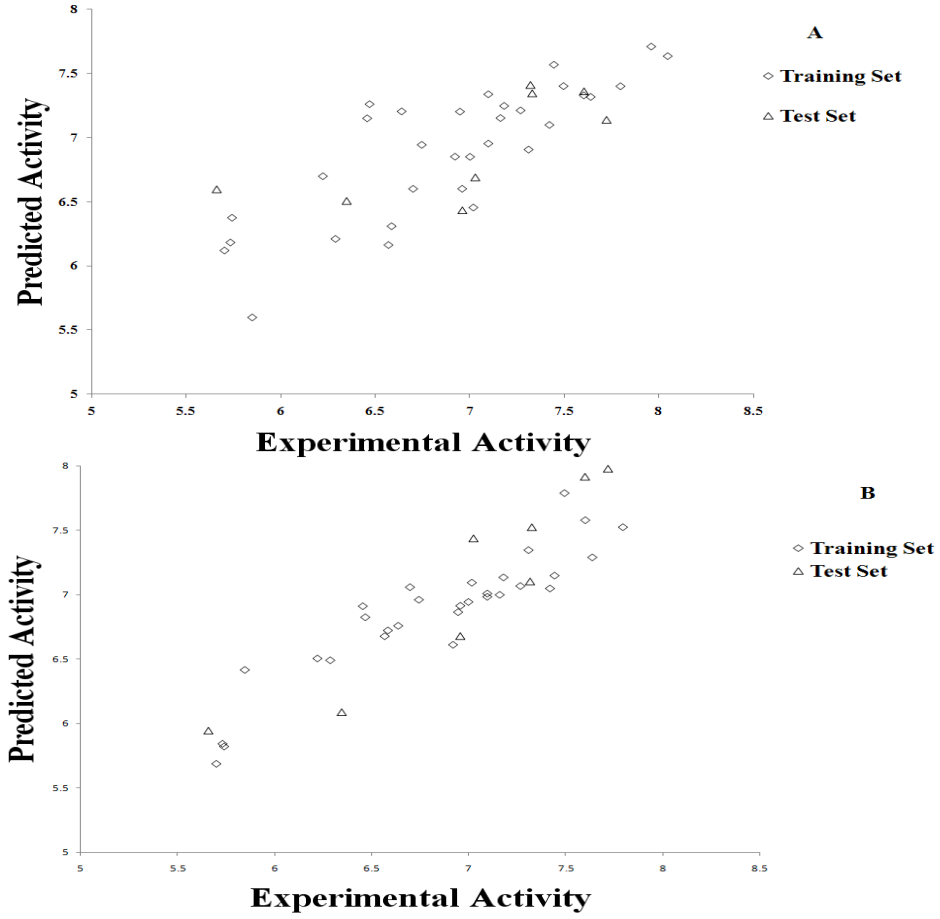


Fig. 4. pIC_{50} predicted values versus experimental values for training and test setsby: A; PCR model, B; least square support vector machine.

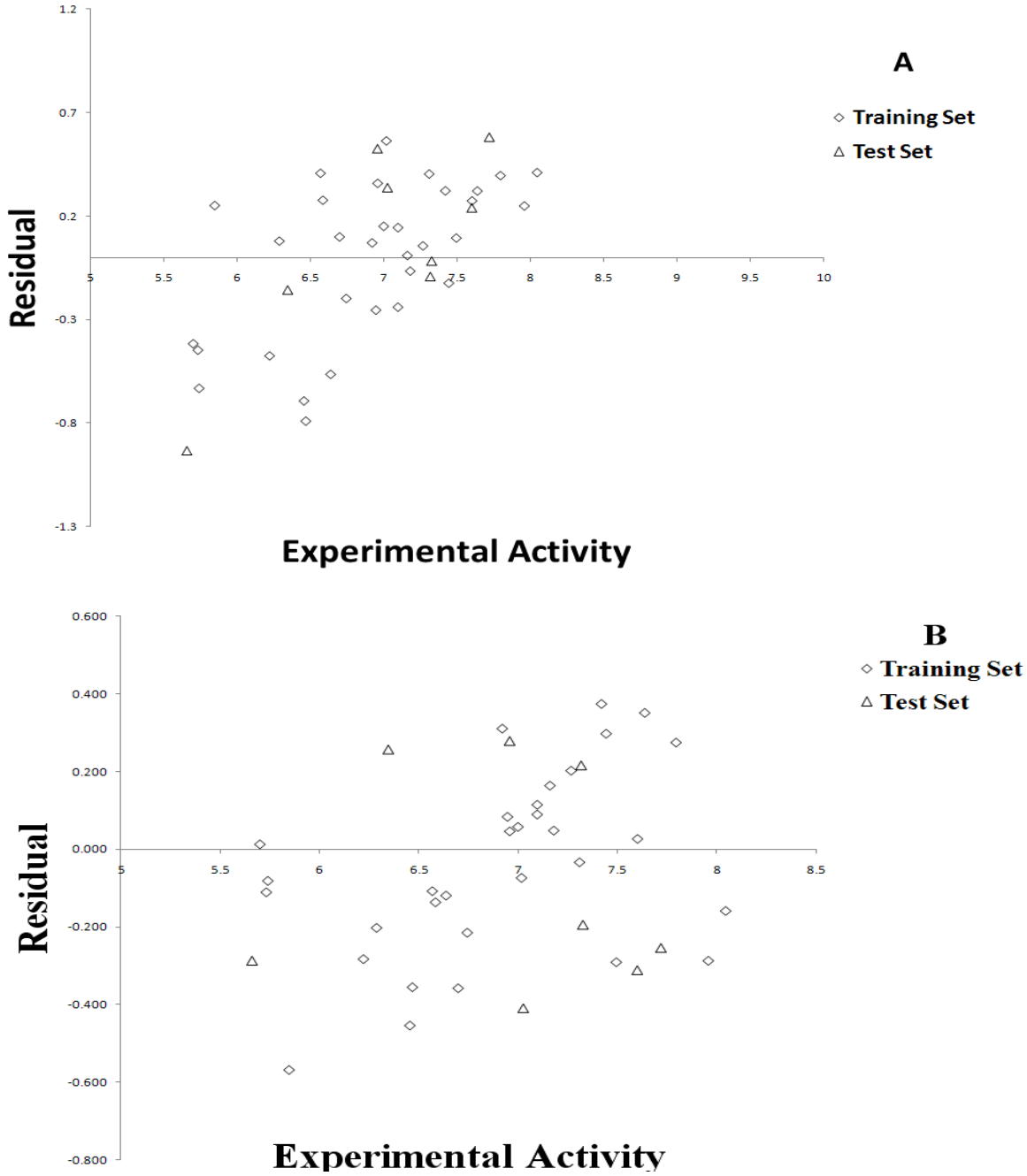


Fig. 5. Scatter plots of the residuals *vs*. experimental activity for A; PCR B; least square support vector machine .

Table 4. Statistic parameters of developed models**.**

Statistics	PCR		LS-SVM	
	Training set	Test set	Training set	Test set
N	32	8	32	8
R^2	0.659	0.506	0.859	0.865
RMSET	0.413		0.256	
RMSEP		0.460		0.282
PRESS	4.274	1.700	1.839	0.639
Q^2	0.523		0.846	
$RMSE_{CV}$	0.765		0.894	

The residuals of the predicted values obtained by PCR are plotted against the experimental values in Fig. 5A. The spread of residuals in both sides of zero indicate that no systematic error exists in the development of the linear regression method.

The developed PCR model was validated by some statistics parameters such as PRESS and RMSE for test and training sets and results are reported in the Table 4. It is clear that this model on the basis of statistics must be rejected.With respect to these results we decided to try a nonlinear regression method, least squares-support vector machine, to obtain robust and predictive model able to describe a relationship between the structure and p38 inhibitory activity of the compounds of interest. On the other words, another way to find a relationship between the biological activity and PCs is a nonlinear modeling using PCs as input and LS-SVM as a regression tool.

As explained in earlier section, as a linear method for dimensionality reduction, PCA can transform the input data set from its original form to its new form. In the case of a high number of input vectors (e.g descriptors or PCs), irrelevant, redundant, and noisy vectors might be included in the data set, simultaneously, meaningful vectors could be hidden (40). For a large number of input vectors, the probability of chance correlation also increases (41). Moreover, high number of input variables may prevent a nonlinear regression model (such as LS-SVM) from finding optimized models (42). Therefore, PCA input selection is essential in order to improve the accuracy rate of pattern recognition analysis with LS-SVM. After the PCA pre-processing procedure to the input vectors, all the PCs of a training set can be acquired. Then, the PCs were input to the LS-SVM in sequence, i.e., the largest PC was employed as the input vector of the corresponding LS-SVM at first, and then the largest and the second largest one was employed as LS-SVM input data set. In the third step, the third largest one was also included in the input data set of LS-SVM, and so forth. The processes continued until all the PCs represented nearly all the variability of the training set were included in the input set. The optimum number of PCs that gives the best pattern recognition results was adopted to carry out the regression.

In order to determine the optimum number of PCs for the LS-SVM model, the cross validation procedure was applied. There are several cross validation routines and “leave one subject out” was used in our experiments. As the training set was performed with 32 molecules, the modeling procedure was carried out on 31 of them. The process was repeated 32 times and predicted and experimental biological activities were compared. The root mean of square of errors for cross validation (RMSECV) was computed. A plot of the RMSECV against the number of PCs for each individual component indicates a minimum value for optimal number of factors (3PCs) (Fig. 6).

LS-SVM was carried out with radial basis function (RBF) as a kernel functions. In the model development phase using RBF kernel, γ and $\sigma 2$ parameters were a manageable task, similar to the process employed to select the number of factors for PCA-based regression model, but in this case for a two-dimensional problem. An advantage of LS-SVM over classical, SVM is that only these two parameters (but not three parameters as in SVM) are to be optimized, and during the grid-searching process, the mesh plot of the RMSCV changing could be visualized easily. It must be noted that the quality of developed LS-SVM model for regression depends on gamma and sig2 parameters. Gamma is a regularization parameter. Sig2 is a kernel parameter that must be optimized. To find out the optimal values of the parameters, a grid search was performed based on root mean square error for prediction set (RMSEP) and also root mean square error of cross validation (RMSECV). This grid search was performed on the original training set for all parameter combinations of gamma and sig2 from 1 to 400 and 1 to 200, with increment steps of 1 for both of them. These ranges were selected on the basis of previous studies. A robust model is attained by selecting parameters that give the lowest error. The surface plots of RMSECV and RMSEP as a function of gamma and sig2 are shown in Fig. 7.

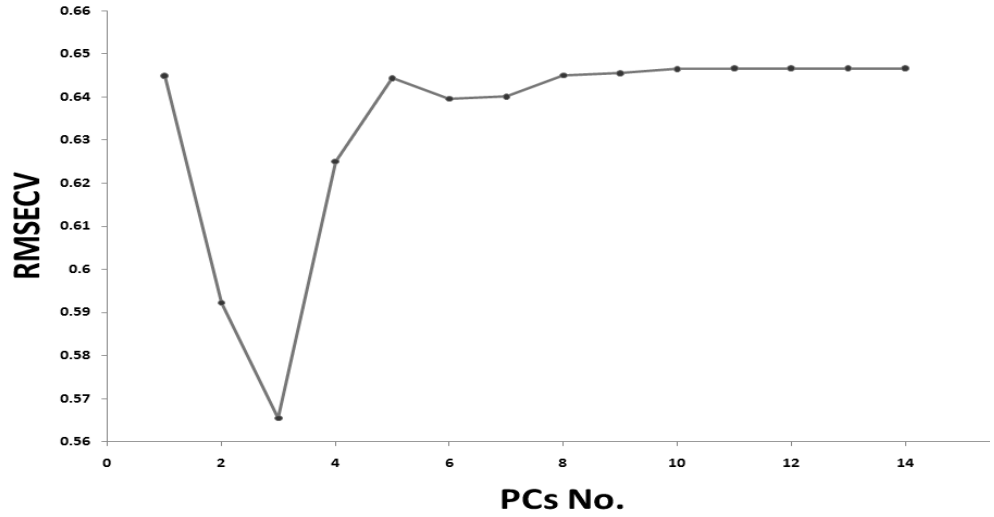


Fig. 6. Optimization number of principal components using root mean square error of cross validation.

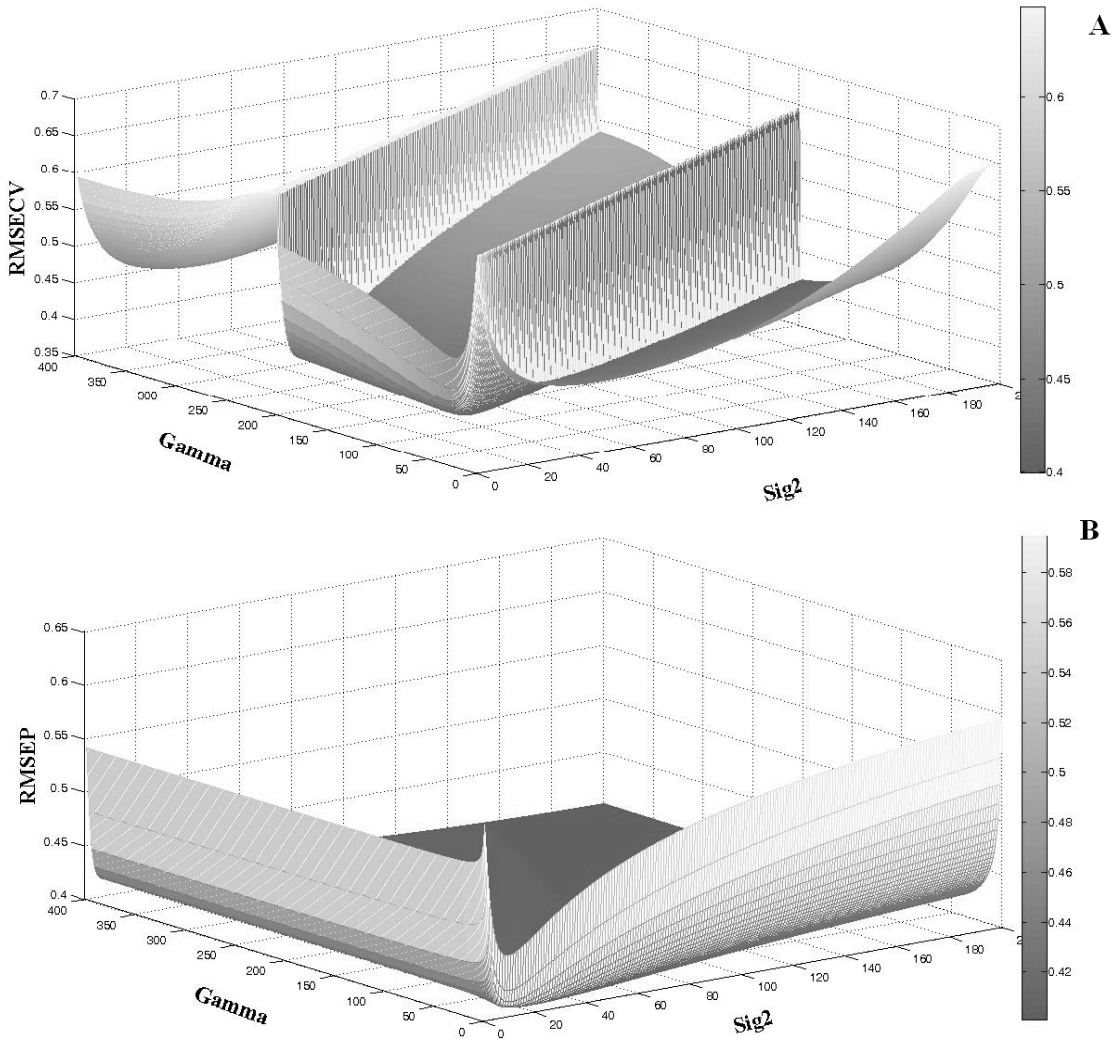


Fig. 7. Optimization the values of gamma and sig2 using A; root mean of square of errors for cross validation and B; root mean square error for prediction set in root mean square error for prediction set.

The results indicate that an LS-SVM with gamma of 28, and sig2 of 17 resulted in the optimum LS-SVM performance.

The nonlinear regression method was trained using training objects and it was evaluated by test molecules. Predicted activities and relative error of prediction by model (RE) for training and test sets are listed in Table 3. Low RE confirms high predictivity of the model.Fig. 4B depicts the plots of observed versus predicted values for training and test sets.

The residuals of the LS-SVM predicted values are plotted against the experimental values in Fig. 5B and show no systematic error in developed model.

The Williams plot for PCR and LS-SVM models are reported in Fig. 8.

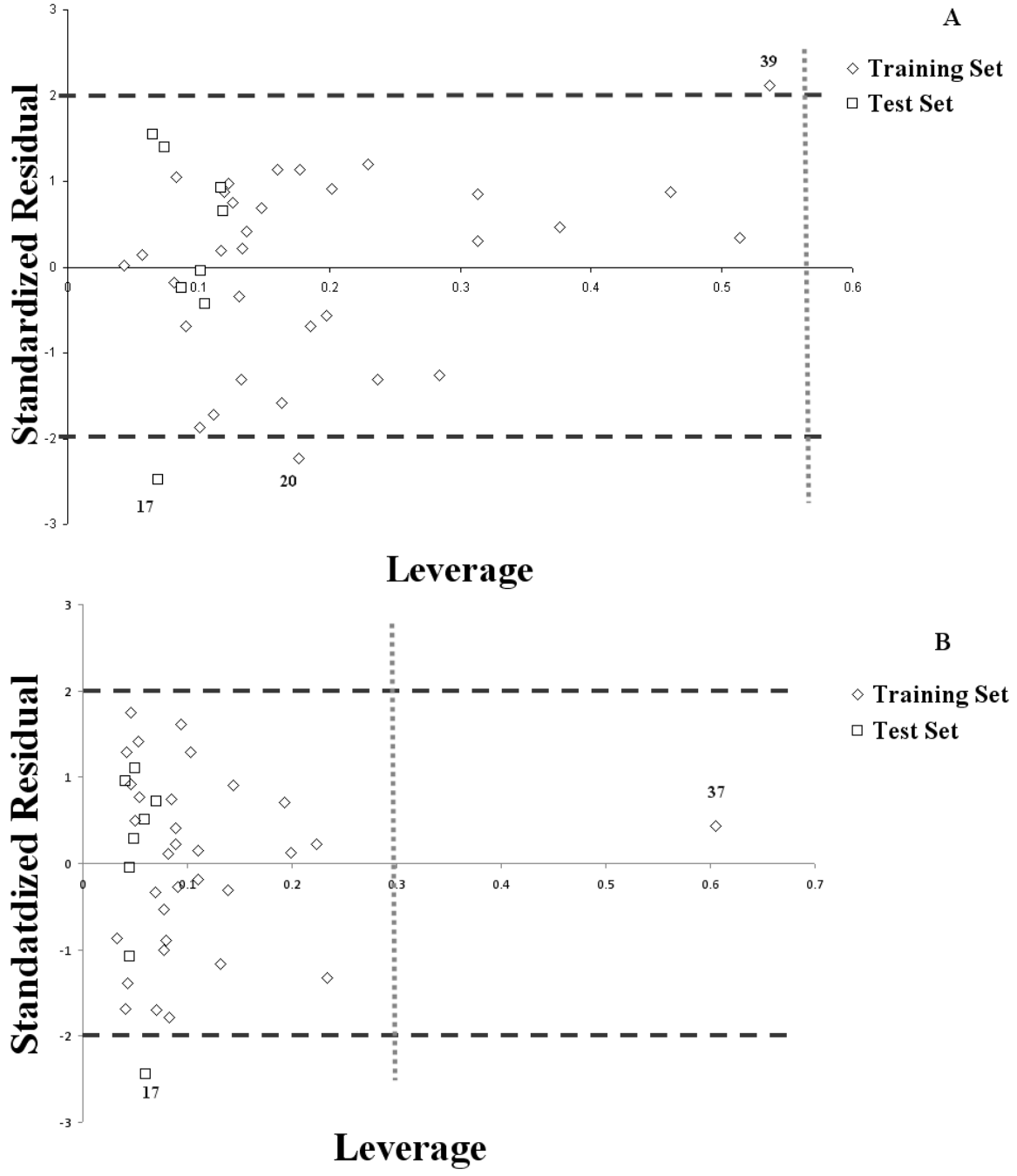


Fig. 8. Williams plot of A; PCR and B; least square support vector machine models**.**

DISCUSSION

The advantages of LS-SVM are the robustness and ability to handle nonlinear responses more effectively in the regression step. The disadvantage is that it is somewhat computationally demanding, even after the model is built. However, with the ever lower price of computers with high performance, this is no longer a major limiting factor. Results of the two LS-SVM and PCR models have been reported in Table 4.

An important validation test of the QSAR model can be accomplished by calculating predicted (Y_{pred}) activity for the test set and comparing such estimates with the corresponding "experimental" values (Y_{exp}). A regular way to quantify the predictivity in a typical QSAR study is through the RMSEP statistic. RMSEP is calculated as SQRT (Σ ((Yexp − Ypred)2)/N),where N is the number of molecules featuring in the test set. An equivalent statistics is obtainable for the training set, however, we here call it root mean square error of training (RMSET) to distinguish it from RMSEP. Table 4 summarizes the prediction results.

The R^2 for test set estimated by PCR and LS-SVM were 0.506 and 0.865, correspondingly, so these models explained the 50% and 86% of the variance for the experimental values of p38 inhibitory activity.

As shown in Table 4, the RMSEP values match very well their RMSET counterparts. RMSEP and RMSET for LS-SVM model is significantly smaller than RMSEP and RMSET for PCR model.

These are very encouraging results. As seen in Table 4, the RMSET, PRESS and RMSEP have decreased by using LS-SVM method. This means that LS-SVM is able to remove unqualified variables and noises; hence, the positive effect of the LS-SVM is more sensible.

It can be seen that the LS-SVM model has higher square of correlation coefficient for training and test sets and fewer errors than the PCR model. Thus, the LS-SVM model produced more accurate results. On the other hand, the PCR model achieves faster training speed.

Outliers in a typical QSAR study are molecules which are extensively separated from the main body of molecules in a dataset of interest. They are a common feature of many real data sets. The presence of outliers can have a deleterious effect on any further processing or regression of the data. A molecule in a given data set may be an outlier with respect to the independent vectors (e.g. PCs) and/or with respect to the dependent vector (e.g. biological activities). Regarding the first aspect, the leverage matrix, H, also called influence matrix, is an important tool in regression diagnostics containing information on the independent variables on which the model is constructed (43). The leverage matrix, H, is a symmetric matrix defined as:

$$H = X(X^T X)^{-1} X \tag{9}$$

Where, X is the matrix consist of PCs of interest for model building, i.e. a matrix with n rows (where n is the number of molecules) and p columns (where p' is the number of model PCs). Molecules whose h_{ii} values are greater than a warning leverage limit (WLL) h^* can be considered as having a great influence (leverage) on the developed model. The warning leverage limit h^* is defined as:

$$h^* = \frac{3(p+1)}{n} \tag{10}$$

Warning leverage limit for PCR and LS-SVM models is 0.562 and 0.300 respectively. A leverage greater than the warning lever-age h^* means that the compound-predicted biological activity can be extrapolated from the model, and therefore, the predicted value must be applied with great care.

Regarding the second aspect, the standardized residuals in prediction can be calculated as the ordinary residuals in prediction divided by the residual standard deviation:

$$\hat{e}_i = \frac{\hat{y}_{i/i} - y_i}{s.\sqrt{1 - h_{ii}}} \tag{11}$$

where, $\hat{e}_i$ is the standardized residual in prediction of the *ith* molecule, $\hat{y}_{i/i}$ and y_i are, respectively, the predicted and the experimental activity of the *ith* molecule, h_{ii} is

the leverage value of the *ith* molecule and *s* is the standard error of the estimate:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)}{n - p}} \quad (12)$$

where, $\hat{y}_i$ is the predicted biological activity of the *ith* molecule. Objects whose $|\hat{e}_i|$ value is greater than 2^{42} can be considered as outliers with respect to the dependent variable (biological activity).

Leverage and standardized residuals in prediction are employed to build graphics (Williams Plot) for the detection of outliers and/or molecules with high influence on the results. It is used to visualize the applicability of domain (a theoretical zone in chemical space, defined by the model input vectors and modeled response, and thus by the nature of the compounds considered in the training set, as represented in each model by specific molecular PCs) of a QSAR model.

The Williams plot, the plot of the standardized residuals versus the leverage (h_{ii}), defined the domain of applicability of the model as a squared area within ± 2 band for residuals and WLL for models. This plot can be used for an immediate and simple graphical detection of both the response outliers (i.e., molecules with standardized residuals greater than two standard deviation units, >2) and structurally influential compounds in a model (h>h*).

As mentioned above, the Williams plot for PCR and LS-SVM models are depicted in Fig. 8. For PCR model, as seen (Fig. 8A), all molecules had acceptable leverage values while few have standardized residuals in prediction higher than the critical value (molecules 17, 20, and 39). For LS-SVM model, as can be seen in Fig. 8B, almost all molecules of interest used lie within area defined by standardized residual threshold and leverage.

Actually, compound 37 has leverage higher than the WLL but show standardized residual within the limits. That is to say, it is any compounds completely outside the AD of the model, as defined by the vertical line (WLL). Thus, there are not any compounds that are both a response outlier and a high leverage chemical.

CONCLUSION

For the production of potent p38 inhibitor compounds it is necessary to have reliable data of biological activity and potency. Unfortunately, the availability of experimentally obtained data is very limited to be useful for screening purposes. QSAR modeling is an alternative approach applicable for filling data gaps, ranking compounds and thus producing p38 inhibitory activity lists. A promising regression model was developed to determine p38 inhibitory activity by 3, 4-dihydropyrido [3,2-d] pyrimidone derivatives in a reproducible and reliable way by using LS-SVM based on PCA.

The resulting procedure obtained for the direct determination of the p38 inhibitory activity without additional preprocessing is attractive, showing the potentiality of the LS-SVM method that permits generation of simple models with no degradation in prediction and validation ability. Further investigations have to be performed to confirm the potentiality of this procedure.

REFERENCES

1. Foster ML, Halley F, Souness JE. Potential of p38 inhibitors in the treatment of rheumatiod arthritis. Drug News and Perspectives. 2000;13:488-497.
2. Rutgeerts P, D'Haens G, Targan S, Vasiliauskas E, Hanauer SB, Present DH, *et al*. Efficacy and safety of retreatment with anti-tumor necrosis factor antibody (infliximab) to maintain remission in Crohn's disease. Gastroenterol. 1999;117:761-769.
3. Paul A, Wilson S, Belham CM, Robinson CJM, Scott PH, Gould GW, *et al*. Stress-activated protein kinases: Activation, regulation and function. Cell Sign. 1997;9;403-410.
4. Han J, Lee JD, Jiang Y, Li Z, Feng L,Ulevitch RJ. Characterization of the structure and function of a novel MAP kinase kinase (MKKG). J Biol Chem. 1996;271:2886-2891.
5. Raingeaud J, Gupta S, Rogers JS, Martin Dickens M, Han J, Ulevitch RJ *et al.* Pro-inflammatory cytokines and environmental stress cause p38 mitogen- activated protein kinase activation by dual phosphorylation on tyrosine and threonine. J Biol Chem. 1995;270:7420-7426.
6. Hansch C, Leo A. Substituent constants for correlation analysis in chemistry and biology. New York: John Wiley & Sons; 1979:65–167

7. Hasegawa K, Yokoo N, Watanabe K, Hirata M, Miyashita Y,Sasaki S. Multivariate Free-Wilson analysis of α-chymotrypsin inhibitors using PLS. Chemometr intell lab sys. 1996;33:63-69.
8. Shahlaei M, Fassihi A, Nazemi A. QSAR study of some 5-methyl/trifluoromethoxy-1H-indole-2, 3-dione-3-thiosemicarbazone derivatives as tubercular agents. Res Pharm Sci. 2009;4:123-131
9. Shahlaei M, Fassihi A, Saghaie L, Zare AR. Prediction of partition coefficient of some 3-hydroxy pyridine-4-one derivatives using combined partial least square regression and genetic algorithm. Res Pharm Sci. 2014;9:143-153.
10. Sabet R, Shahlaei M, Fassihi A. QSAR study of anthranilic acid sulfonamides as inhibitors of methionine aminopeptidase-2 using different chemometrics tools. World scientific and engineering academy and society (WSEAS);2009. p.119-125.
11. Saghaie L, Sakhi H, Sabzyan H, Shahlaei M, Shamshirian D. Stepwise MLR and PCR QSAR study of the pharmaceutical activities of antimalarial 3-hydroxypyridinone agents using B3LYP/6-311++ G** descriptors. Med Chem Res. 2013;22:1679-1688.
12. Saghaie L, Shahlaei M, Fassihi A, Madadkar-Sobhani A, Gholivand MB, Pourhossein A. QSAR Analysis for Some diaryl-substituted Pyrazoles as CCR2 inhibitors by GA-Stepwise MLR. Chem Biol Drug Des. 2011;77:75-85.
13. 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 Modell. 2010;29:518-528.
14. 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.
15. Shahlaei M, Fassihi A, Saghaie L, Arkan E, Madadkar-Sobhani A, Pourhossein A. Computational evaluation of some indenopyrazole derivatives as anticancer compounds; application of QSAR and docking methodologies. J Enzym Inhib Med Chem. 2013;28:16-32.
16. Shahlaei M, Madadkar-Sobhani A, Fassihi A, Saghaie L, Arkan E. QSAR study of some CCR5 antagonists as anti-HIV agents using radial basis function neural network and general regression neural network on the basis of principal components. Med Chem Res. 2011;1-17.
17. Shahlaei M, Madadkar-Sobhani A, Fassihi A, Saghaie L, Shamshirian D,Sakhi H. Comparative quantitative structure–activity relationship study of some 1-aminocyclopentyl-3-carboxyamides as CCR2 inhibitors using stepwise MLR, FA-MLR, and GA-PLS. Med Chem Res. 2012;21:100-115.
18. Shahlaei M, Madadkar-Sobhani A, Saghaie L, Fassihi A. Application of an expert system based on genetic algorithm–adaptive neuro-fuzzy inference system (GA–ANFIS) in QSAR of cathepsin K inhibitors. Expert Syst Appl. 2012;39: 6182–6191
19. 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.
20. Shahlaie M, Fassihi A, Pourhossein A, Arkan E. Statistically validated QSAR study of some antagonists of the human CCR5 receptor using least square support vector machine based on the genetic algorithm and factor analysis. Med Chem Res. 2013; 22: 1399-1414.
21. Cortes C,Vapnik V. Support-vector networks. Mach learn. 1995;20:273-297.
22. Chen C, Zhou X, Tian Y, Zou X, Cai P. Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem. 2006;357:116-121.
23. Shen Q, Shi WM, Kong W, Ye BX. A combination of modified particle swarm optimization algorithm and support vector machine for gene selection and tumor classification. Talanta. 2007;71:1679-1683.
24. Suykens JAK,Vandewalle J. Least squares support vector machine classifiers. Neural Processing Letters. 1999;9:293-300.
25. Huo HB, Zhu XJ, Cao GY. Nonlinear modeling of a SOFC stack based on a least squares support vector machine. J Power Sources. 2006;162:1220-1225.
26. Asl BM, Setarehdan SK, Mohebbi M. Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal. Artif Intell Med. 2008;44;51-64.
27. Polat K,Güneş S. Detection of ECG Arrhythmia using a differential expert system approach based on principal component analysis and least square support vector machine. Appl Math Comput. 2007;186:898-906.
28. Liu L, Stelmach JE, Natarajan SR, Chen M-H, Singh, SB, Schwartz CD, et al. SAR of 3,4-Dihydropyrido [3,2-d] pyrimidone p38 Inhibitors. Bioorg Med Chem Lett. 2003;13:3979–3982.
29. Hyperchem. Hyperchem, Molecular Modeling System. In: Developed by Hyper Cube I, editor.: Hyper Cube, Inc. and Auto Desk, Inc.
30. Todeschini R, Consonni V, Mauri A, Pavan M. DRAGON—Software for the calculation of molecular descriptors. In. 5 ed; 2004.
31. Todeschini R, Consonni V. Handbook of molecular descriptors. Weinheim, Germany: Wiley-VCH; 2000.
32. Burges CJC. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery. 1998;2:121-167.
33. Vapnik V. Statistical learning theory. 1998. Wiley, New York; 1998.
34. Caballero J, Fernández L, Garriga M, Abreu JI, Collina S, Fernández M. Proteometric study of ghrelin receptor function variations upon mutations using amino acid sequence autocorrelation vectors and genetic algorithm-based least square support vector machines. J Mol Graph Modell. 2007; 26:166-178.

35. Suykens JAK,Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett. 1999;9:293-300.
36. Tropsha A, Gramatica P,Gombar V. The Importance of Being Earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci. 2003;22:69-77.
37. Atkinson AC. Plots, Transformations and Regression. Oxford, UK: Clarendon Press; 1985.
38. Golbraikh A, Tropsha A. Beware of q2. J Mol Graph Model. 2002;20:269-276.
39. He Y, Li X, Deng X. Discrimination of varieties of tea using near infrared spectroscopy by principal component analysis and BP model. J Food Eng. 2007;79:1238-1242.
40. Seasholtz MB, Kowalski B. The parsimony principle applied to multivariate calibration. Anal Chim Acta. 1993;277:165-177.
41. Livingstone DJ, Manallack DT. Statistics using neural networks: chance effects. J Med Chem. 1993;36:1295-1297.
42. Broadhurst D, Goodacre R, Jones A, Rowland JJ, Kell DB. Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. Anal Chim Acta. 1997;348:71-86.
43. Cook R, Weisberg S. Residuals and influence in regression. NY: Chapman & Hall; 1982.